Artificial Intelligence Review (2024) 57:51
https://doi.org/10.1007/s10462-023-10675-1

Gene selection for high dimensional biological datasets using hybrid island binary artificial bee colony with chaos game optimization

Maha Nssibi[1,2] · Ghaith Manita[1,3] · Amit Chhabra[4] · Seyedali Mirjalili[5,6] · Ouajdi Korbaa[7]

Accepted: 20 December 2023 / Published online: 13 February 2024

Abstract

Microarray technology, as applied to the fields of bioinformatics, biotechnology, and bioengineering, has made remarkable progress in both the treatment and prediction of many biological problems. However, this technology presents a critical challenge due to the size of the numerous genes present in the high-dimensional biological datasets associated with an experiment, which leads to a curse of dimensionality on biological data. Such high dimensionality of real biological data sets not only increases memory requirements and training costs, but also reduces the ability of learning algorithms to generalise. Consequently, multiple feature selection (FS) methods have been proposed by researchers to choose the most significant and precise subset of classified genes from gene expression datasets while maintaining high classification accuracy. In this research work, a novel binary method called *i*BABC-CGO based on the island model of the artificial bee colony algorithm, combined with the chaos game optimization algorithm and SVM classifier, is suggested for FS problems using gene expression data. Due to the binary nature of FS problems, two distinct transfer functions are employed for converting the continuous search space into a binary one, thus improving the efficiency of the exploration and exploitation phases. The suggested strategy is tested on a variety of biological datasets with different scales and compared to popular metaheuristic-based, filter-based, and hybrid FS methods. Experimental results supplemented with the statistical measures, box plots, Wilcoxon tests, Friedman tests, and radar plots demonstrate that compared to prior methods, the proposed *i*BABC-CGO exhibit competitive performance in terms of classification accuracy, selection of the most relevant subset of genes, data variability, and convergence rate. The suggested method is also proven to identify unique sets of informative, relevant genes successfully with the highest overall average accuracy in 15 tested biological datasets. Additionally, the biological interpretations of the selected genes by the proposed method are also provided in our research work.

Keywords Feature selection · Island model · Binary representation · Artificial bee colony · Gene selection · Chaos game optimization

Extended author information available on the last page of the article

1 Introduction

In the domains of bioinformatics and biotechnology, DNA microarray technology is regarded as a valuable asset. This technology's breakthrough, combined with gene expression techniques, has revealed various mysteries about the biological characteristics of all live cells, enabling the examination and therapy of such endogenous genetic expression. However, conducting an experiment on hundreds of genes simultaneously (Yang et al. 2006), remains challenging work, not only because of the extensive number of genes but also because of the redundancy and irrelevance of some of them, which reduces and affects the performance of the microarray technology. As a result, the gene selection issue is being addressed using the Feature Selection (FS) technique (Dash and Liu 1997). By using gene expression data, the feature selection strategy can decrease the number of features, shorten calculation times, improve accuracy, and make data easier to understand.

The fundamental goal of FS process is the reduction of the number of features (*NFs*) by eliminating unnecessary ones. It could facilitate model inference and increase the classification model's precision. Filter, wrapper, and embedding are the different feature selection techniques (Chandrashekar and Sahin 2014). Wrapper and embedded methods (Chen and Chen 2015; Lal et al. 2006; Yan et al. 2019; Erguzel et al. 2015) are based on classification techniques, and gene selection is a part of the training phase of the learning algorithms. In contrast, filter methods are based on ranking techniques to select genes that are ranked above the fixed threshold (Sánchez-Maroño et al. 2007). Due to the size of the solution space and the overall *NFs* count, Selecting the optimal subset of attributes is viewed as a hard combinatorial optimization problem (OP), with NP-completeness (Wang et al. 2007).

The search for the nearly ideal subset of characteristics is a crucial consideration while building an FS algorithm. The standard comprehensive approaches, such as breadth searches, depth searches, and others, are impractical for choosing the optimal subset of characteristics in large datasets. The production and evaluation of 2N subsets of a dataset with N features are required by wrapper-based methods like neural networks (Guyon and Elisseeff 2003), which is a computationally demanding task, especially when assessing subsets separately. FS is therefore thought of as an NP-hard optimization issue. Its primary purpose is to pick the fewest possible characteristics while maintaining the highest level of classification accuracy. To overcome this difficulty, FS is often built as a single-objective OP by merging these two goals using the weighted-sum approach, or as a multi-objective OP to discover compromise alternatives between the two competing goals (Xue et al. 2015). In the literature, two main objectives are frequently used: minimizing the *NFs* count and the classification error rate, which do not necessarily conflict with one another. For instance, in some subspaces, reducing the *NFs* count also reduces the classification error rate (CER) because redundant features are eliminated (Xue et al. 2012; Vieira et al. 2009; Al-Tashi et al. 2018). Due to this reason, we choose to use a single-objective strategy rather than a multi-objective one.

Furthermore, metaheuristic strategies are incorporated into feature selection since they are superior methodologies in many NP-hard issues. It can be explained by the fact that the search for the best subset in feature space is also an NP-hard problem (Yusta 2009). Additionally, compared to conventional optimization techniques, the wrapper approach based on meta-heuristic algorithms may adjust classifier parameters and choose the best feature subset, both of which can enhance classification outcomes (Huang 2009; Oliveira et al. 2010; Saha et al. 2009). Generally, metaheuristics are inspired by nature, biological and social behaviour, several approaches were proposed to handle the FS problem, such as Simulated Annealing (SA) (Meiri and Zahavi 2006), Genetic Algorithm (GA) (Oliveira et al. 2003), Ant Colony Optimization (ACO) (Chen et al.

2010; Liu et al. 2013), Differential Evolution (DE) (Hancer et al. 2018), Artificial Bee Colony (ABC) (Rao et al. 2019), Particle swarm optimization (PSO) (Wang et al. 2018), Crow Search Algorithm (CSA) (Al-Thanoon et al. 2021), Chaotic Binary Black Hole Algorithm (CBBHA) (Qasim et al. 2020), Tabu Search (TS) (Oduntan et al. 2008), etc. However, these mentioned approaches demand continuous adjustment of parameters and lack rapidity in execution time.

On the other hand, all metaheuristic algorithms must strike a balance between the exploitation and exploitation phases to prevent being caught in local optima or failing to converge (Del Ser et al. 2019). These issues have risen on by the MH algorithms' unpredictable solution-seeking process. In this situation, it is necessary to combine concepts from many scientific disciplines. Hybridization, which combines the best features of several algorithms into a single improved technique, can result in an algorithm with higher performance and accuracy (Talbi 2002). The literature claims that hybrid algorithms outperformed single ones. However, as per the No Free Lunch (NFL) theorem, no strategy is better than all others in all feature selection tasks (Wolpert and Macready 1997). As a result, in order to better handle feature selection challenges, new algorithms must be developed or existing algorithms must be improved by modifying some of their variables. Consequently, we proposed a novel hybrid FS technique, named *i*ABC-CGO, which is further adapted to the discrete FS problem by developing the binary variant *i*BABC-CGO.

The main aims and contributions of this research work are summarized as follows:

(1) This paper proposes an enhanced binary version of an island-based model of artificial bee colonies (*i*BABC) combined with Chaos Game Optimization (CGO) to tackle the FS problem using gene expression data. The integration of CGO principles in the migration process aims to improve the convergence behaviour of *i*BABC and helps in escaping local optimum.

(2) The performance of the suggested variants, *i*BABC-CGO-S and *i*BABC-CGO-V, is first verified by extensive testing on 15 challenging biological datasets. With the use of average accuracy values and the average number of selected features, performance comparisons have been done with a variety of current metaheuristics. Results reveal that the suggested strategy maintains performance accuracy despite controlling a large number of genes and yielding the most important subset of attributes.

(3) The fitness of proposed *i*BABC-CGO variants for different biological data sets is further tested and analysed with the standard statistical measures, Wilcoxon tests, Friedman tests, Quade test, box plots, and radar plots. All of these tests and plots further demonstrate the superiority of the proposed method in terms of average accuracy over the compared metaheuristics for most of the datasets.

(4) Finally, we also provided the biological interpretations of the genes, which are selected by the proposed *i*BABC-CGO method.

Accordingly, this paper is structured as follows: Section 2 provides insight into the established prior works of the FS problem using gene expression data, Section 3 explains the methodology of the island artificial bee colony (*i*ABC) for global optimization metaheuristic, Section 4 details the proposed binary version of *i*ABC-CGO named *i*BABC with the two versions *i*BABC-CGO-S and *i*BABC-CGO-V. Section 5 provides the experimental results based on several gene expression datasets, and Section 6 presents the biological interpretations of the selected genes by the proposed *i*BABC-CGO method. Section 7 provides insights into the pros and cons of the proposed approach. Finally, Section 8 presents the paper's conclusion and future directions.

2 Related works

Metaheuristic techniques have emerged as powerful tools in the realm of optimization problems. They have provided fresh insights into various optimization challenges, including large-scale optimization functions, combinatorial optimization, and bioinformatics.

In large-scale optimization problems, the dimensionality of the solution space becomes prohibitively high. Traditional optimization techniques often struggle with such problems due to the sheer number of variables involved, leading to computational inefficiencies or even infeasibility. Metaheuristics, with their iterative and heuristic-based approach, offer a more scalable solution. They can handle a large number of variables and constraints more effectively, making them particularly well-suited for problems where the search space is vast and the optimal solution is difficult to locate.

Metaheuristics are strategies employed to solve optimization problems by targeting near-optimal solutions for specific problems. This search process can involve multiple agents who work together, applying mathematical equations or rules iteratively until a predefined criterion is met. This point of near-optimality is known as convergence (Yang et al. 2014).

Unlike exact methods that provide optimal solutions at the cost of high computational time, heuristic methods yield near-optimal solutions more quickly but are typically problem-specific. Metaheuristics, being a level above heuristics, have gained popularity due to their ability to deliver solutions with reasonable computational costs. Combining effective heuristics with established metaheuristics can produce high-quality solutions for a wide range of real-world problems.

In order to comprehend metaheuristics, it's crucial to understand the foundational terms in metaheuristic computing, an approach that employs adaptive intelligent behavior. According to Wang (2010), these terms can be defined as follows:

- A heuristic is a problem-solving strategy based on trial-and-error.
- A metaheuristic is a higher-level heuristic used for problem-solving.
- Metaheuristic computing is adaptive computing that uses general heuristic rules to solve a variety of computational problems.

Wang (2010) provides a generalized mathematical representation of a metaheuristic as:

$$MH = (O, A, Rc, Ri, Ro) \tag{1}$$

where:

- O is a set of metaheuristic methodologies (metaheuristic, adaptive, automotive, trial-and-error, cognitive, etc.)
- A is a set of generic algorithms (e.g., genetic algorithm, particle swarm optimization, evolutionary algorithm, ant colony optimization, etc.)
- $Rc = O \times A$ is a set of internal relations.
- $Ri \subseteq A \times A, A \wedge A$ is a set of input relations.
- $Ro \subseteq c \times C$ is a set of output relations.

Additional concepts such as neighborhood search, diversification, intensification, local and global minima, and escaping local minima are also important. Fundamental strategies in metaheuristics involve balancing exploration and exploitation, identifying promising neighbors, avoiding inefficient ones, and limiting searches to unpromising areas.

The exploration versus exploitation dichotomy, central to optimization methods, requires expanding the search to explore unvisited areas (exploration) and focusing on promising regions based on accumulated search experience for optimal utilization and convergence (exploitation) (Črepinšek et al. 2013).

Other key considerations in metaheuristics include local versus global search, single versus population-based algorithms, and hybrid methods. The choice between single-solution or population-based algorithms depends on whether the metaheuristic is more inclined towards exploitation or exploration. Ultimately, balancing exploration and exploitation is vital for both types of algorithms (Abualigah et al. 2022).

Bioinformatics is a multidisciplinary field that combines biology, computer science, and statistics to analyze and interpret biological data. With the advent of high-throughput technologies, bioinformaticians now deal with massive datasets that require efficient processing and analysis. Metaheuristic techniques have found applications in areas such as protein folding, sequence alignment, and phylogenetic tree reconstruction. They offer the ability to handle complex biological data, accommodate uncertainties, and provide solutions that are both computationally feasible and biologically meaningful.

Gene selection plays a crucial role in profiling and predicting different forms of abnormalities in the field of bioinformatics, particularly within the scope of optimization problems. Given the vast amount of gene expression data available, there is a need for a reliable and accurate approach to selecting the most pertinent genes. The feature selection approach serves this purpose by reducing the dimensionality and removing redundancy in gene expression data.

Gene expression data typically involves thousands of genes measured across various experimental conditions or time points. However, not all these genes are equally relevant for understanding or predicting biological processes or abnormalities. Many genes may exhibit similar expression patterns, leading to redundancy in the data, while others may be irrelevant to the specific condition under investigation. This high-dimensional and redundant data poses significant challenges for analysis, making gene selection an essential step.

Feature selection aims to identify a subset of the most informative genes that contribute significantly to the condition under study. This process involves three key steps: evaluation, search, and validation. In the evaluation step, each gene is assessed based on its relevance and contribution to the target condition. Statistical tests, information theory, or machine learning techniques are commonly used to assign scores or ranks to genes according to their importance.

In the search step, different algorithms, such as sequential forward selection, sequential backward elimination, or metaheuristic algorithms, are applied to identify the optimal subset of genes. These algorithms operate iteratively, adding or removing genes based on their scores, and aim to balance the trade-off between including informative genes and minimizing redundancy.

In the validation step, the selected subset of genes is tested on independent data to assess its performance in profiling or predicting the condition under investigation. Techniques like cross-validation or bootstrapping are used to ensure that the selected genes are robust and generalizable across different datasets.

The feature selection approach offers several advantages. By reducing the dimensionality of the data, it facilitates easier and more interpretable analysis. Eliminating redundancy leads to more robust and reliable results, as irrelevant or correlated genes are removed.

Additionally, selecting a smaller subset of genes helps in reducing the cost and time associated with experimental validation.

Researchers have proposed a variety of techniques for applying Feature Selection (FS) to gene expression data. Lu et al. (2017) proposed a hybrid FS approach that integrates the mutual information maximization (MIM) and the adaptive genetic algorithm (AGA) to classify gene expression data. This method reduces the size of the original gene expression datasets and eliminates data redundancies. First, the authors used the MIM filter technique to identify relevant genes. Subsequently, they applied the AGA algorithm, coupled with the extreme learning machine (ELM) as a classifier.

In another study, Dhrif et al. (Dhrif et al. 2019) introduced an improved binary Particle Swarm Optimization (PSO) algorithm with a local search strategy (LS) for feature selection in gene expression data. This approach facilitates the selection of specific features by using the LS to guide the PSO search, thereby successfully reducing the overall number of features. Shukla et al. (2019) presented a hybrid wrapper model for gene selection that combines the gravitational search algorithm (GSA) and a teaching-learning-based optimization algorithm (TLBO). To apply the FS methodology, the authors converted the continuous search space into binary form.

Sharma and Rani (2019) employed a hybrid strategy for gene selection in cancer classification that integrates the Salp Swarm Algorithm (SSA) and a multi-objective spotted hyena optimizer. The authors initially used the filter method to obtain a reduced subset of significant genes. The most pertinent gene subset is then identified using the hybrid gene selection approach, applied to the pre-processed gene expression data. Masoudi et al. (2021) proposed a wrapper FS approach to select the optimal subset of genes based on a Genetic Algorithm (GA) and World Competitive Contests (WCC). This method was applied to 13 biological datasets with diverse features, including cancer diagnostics and drug discovery, and demonstrated better performance than many currently used solutions.

Ghosh et al. (2019) developed a modified version of the Memetic Algorithm (MA), called the Recursive MA (RMA), for gene selection in microarray data. The proposed method outperformed the original MA and GA-based FS algorithms on seven microarray datasets using SVM, KNN, and Multi-layer Perceptron (MLP) classifiers. Kabir et al. (2012) proposed a wrapper FS methodology that employed an enhanced Ant Colony Optimization (ACO) variant, termed ACOFS, as a search method. ACOFS showed remarkable results compared to popular FS approaches when tested on multiple biological datasets. For high-dimensional microarray data classification, authors in (Apolloni et al. 2016) developed two hybrid FS algorithms by combining a Binary Differential Evolution (BDE) algorithm with a rank-based filter methodology. These BDE-based FS algorithms were tested for robustness using SVM, KNN, Naive Bayes (NB), Decision Trees (C4.5), and six high-dimensional microarray datasets. It was observed that the proposed FS methods produced substantially equivalent classification results with more than a 95% reduction in the initial number of features/genes. In a recent study (Shukla et al. 2020), various feature selection methods were investigated using gene expression data.

In the paper (Yaqoob et al. 2023), the authors address the challenge of dimensionality in high-dimensional biomedical data, which complicates the identification of significant genes in diseases like cancer. They explore new machine learning techniques for analyzing raw gene expression data, which is crucial for disease detection, sample classification, and early disease prediction. The paper introduces two dimensionality reduction methods, feature selection and feature extraction, and systematically compares several techniques for analyzing high-dimensional gene expression data. The authors present a review of popular nature-inspired algorithms, focusing on their underlying principles and applications in cancer classification and prediction. The paper also evaluates the pros and cons of using nature-inspired algorithms for biomedical data. This review offers guidance for researchers seeking the most effective algorithms for cancer classification and prediction in high-dimensional biomedical data analysis.

The article (Aziz 2022) addresses the challenge of designing an optimal framework for predicting cancer from high-dimensional and imbalanced microarray data, a common problem in bioinformatics and machine learning. The authors focus on the independent component analysis (ICA) feature extraction method for Naïve Bayes (NB) classification of microarray data. The ICA method effectively extracts independent components from datasets, satisfying the classification criteria of the NB classifier. The authors propose a novel hybrid method based on a nature-inspired metaheuristic algorithm to optimize the genes extracted using ICA. They employ the cuckoo search (CS) and artificial bee colony (ABC) algorithms to find the best subset of features, enhancing the performance of ICA for the NB classifier. According to their research, this is the first application of the CS-ABC approach with ICA to address the dimensionality reduction problem in high-dimensional microarray biomedical datasets. The CS algorithm improves the local search process of the ABC algorithm, and the hybrid CS-ABC method provides better optimal gene sets, improving the classification accuracy of the NB classifier. Experimental comparisons show that the CS-ABC approach with the ICA algorithm performs a deeper search in the iterative process, avoiding premature convergence and producing better results compared to previously published feature selection algorithms for the NB classifier.

The article (Chen et al. 2023) addresses the issue of high-dimensional genetic data in contemporary medicine and biology. The authors propose a new wrapper gene selection algorithm called the artificial bee bare-bone hunger games search (ABHGS), which integrates the hunger games search (HGS) with an artificial bee strategy and a Gaussian bare-bone structure. The performance of ABHGS is evaluated by comparing it with the original HGS and a single strategy embedded in HGS, as well as six classic algorithms and ten advanced algorithms using the CEC 2017 functions. Experimental results show that ABHGS outperforms the original HGS. In comparison to other algorithms, it improves classification accuracy and reduces the number of selected features, indicating its practical utility in spatial search and feature selection.

The work proposed by (Coleto-Alcudia and Vega-Rodríguez 2020), introduces a new hybrid method for gene selection in cancer research, aimed at classifying tissue samples into different classes (normal, tumor, tumor type, etc.) effectively with the fewest number of genes. The proposed approach comprises two steps: gene filtering and optimization. The first step employs the Analytic Hierarchy Process, using five ranking methods to select the most relevant genes and reduce the number of genes for consideration. In the second step, a multi-objective optimization approach is applied to achieve two objectives: minimize the number of selected genes and maximize classification accuracy. An Artificial Bee Colony based on Dominance (ABCD) algorithm is proposed for this purpose. The method is tested on eleven real cancer datasets, and results are compared with several multi-objective methods from the scientific literature. The approach achieves high classification accuracy with small subsets of genes. A biological analysis on the selected genes confirms their relevance, as they are closely linked to their respective cancer datasets.

The paper (Pashaei and Pashaei 2022) presents a new wrapper feature selection method based on the chimp optimization algorithm (ChOA) for the classification of high-dimensional biomedical data. Due to the presence of irrelevant or redundant features in biomedical data, classification methods struggle to accurately identify patterns without a feature selection algorithm. The ChOA is a newly introduced metaheuristic algorithm, and this work explores its potential for feature selection. Two binary variants of the ChOA are proposed, using transfer functions (S-shaped and V-shaped) or a crossover operator to convert the continuous version of ChOA to binary. The proposed methods are validated on five high-dimensional biomedical datasets, as well as datasets from life, text, and image domains. The performance of the proposed approaches is compared to six well-known wrapper-based feature selection methods (GA, PSO, BA, ACO, FA, FP) and two standard filter-based methods using three different classifiers. Experimental results show that the proposed methods effectively remove less significant features, improving classification accuracy, and outperforming other existing methods in terms of selected genes and classification accuracy in most cases.

The paper (Pashaei and Pashaei 2021) presents a new hybrid approach (DBH) for gene selection that combines the strengths of the binary dragonfly algorithm (BDF) and binary black hole algorithm (BBHA). The proposed approach aims to identify a small and stable set of discriminative genes without sacrificing classification accuracy. The approach first applies the minimum redundancy maximum relevancy (MRMR) filter method to reduce feature dimensionality and then uses the hybrid DBH algorithm to select a smaller set of significant genes. The method was evaluated on eight benchmark gene expression datasets and compared against the latest state-of-art techniques, showing a significant improvement in classification accuracy and the number of selected genes. Furthermore, the approach was tested on real RNA-Seq coronavirus-related gene expression data of asthmatic patients to select significant genes for improving the discriminative accuracy of angiotensin-converting enzyme 2 (ACE2), a coronavirus receptor and biomarker for classifying infected and uninfected patients. The results indicate that the MRMR-DBH approach is a promising framework for identifying new combinations of highly discriminative genes with high classification accuracy.

The paper (Pashaei 2022) addresses the challenge of microarray data classification in bioinformatics. The authors propose a new wrapper gene selection method called mutated binary Aquila Optimizer (MBAO) with a time-varying mirrored S-shaped (TVMS) transfer function. This hybrid approach employs the Minimum Redundancy Maximum Relevance (mRMR) filter to initially select top-ranked genes and then uses MBAO-TVMS to identify the most discriminative genes. TVMS is used to convert the continuous version of Aquila Optimizer (AO) to binary, and a mutation mechanism is added to the binary AO to enhance global search capabilities and avoid local optima. The method was tested on eleven benchmark microarray datasets and compared to other state-of-the-art methods. The results indicate that the proposed mRMR-MBAO approach outperforms the mRMR-BAO algorithm and other comparative gene selection methods in terms of classification accuracy and the number of selected genes on most of the medical datasets.

To tackle the challenges associated with analyzing gene expression data generated by DNA microarray technology, authors in (Alomari et al. 2021) propose a new hybrid filter-wrapper gene selection method combining robust Minimum Redundancy Maximum Relevancy (rMRMR) as a filter approach to select top-ranked genes, and a Modified Gray Wolf Optimizer (MGWO) as a wrapper approach to identify smaller, more informative gene sets. The MGWO incorporates new optimization operators inspired by the TRIZ-inventive solution, enhancing population diversity. The method is evaluated on nine well-known microarray datasets using a support vector machine (SVM) for classification. The impact of TRIZ optimization operators on MGWO's convergence behavior is examined, and the results are compared to seven state-of-the-art gene selection methods. The proposed method achieves the best results on four datasets and performs remarkably well on the others. The experiments confirm the effectiveness of the method in searching the gene search space and identifying optimal gene combinations.

However, these methods still have several drawbacks, such as local optima stagnation, premature convergence, and onerous criteria and execution times (Abu Khurma et al. 2022; Agrawal et al. 2021; Abiodun et al. 2021). To mitigate these shortcomings, this paper introduces a new, enhanced wrapper algorithm based on the island model of the ABC metaheuristic. Over the years, numerous extensions and applications of the ABC algorithm have been proposed. For instance, the study in (Zorarpacı and Özel 2016) presented a hybrid approach that integrates differential evolution with the ABC method for feature selection. In (Garro et al. 2014), distance classifiers and the ABC method were employed to classify DNA microarrays. The study in (Alshamlan et al. 2019) introduced an ABC-based technique for accurate cancer microarray data classification, utilizing an SVM as a classifier. Various other hybridizations and expansions of the ABC algorithm have been proposed over time. Recently, (Awadallah et al. 2020) introduced the island model to the standard ABC algorithm for global optimization (denoted as *i*ABC), achieving impressive success compared to its competitors.

Building upon these developments, the objective of this study is twofold. First, we propose a hybridized version of *i*ABC, termed *i*ABC-CGO, which incorporates Chaos Game optimization to tackle the standard ABC's slow convergence problem, which arises from modifying only a single decision vector dimension. Subsequently, we introduce a binary variant of *i*ABC-CGO, referred to as *i*BABC-CGO, designed to address the feature selection (FS) problem in the context of gene expression data. This hybrid approach aims to harness the strengths of both Chaos Game optimization and the island model of the ABC metaheuristic, potentially offering improved convergence rates and enhanced feature selection capabilities.

3 Preliminaries

3.1 Overview of island ABC (*i*ABC) for optimization

ABC algorithm is a swarm-based metaheuristic motivated by how bees forage for food, where the location of the food source indicates potential ideal solutions and the quantity of nectar suggests the quality of the solution (Rao et al. 2019). ABC intrigued the authors of (Awadallah et al. 2020) due to its many benefits, and they suggested a new version of ABC paired with the island model for enhancing the convergence speed and diversity of solutions. However, the original ABC algorithm still struggles with three major deficiencies in the search behaviour, which are: (i) the search equation favours exploration over exploration (Zhu and Kwong 2010; (ii) numerous fitness evaluations (Mernik et al. 2015); and (iii) tendency to stuck in local optima due to premature convergence, especially while applying to complex optimization problems (Karaboga et al. 2014).

Therefore, the island model concept (Wu et al. 2019) has been introduced mainly to address the lack of heterogeneity from which most population-based algorithms suffer. The original algorithm is independently run, either synchronously or asynchronously, on each island. Therefore, a migration mechanism is used to shift certain individuals from one island to another in order to increase the algorithm's effectiveness. This procedure might adhere to a number of strategies that guarantee the spatial exploration of newer areas of the search area (Tomassini 2006). The wide use of this technique can be explained by its balance between exploration and exploitation. Moreover, dividing individuals (potential solutions) into different islands reduces the computational time and increases the probability that weak solutions reach the best values (Whitley et al. 1997).

Several island-based metaheuristic studies were proposed such as island ACO (Mora et al. 2013), island bat algorithm (Al-Betar and Awadallah 2018), island GA (Corcoran and Wainwright 1994; Whitley et al. 1997; Palomo-Romero et al. 2017), island PSO (Abadlia et al. 2017), island harmony search algorithm (Al-Betar et al. 2015), and island crow search algorithm (Turgut et al. 2020). The proposed approaches succeeded in reducing the computational requirements and providing good results. However, choosing the adequate values of different parameters and adopting the suitable migration policy for the island-based technique greatly impact the final results as proved by many studies (Cantú-Paz et al. 1998; Fernandez et al. 2003; Skolicki and De Jong 2005; Tomassini 2006; Ruciński et al. 2010).

The island model requires at first two parameters: the quantity also referred to as the number of islands (I_n) and the size of islands (I_s). Then, the four key variables that determine the migration process between the islands are the migration rate, the migration frequency, the migration policy, and the migration topology. The migration rate (R_m) represents the number of solutions transferred between islands. The migration frequency (F_m) defines the periodic time for the exchange. The migration topology structures the path of exchanging solutions among islands. Several topologies were proposed in the literature and mainly categorized into two sets, one for static and the second for dynamic. Static topologies include ring, mesh, and star, where the structured paths are predefined and remain static during the migration process (da Silveira et al. 2019). Dynamic topologies, as the name implies, randomly define paths and changes in every migration process (Duarte et al. 2017). The migration policy determines which solutions are exchanged between islands. Researchers introduced different policies based on greed or random selection. The most used policies are the best-worst policy dealing with replacing the worst solutions of one island with the best solutions from the other(Kushida et al. 2013). The random policy consists of migrating solutions randomly(Araujo and Merelo 2011). Finally, the migration process with all its factors can be carried out in two ways: synchronously or asynchronously. In *i*ABC algorithm, the artificial bee colony population is divided into islands, and solutions are enhanced separately and locally on each island. Once a certain number of iterations are over, a migration procedure based on random ring topology is applied to exchange solutions within islands. The flowchart of *i*ABC is presented in Fig. 1 and detailed in (Awadallah et al. 2020).

3.2 Overview of choas game optimization (CGO)

Chaos game optimization (CGO) (Talatahari and Azizi 2021), which incorporates both game theory and the mathematical idea of fractals, is a recent innovative optimization metaheuristic method. Fractals are built using a polygon form that begins with a random beginning point and an affine function. The iterative series of points produces the fractal shape by continually applying the chosen function to a new point. The CGO algorithm seeks to produce a Sierpinski Triangle based on characteristics of fractals in chaos theory. There are many solution candidates (X) in the CGO's initial population. Each potential solution (X_i), which comprises the decision variables (xi, j), offers a Sierpinski triangle as an eligible point in the search space. The following criteria are used to generate the eligible points:

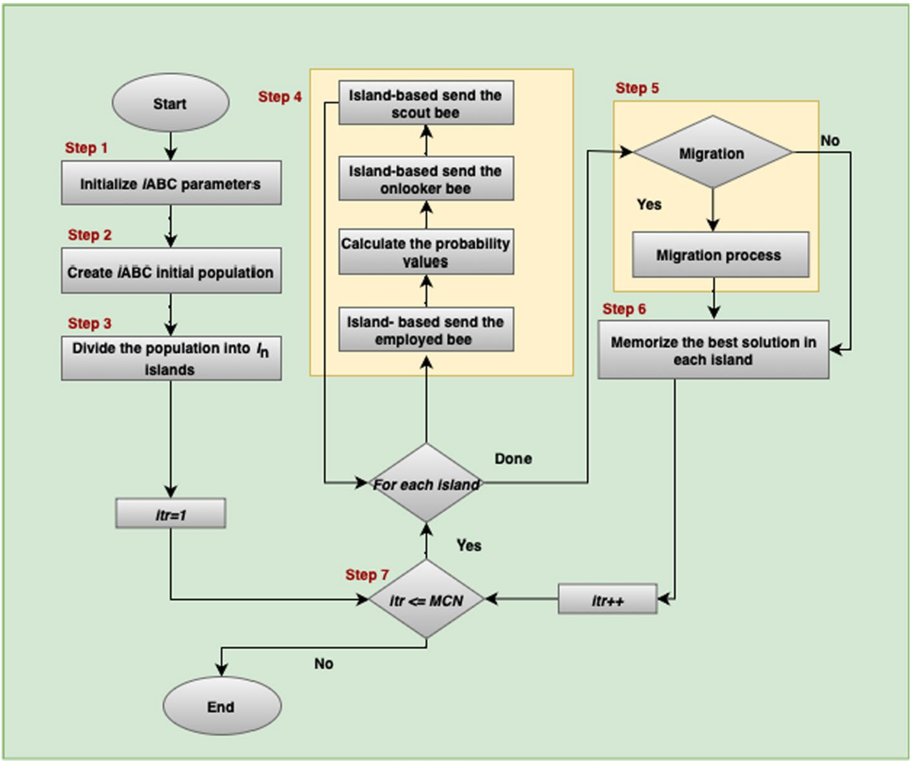

Start
Step 1
Initialize iABC parameters
Step 2
Create iABC initial population
Step 3
Divide the population into I_n islands
Itr=1
Step 4
Island-based send the scout bee
Island-based send the onlooker bee
Calculate the probability values
Island- based send the employed bee
For each island
Done
Yes
Step 7
Itr <= MCN
No
End
Step 5
Migration
No
Yes
Migration process
Step 6
Memorize the best solution in each island
Itr++

Fig. 1 The flowchart of *i*ABC algorithm

$$x_i^j(0) = x_{i,min}^j + \text{rand.}\left(x_{i,max}^j - x_{i,min}^j\right), \quad \begin{cases} i = 1,2,\ldots,n \\ j = 1,2,\ldots,d \end{cases} \tag{2}$$

The dimension of each solution is given by d, where n is the total number of candidates for eligible solutions. The beginning locations of the solutions are indicated by the variable $x_i^j(0)$. The maximum and minimum values of the $x_{i,max}^j$ and $x_{i,min}^j$ variables serve as the j^{th} decision variable of the i^{th} solution, while the *rand* variable denotes a random value in the range of [0, 1] .

For the initial search, a temporary Sierpinski triangle is created inside the search space for each potential solution depending on the locations of three vertices as follows:

- The Global Best (GB).
- The Mean Group (MG_i).
- The i^{th} solution candidate (X_i).

Four different approaches are used to update positions. The first one mimics how the solution X_i moves to GB and MG_i using the following mathematical formulation:

$$\text{Seed}_i^1 = X_i + \alpha_i \times \left(\beta_i \times GB - \gamma_i \times MG_i\right), \quad i = 1,2,\ldots,n \tag{3}$$

where αi denotes a factorial formed at random, βi and γi denote a random number equal to 0 or 1, respectively. The following equation models how GB moves in relation to X_i and MG_i:

$$\text{Seed}\,_i^2 = GB + \alpha_i \times \left(\beta_i \times X_i - \gamma_i \times MG_i\right), \quad i = 1, 2, \dots, n \tag{4}$$

The tertiary technique shows how MG_i moves in the direction of X_i and GB. This technique is mathematically represented:

$$\text{Seed}\,_i^3 = MG_i + \alpha_i \times \left(\beta_i \times X_i - \gamma_i \times GB\right), \quad i = 1, 2, \dots, n \tag{5}$$

The exploitation phase of the CGO optimization process is defined by the position update described by the first three approaches. The mutation for the exploration phase is shown in the fourth technique, which is mathematically represented as follows:

$$\text{Seed}\,_i^4 = X_i\left(x_i^k = x_i^k + R\right), \quad k = [1, 2, \dots, d] \tag{6}$$

where k is a randomly generated integer in the range of [1, d] and R is a uniformly distributed random (UDR) number in the range of [0, 1]. Four possible formulations for α_I are considered to adjust the global and local search rates of the CGO.

$$\alpha_i = \begin{cases} \text{Rand} \\ 2 \times \ \text{Rand} \\ (\delta \times \ \text{Rand}\ ) + 1 \\ (\varepsilon \times \ \text{Rand}\ ) + (\sim \varepsilon) \end{cases}, \tag{7}$$

The variables δ and ε denote random numbers in the interval [0, 1], where *Rand* is a UDR number in the range of [0, 1]. Figure 2 shows the CGO algorithm's flowchart.

The CGO algorithm offers an efficient optimization technique using multigroup behaviour as a basis. Easy to implement and simple to understand is what best describes the CGO algorithm, where it performs well in many optimization problems (Talatahari and Azizi 2020; Ponmalar and Dhanakoti 2022; Ramadan et al. 2021). However, it converges early due to a talent imbalance between exploration and exploitation. In fact,

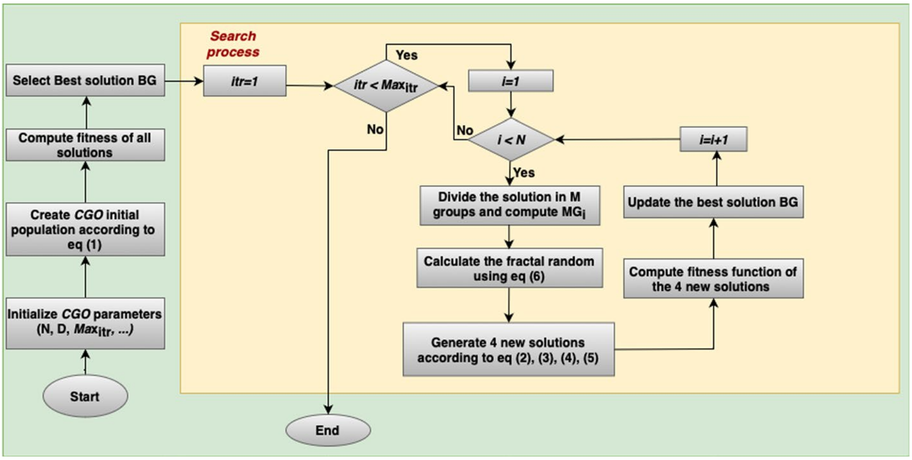

Search process
Select Best solution BG
Compute fitness of all solutions
Create CGO initial population according to eq (1)
Initialize CGO parameters (N, D, Max$_{itr}$, ...)
Start
itr=1
itr < Max$_{itr}$
Yes
No
End
i=1
No
i < N
Yes
Divide the solution in M groups and compute MG$_i$
Calculate the fractal random using eq (6)
Generate 4 new solutions according to eq (2), (3), (4), (5)
Compute fitness function of the 4 new solutions
Update the best solution BG
i=i+1

Fig. 2 The flowchart of CGO algorithm

excessive exploration wastes time and has a poor convergence rate, whereas high exploitation destroys variety and traps people in local optimums (Roshanzamir et al. 2020).

4 Proposed approach

FS can be formulated as a binary optimization problem since it deals with the binary decision of whether to select a feature or not. Consequently, to adapt the FS problem using gene expression data, a hybrid binary version of *i*ABC algorithm with CGO algorithm is suggested in this work named *i*BABC-CGO. The proposed *i*BABC-CGO approach combines the swarm intelligence ABC metaheuristic, the island model concept, the chaos game optimization principles and the binary version to handle the large dimension of the gene expression data and the considerable number of irrelevant and redundant genes.

4.1 *i*ABC-CGO

The island model offers one of the most useful techniques for partitioning the population into a number of sub-population. The searching activities of the metaheuristic algorithm are repeated synchronously or asynchronously on each island. In order to exchange migrants between islands, a migration mechanism must be implemented. This procedure allows the metaheuristic algorithm to rigorously explore the search space resulting in performance improvement of the final solution (Tomassini 2006). The algorithm can arrive at a global optimal solution due to the increased population diversity because of the exchange of individuals with diverse fitness values across islands (Tomassini 2006). Indeed, the count of islands in the model or the frequency of inter-island migrant transfers impacts the efficiency of the algorithm.

The migration policy, which specifies who will be relocated off the island and decides their location on the other island, influences the efficacy of the island model (Araujo and Merelo 2010; Ruciński et al. 2010). Numerous academics have examined the influence of the chosen migration topology on the algorithm efficiency (Cantú-Paz et al. 1998; Tomassini 2006; Ruciński et al. 2010; Fernandez et al. 2003). Therefore, we decide to introduce a new migration policy based on the principles of the CGO algorithm.

Indeed, the CGO method relies on the Sierpinski triangle, which consists of three positional vertices of the global Best, the mean group, and the i^{th} solution candidate (X_i). Since the CGO method is developed on the concept of sub-populations, the idea of replacing the traditional random ring topology with the CGO position update formulas emerged.

Therefore, the $R_{\mathrm{m}} \times I_{\mathrm{s}}$ solutions are to be exchanged among islands after a predefined number of iteration F_{m}. Those solutions are selected using the roulette wheel technique (Lipowski and Lipowska 2012).

The detailed proposed *i*ABC-CGO algorithm is represented by the flowchart in Fig. 3.

4.2 Binary version of *i*ABC-CGO (*i*BABC-CGO)

In the continuous version of the proposed *i*ABC-CGO algorithm, the bees change their position in a continuous search space. Therefore, a feature subset is encoded as a one-dimensional vector with the same length as the *NFs* count in the binary optimization, where limits on the search agents' positions (the bees) are enforced in the order of 0, 1 values. Non-chosen features are set to 0, while those selected are set to 1.

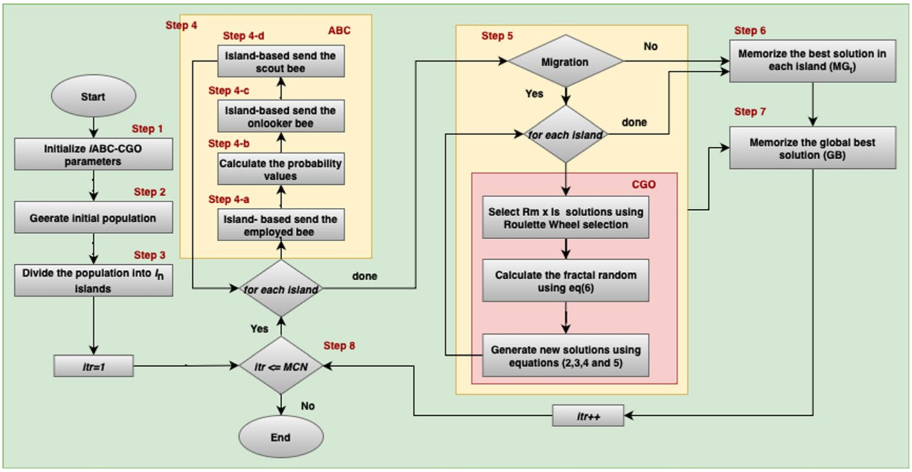

Step 4
Step 4-d
ABC
Island-based send the scout bee
Step 4-c
Island-based send the onlooker bee
Step 4-b
Calculate the probability values
Step 4-a
Island- based send the employed bee
Start
Step 1
Initialize iABC-CGO parameters
Step 2
Geerate initial population
Step 3
Divide the population into I_n islands
Itr=1
for each island
done
Yes
Itr <= MCN
Step 8
No
End
Step 5
Migration
No
Yes
for each island
done
CGO
Select Rm x Is solutions using Roulette Wheel selection
Calculate the fractal random using eq(6)
Generate new solutions using equations (2,3,4 and 5)
Step 6
Memorize the best solution in each island (MG_i)
Step 7
Memorize the global best solution (GB)
Itr++

Fig. 3 The flowchart of *i*ABC-CGO

Afterward, the conversion of the continuous optimization algorithm into a binary version is performed using transfer functions (TFs) which are an effective tool to perform this conversion (Mirjalili and Lewis 2013). In this work, two commonly used S-shaped and V-shaped TFs are chosen to develop the two binary versions of *i*ABC, namely *i*BABC-CGO-S and *i*BABC-CGO-V. The probability of updating an element's position in the TFs *S* and *V* in binary form to be 1 or 0 is given by Eq (7):

$$P(x_i^d(t)) = TF(x_i^d(t)), \tag{8}$$

where TF is the transfer function that can be Sigmoid for the S-shaped TF given in Eq. (8) or Hyperbolic tang for V-shaped TF defined in Eq. (9), and $x_i^d(t)$ presents the i^{th} bee position in dimension d^{th} at iteration t.

$$TF(x_i^d(t)) = \frac{1}{1 + e^{-x_i^d(t)}} \tag{9}$$

$$TF(x_i^d(t)) = |\tanh(x_i^d(t))| \tag{10}$$

The probability value produced by Eq. (7) is then applied to Eq. (10) to generate the binary value for the S-shaped transfer function or Eq. (11) for the V-shaped TF in order to update the position vectors of bees.

$$x_i^d(t+1) = \left\{ \begin{array}{lll} 0 & \text{if} & \text{rand} < P(x_i^d(t)) \\ 1 & \text{if} & \text{rand} \geq P(x_i^d(t)) \end{array} \right\} \tag{11}$$

$$x_i^d(t+1) = \left\{ \begin{array}{lll} x_i^d(t)^{-1} & \text{if} & \text{rand} < P(x_i^d(t)) \\ x_i^d(t) & \text{if} & \text{rand} \geq P(x_i^d(t)) \end{array} \right\} \tag{12}$$

with rand as a random vector in [0, 1].

The general steps of *i*BABC-CGO are presented in Algorithm 1 and followed by a flowchart as illustrated in Fig. 4 and detailed as follows:

- Step 1: Set *i*BABC-CGO algorithm's parameters to initial values. These parameters are Solution number (SN), Maximum cycle number (MCN), limit, Island number (I_n), Island size (I_s), Migration frequency (F_m) and Migration rate (R_m).
- Step 2: Generate randomly the initial population. Each Bee (i.e. Solution) is spawned according to this formula: $X_i^j = r * (Ub_j - Ul_j) + Ul$, where r is a uniform random number between 0 and 1, and Ub_j and Lb_j are respectively the upper bound and the lower bound of the j^{th} dimension.
- Step 3: Create a set of I_n islands from the *i*BABC-CGO population.
- Step 4: The search process according to ABC principles. The four stages (Steps 4a–4d) of this procedure are as follows:
  - Step 4a: Send the employed bees using island based mechanism as per lines no. 19–31 in the pseudo-code (Algorithm 1). This stage involves doing a binary transformation in accordance with the function *binary_transformation*'s pseudo code.
  - Step 4b: Calculate the probability values as per line no. 32–36 (Algorithm 1). In this stage, a binary transformation is carried out in accordance with the function *binary_transformation*'s pseudo code.
  - Step 4c: Send the onlooker bee as per line no. 37–46 (Algorithm 1). In accordance with the pseudo-code of the function *binary_transformation*, a binary transformation is carried out in this phase.
  - Step 4d: Send the scout bee as per line no. 53–57 (Algorithm 1). During this step, a binary transformation is performed as per the pseudo-code of the function *binary_transformation*.
- Step 5: The migration procedure of the *i*BABC-CGO algorithm is in charge of transferring food supplies across islands, which starts when a certain number of iterations (*Fm*) are finished. This process is described in Algorithm 1 as shown in Lines 58–71.
- Step 6: Record the best solution in each island (MG_t).
- Step 7: Memorize the best solution (GB) of all islands.
- Step 8: Stop condition. Repeat steps 4, 5, 6, and 7 until the MCN is reached.

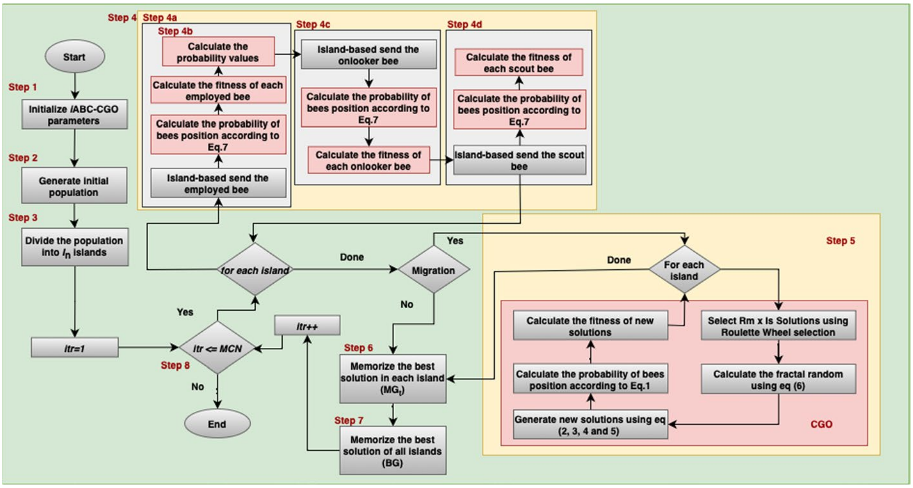

Step 4
Step 4a
Step 4b
Step 4c
Step 4d
Start
Step 1
Initialize iABC-CGO parameters
Step 2
Generate initial population
Step 3
Divide the population into I_n islands
itr=1
Calculate the probability values
Calculate the fitness of each employed bee
Calculate the probability of bees position according to Eq.7
Island-based send the employed bee
Island-based send the onlooker bee
Calculate the probability of bees position according to Eq.7
Calculate the fitness of each onlooker bee
Calculate the fitness of each scout bee
Calculate the probability of bees position according to Eq.7
Island-based send the scout bee
for each island
Done
Migration
Yes
No
Yes
itr <= MCN
Step 8
No
End
itr++
Step 6
Memorize the best solution in each island (MG_t)
Step 7
Memorize the best solution of all islands (BG)
Step 5
Done
For each island
Calculate the fitness of new solutions
Select Rm x Is Solutions using Roulette Wheel selection
Calculate the probability of bees position according to Eq.1
Calculate the fractal random using eq (6)
Generate new solutions using eq (2, 3, 4 and 5)
CGO

Fig. 4 The flowchart of *i*BABC-CGO algorithm

Algorithm 1: Pseudo code of iBABC-CGO algorithm

```
Initialization of $i$ABC and problem parameters // Step 1
Set the ABC and the island parameters $(SN, MCN, limit, I_s, I_n, R_m, F_m)$
Define the fitness function min or max $f_(x), x = (x^1, x^2x, ..., x^{SN})$
Initialize TF to use // TF:transfer function
Construct ABC population based on binary initialization.// Step 2
Calculate fitness values $f(x^j), \forall j = (1, 2, ..., SN)$
Divide $i$ABC into $I_n$ islands // Step 3
$Flag(j) = False, \forall j = (1, 2, ..., SN)$
for $t \leftarrow 1$ to $I_n$ do
    for $i \leftarrow 1$ to $I_s$ do
        $select \quad j, where \quad j \in \{1, 2, ..., SN\}$
        while $Flag(j)$ $is$ $true$ do
            $select \quad j, where \quad j \in \{1, 2, ..., SN\}$
        end
        $Added \quad x^j \quad to \quad island \quad \zeta_t$
        $flag(j) = true$
    end
end
// Step 4
// Step 4a:  Island based send the employed bee
for $t \leftarrow 1$ to $I_n$ do
    for $i \leftarrow 1$ to $I_s$ do
        $select \quad x^k \in \zeta_t where \quad k \neq i$
        $select \quad j \quad where \quad j \in \{1, 2, ..., d\}$
        $x\prime_i^j(\zeta_t) = MG_t + U(0, 1) \times (x_i^j(\zeta_t) - x_k^j(\zeta_t))$
        $x\prime_i^j(\zeta_t) = binarytransformation(TF, x\prime_i^j(\zeta_t))$// Binary transformation
        if $(f(x\prime_i(\zeta_t) < (f(x_i)(\zeta_t))$ then
            $x_i(\zeta_t) = x\prime_i(\zeta_t)$
            $trial_i(\zeta_t) = 0$
        else
            $trial_i(\zeta_t) = trial_j(\zeta_t) + 1$
    end
end
// Step 4b:  Calculate the probability values
for $t \leftarrow 1$ to $I_n$ do
    for $i \leftarrow 1$ to $I_s$ do
        $p_i(\zeta_t) = \frac{f(x_i(\zeta_t))}{\sum_{t=1}^{I_n} \sum_{i=1}^{I_s} f(x_i(\zeta_t))}$
    end
end
// Step 4c:  Island-based send the onlooker bee
$r = rand(0, 1)$
$sum\_prob = 0$
for $t \leftarrow 1$ to $I_n$ do
    for $i \leftarrow 1$ to $I_s$ do
        $sum\_prob = sum\_prob + p_i(\zeta_t)$
        if $sum\_prob \geq r$ then
            break loop
$select\ x^k \in \zeta_t, where \quad k \neq i$
select j, where $i \in \{1, 2, ..., d\}$
$x\prime_i^j(\zeta_t) = MG_t + U(0, 1) \times (x_i^j(\zeta_t) - x_k^j(\zeta_t))$
$x\prime_i^j(\zeta_t) = binarytransformation(TF, x\prime_i^j(\zeta_t)$// Binary transformation
if $(f(x\prime_i(\zeta_t) < (f(x_i(\zeta_t)))$ then
    $x_i(\zeta_t) = x\prime_i(\zeta_t)$
    $trial_i(\zeta_t) = 0$
else
    $trial_i(\zeta_t) = trial_j(\zeta_t) + 1$
// Step 4d:  Island based send the scout bee
if $trial_i(\zeta_t) \geq limit$ then
    $x_i^j(\zeta_t) = x^j(min) + (x^j(max) - x^j(min)) \times U(0, 1), \forall j = 1, 2, ..., d$
    $x_i^j(\zeta_t) = binary\_transformation(TF, x_i^j(\zeta_t))$// Binary transformation
    $Calculate \quad f(x\prime^j(\zeta_t))$
    $trial_j(\zeta_t) = 0$
// Step 5:  Migration Process
if $itr$ mod $F_m = 0$ then
    for $t \leftarrow 1$ to $I_n$ do
        $j = 1$
        while $j \leq R_m \times I_s$ do
            // Apply CGO algorithm using its equations
            $x_s = RouletteWheel(Pop_t)$
            $\mathrm{x}_{new}^1 = x_s + \alpha_s \times (\beta_s \times GB - \gamma_s \times MG_t)$
            $\mathrm{x}_{new}^2 = GB + \alpha_s \times (\beta_s \times x_s - \gamma_i \times MG_t)$
            $\mathrm{x}_{new}^3 = MG_t + \alpha_s \times (\beta_s \times x_s - \gamma_s \times GB)$
            $\mathrm{x}_{new}^4 = x_s \left(x_s^k = x_s^k + R\right)$
            // Binary transformation
            $x_{new}^1 = binary\_transformation(TF, x_{new}^1)$
            $x_{new}^2 = binary\_transformation(TF, x_{new}^2)$
            $x_{new}^3 = binary\_transformation(TF, x_{new}^3)$
            $x_{new}^4 = binary\_transformation(TF, x_{new}^4)$
            Evaluate fitness values of $x_{new}^1$, $x_{new}^2$, $x_{new}^3$ and $x_{new}^4$
// Step 6:  Memorize the best solution in each island
Memorize Best solution $MG_t$ in island $\zeta_t$
// Step 7:  Memorize the best solution of all islands
Memorize Best solution $GB$ of all islands
// Step 8:  Stop condition
while $time \leq MCN$ do
    Repeat step 4 to step 6

Function Binary_transformation($TF$, $X_i$):
    calculate the probability of updating position as Eq. (1).
    if $(TF ==' S')$ then
        Apply Eq. (10)
    else
        Apply Eq. (11)
    End if
    return $X_i$
```

4.3 *i*BABC-CGO complexity

According to the original *i*ABC, the time complexity is $\mathcal{O}(MCN \times I_s \times I_n \times d)$ by neglecting the compute time for calculating the objective function. However, for our proposed approach *i*BABC, the fitness assessment takes more time than other modules and we note it as '*fe*'. In addition, the slight variation between the original *i*ABC and the binary variant *i*BABC lies in the integration of the transfer function (TF) calculation. Overall, the complexity of *i*BABC is $\mathcal{O}(MCN \times I_s \times I_n \times d \times fe \times TF)$. The latter could be simplified by neglecting the part **TF** since the function used in this study is simple and fast to calculate.

However, the migration method for *i*BABC-CGO is the main difference from that of *i*BABC. In this stage, the proposed approach uses the principles of the CGO algorithm by updating a randomly selected solution using roulette wheel selection. This update procedure concerns $R_m * I_s$ solutions on each island as same as the *i*BABC. The used migration mechanism in *i*BABC is the random ring topology which is based on replacing the worst solution of an island t with the best solution in the previous island $t-1$. Since the computational complexity of the newly proposed migration mechanism is slightly more time-consuming than the random ring topology, we decide to consider them equivalent. Therefore, the complexity of *i*BABC-CGO does not differ from the complexity of *i*BABC which is $\mathcal{O}(MCN \times I_s \times I_n \times d \times fe)$.

4.4 *i*BABC for feature selection using gene expression data

As previously stated, FS is a binary optimization problem that involves identifying and evaluating a subset of significant features while maintaining good accuracy. The best relevant subset of features is chosen based on two objectives: the lowest CER and the fewest features, as expressed in the objective function Eq. (12). At the core of the fitness function, each determined solution is evaluated with the KNN machine learning classifier.

$$\downarrow fitness = \alpha\gamma_R(D) + \ (1-\alpha)\frac{|R|}{|C|}, \tag{13}$$

where $\gamma_R(D)$ denotes the CER of the classifier R with respect to the decision D as formulated in Eqs. (13) and (14), $|R|$ defines the length of the selected feature subset, $|C|$ shows the total features count, and $\alpha \in [0, 1]$, $(1-\alpha)$ parameters defines the importance of the classification quality and the length of the subset as adopted from literature (Emary et al. 2016). During this study, we set the value of α to 0.1.

$$\gamma_R(D) = 1 - Acc \tag{14}$$

$$Acc = \frac{C_{\text{num}}}{C_{\text{num}} + I_{\text{num}}} * 100\%, \tag{15}$$

where C_{num} and I_{num} denote the count of correctly and incorrectly classified labels, respectively.

5 Experimental results

The proposed approaches were implemented using $MATLABR2020a$ and run on an Intel Core i7 machine, 2.6 GHz CPU and $16GB$ of RAM. Experiments were repeated 51 independent times to obtain statistically meaningful results. For evaluation purposes, the k–fold (Fushiki 2011) cross-validation method is used to evaluate the consistency of the generated results. The data is partitioned into K equivalent folds, with $K-1$ folds used for training the classifier during the optimization phase and the remaining fold used for testing and validating the classifier. Using different folds as a test set, the method runs K times. The standard form of this technique in wrapper feature selection algorithms is based on the KNN (k-nearest neighbor) classifier with $K=5$ (Friedman et al. 2001), which will be held during this study.

5.1 Datasets description and parameter settings

For evaluating the proposed iBABC-CGO, fifteen biological gene expression datasets are used for experiments detailed in Table 1. The studied datasets express different types of cancerous diseases and contain two different types: binary-class (BC) and multi-class (MC) datasets. For performance comparison, seven metaheuristic feature selection approaches are chosen, namely: original ABC (Rao et al. 2019), Island-based ABC (iABC) (Awadallah et al. 2020). In addition, four recent metaheuristics: Binary Atom Search Optimization (BASO) (Too and Rahim Abdullah 2020), Binary Equilibrium Optimizer (BEO) (Gao et al. 2020), and Binary Henry Gas Solubility Optimization (BHGSO) (Neggaz et al. 2020), are chosen. In the end, another state-of-the-art metaheuristic Binary Genetic Algorithm (BGA) (Babatunde et al. 2014) is selected. These selected comparative metaheuristics have performed well in the past in solving different OPs including the feature selection problem. Table 2 summarizes details of all the comparative studies and models examined during this study.The parameters of these methods for the feature selection problem are detailed in Table 3 and selected according to their original respective papers.

5.2 Results for microarray datasets

Table 4 presents the average performance of three classifiers KNN, SVM, and NB across 15 microarray datasets. The performance is measured in terms of four metrics: Accuracy (Acc), Precision (Pr), Recall (Rc), and ROC (Roc). Across the datasets: The "Lymphoma" dataset shows an impressive performance with the KNN classifier, achieving perfection across all metrics. The "9-tumors" dataset appears challenging for all classifiers, though SVM manages to outperform the others slightly. The "Leukemia-1" dataset highlights the strength of the SVM classifier, with nearly perfect scores across the metrics. In the "Ovarian" dataset, while KNN and SVM offer competitive performances, the NB classifier shines with near-perfect results. For the "SRBCT" dataset, the SVM classifier emerges as the dominant one, showcasing an accuracy close to 95.12%. While each classifier exhibits strengths in specific datasets, SVM consistently demonstrates a robust performance across a broader range.

The presented Table 5, details the classification accuracy of three classifiers (KNN, SVM, and NB) when utilizing filter-based gene selection methods. Each classifier was tested with four different filters: mRMR, CMIM, Chi-square, and Relief-F. The experiment kept the number of selected genes constant at 50 for each filter. The data indicates that no single classifier or filter consistently outperforms others across all datasets; the optimal combination varies depending on the specific dataset. For instance, KNN achieves standout results with the CMIM filter on the "11-tumors" dataset and with mRMR for the "Lymphoma" dataset. For the majority of datasets, KNN's performance is quite variable, but in several cases, it either matches or slightly surpasses the other classifiers.

On the other hand, SVM often shines when paired with the Chi-square filter, as seen in its exemplary performance on the "SRBCT" dataset. The variation in classifier performance emphasizes the importance of method selection tailored to individual datasets. NB achieves perfect scores on both the "Lymphoma" and "SRBCT" datasets when paired with the mRMR and Chi-square filters respectively. In several datasets like "Brain tumor 1", "DLBCL", and "Prostate tumor", NB outperforms both KNN and SVM, illustrating its potential strength when the underlying assumptions of the Naive Bayes classifier hold true for the data. As for the mRMR filter method, it shows consistently competitive performance across all three classifiers. Especially for NB, it offers the best results in a significant number of datasets. While the CMIM filter provides leading results in a few datasets (notably for KNN), its performance is not consistently top-tier across all datasets. On the other hand, Chi-square filter especially when paired with SVM and NB, produces some of the highest accuracy across multiple datasets. Relief-F results are mixed. While it yields the best performance in a few cases, it doesn't consistently outperform the other filters. While SVM paired with the Chi-square filter method often yields strong results, it's clear that no single combination of classifier and filter universally outperforms others across all datasets.

Table 6, displays the classification accuracy obtained from various combinations of classifiers and filter-based gene selection methods, using a consistent set of 100 selected genes. Three classifiers KNN, SVM, and NB are evaluated with four filter methods: mRMR, CMIM, Chi-square, and Relief-F. The results are presented for 15 datasets. A single glance reveals that the best-performing combination varies greatly across datasets, highlighting the non-uniform nature of optimal classifier and filter method pairing. For instance, in the "11-tumors" dataset, KNN combined with the CMIM filter shows a peak performance of 88.99%. Meanwhile, in the "SRBCT" dataset, both SVM and NB, when paired with the Chi-square filter, achieve a perfect accuracy of 100%. The fluctuations in performance underscore the dataset specific efficiency of each combination.

Fig. 5, represents the classification accuracy achieved using a fixed set of genes 50 and 100 across various combinations of classifiers and filter-based gene selection methods. When comparing the two sets, it becomes evident that the number of selected genes and the dataset in use significantly influence the optimal choice of classifier and filter method. Across the two configurations, the KNN, SVM, and NB classifiers are evaluated using four filter methods: mRMR, CMIM, Chi-square, and Relief-F for 15 distinct datasets. Notably, for some datasets, an increase in the number of genes from 50 to 100 leads to noticeable changes in classification accuracy. For example, in the "11-tumors" dataset, when 100 genes are used, KNN combined with the CMIM filter achieves a peak performance of 88.99%. On the other hand, when using 50 genes, different combinations might exhibit optimal performance. The variability in results underscores the importance of selecting an appropriate number of genes, classifier, and filter method combinations, tailored to the specific characteristics of each dataset.

Table 1 Details of datasets with Class Distribution (CD)

Dataset	No of instances	No of features	Classes	Description	CD
9-tumors	60	5726	9 MC	NSCLC, Colon, Breast, Ovary, Leukemia, Renal, Melanoma, Prostate, CNS	[8, 8, 8, 8, 7, 6, 6, 6, 3]
11-tumors	174	12533	11 MC	Ovary, Bladder/ureter, Breast, Colorectal, Gastro-esophagus, Kidney, Liver, Prostate, Pancreas, Adeno Lung, Squamous Lung	[27, 26, 25, 23, 14, 14, 12, 11, 8, 7, 7]
Brain-tumor 1	90	5920	5 MC	Medulloblastoma, Malignant glioma, AT/RT, Normal cerebellum, PNET	[24, 14, 19, 23, 10]
Brain-tumor 2	50	10367	4 MC	Classic glioblastomas, Classic anaplastic oligodendrogliomas, Nonclassic glioblastomas, Nonclassic anaplastic oligodendrogliomas	[14, 6, 14, 16]
Breast	97	24481	2 BC	Relapsed Metastases within 5 years, Non-relapsed Metastasis for 5 years	[46, 51]
CNS	60	7129	2 BC	Survivors, Exitus (Deceased)	[21, 39]
Colon	62	2000	2 BC	Normal, Malignant	[22, 40]
DLBCL	77	5469	2 BC	Diffuse Large B-Cell Lymphoma, Follicular Lymphoma	[58, 19]
Leukemia 1	72	7129	2 BC	ALL, AML	[47, 25]
Lung cancer	203	12600	5 MC	Adenocarcinomas, Small-cell lung carcinomas, Squamous cell carcinomas, Carcinoids, Normal lung tissue	[139, 6, 21, 20, 17]
Lymphoma	66	4026	3 MC	Not specified	[22, 21, 23]
MLL	72	12582	3 MC	Not specified	[24, 24, 24]
Ovarian	253	15154	2 BC	Controls, Ovarian cancer	[91, 162]
Prostate tumor	102	10509	2 BC	Normal, Tumor	[50, 52]
SRBCT	83	2308	4 MC	Ewing’s tumors, Neuroblastoma, Burkitt’s lymphoma, Rhabdomyosarcoma	[29, 18, 11, 25]

Table 2 Key to comparative methods

Type	Abbreviation	Explanation	Reference
Classifiers	KNN	k-Nearest Neighbor classifier	Maleki et al. (2021)
	NB	Naive Bayes classifier	Ahmed et al. (2017)
	SVM	Support Vector Machine	Mukherjee (2003)
Validation approaches	LOOCV	Leave-One-Out Cross-Validation	Qi et al. (2019)
	K-fold CV	K-fold Cross-Validation (k = 5 in this study)	
Filter-based feature selection methods	CMIM	Conditional Mutual Information Maximization criterion	Fleuret (2004)
	Relief-F	==========	Robnik-Šikonja and Kononenko (2003)
	Chi-square	==========	Jin et al. (2006)
	mRMR	Maximum Relevance Minimum Redundancy algorithm	Alomari et al. (2017)
Hybrid-based feature selection methods	IG-MBKH	Information Gain and a Modified Binary Krill Herd Algorithm	Zhang et al. (2020)
	rMRMR-MGWO	Robust Maximum Relevance Minimum Redundancy filter and Modified Gray Wolf Optimizer with TRIZ-inspired operators wrapper	Alomari et al. (2021)
	IDGA-F-SVM	Intelligent Dynamic Genetic Algorithm with Fisher using SVM classifier	Dashtban and Balafar (2017)
	IWS^2-MB	Wrapper-based Sequential Forward Selection with Markov Blanket	Wang et al. (2017)
	TLBOSA-SVM	Teaching Learning-Based Optimization, Simulated Annealing, and Support Vector Machine	Shukla et al. (2019)
	F-Score-IDGA-FSVM	Fisher score filter, Intelligent Dynamic Genetic Algorithm (IDGA) wrapper, and Support Vector Machine classifier	Dashtban and Balafar (2017)
	VLPSO-LS-KNN	Variable-Length Particle Swarm Optimization with Local Search and k-nearest neighbor classifier	Tran et al. (2018)
	BCO-KNN	Bacterial Colony Optimization and k-nearest neighbor classifier	Wang et al. (2017)
	PS-NSGA	Problem-Specific Non-dominated Sorting Genetic Algorithm and k-nearest neighbor classifier	Zhou et al. (2021)
	BChOA-KNN	Binary Chimp Optimization Algorithm using KNN	Pashaei and Pashaei (2022)
	BChOA-C-KNN	Binary Chimp Optimization Algorithm with Crossover operator	Pashaei and Pashaei (2022)

Table 2 (continued)

Type	Abbreviation	Explanation	Reference
	CDNC-SVM	Community Detection with Node Centrality using support Vector machine classifier	Rostami et al. (2022)
	PSO-ensemble	Particle Swarm Optimization and an ensemble learning method	Alrefai and Ibrahim (2022)
	BCOOT-CSA	Binary COOT optimization with crossover schema and simulated annealing	Pashaei and Pashaei (2023)
	BAOAC-SA	Binary Arithmetic Optimization Algorithm with crossover schema and simulated annealing	Pashaei and Pashaei (2022)
	CFC-FBBA	Correlation-based Feature Clustering with Fractional-order Binary Bat Algorithm	Esfandiari et al. (2023)

Table 3 Algorithms parameter settings

Algorithm	Parameter	Value
iABC	Limit	0.1*SN*d
	In	5
	Fm	5
	Rm	20%
ABC	limit	0.1*SN*d
BASO	α(depth weight)	50
	β(multiplier weight)	0.2
	Vmax (maximum velocity)	6
BEO	α	1
	β	2
BGA	Crossover rate	0.9
	Mutation rate	0.1
BHGSO	M_1	0.1
	M_2	0.2
	α, β	1
	K	1
All of them	Search agents (atoms, bees, particles,...)	30
	Maximum iterations	100

Table 4 Percentage of average performance using KNN, SVM, and NB classifiers on 15 microarray datasets

Dataset	KNN				SVM				NB			
	Acc	**Pr**	**Rc**	**Roc**	**ACC**	**Pr**	**Rc**	**Roc**	**ACC**	**Pr**	**Rc**	**Roc**
11-tumors	71,10	78,09	62,28	0,99	**89,02**	**90,88**	**83,79**	**1**	74,57	67,41	60,27	**1**
9-tumors	37,29	29,99	32,94	0,97	**49,15**	**55,55**	**43,45**	**1**	35,59	21,66	29,56	0,51
Brain tumor 1	77,53	48,86	40,00	0,98	**83,15**	**55,95**	**50,00**	**1**	80,90	55,53	46,00	1
Brain tumor 2	**73,47**	71,07	**69,33**	0,96	**73,47**	80,86	68,33	0,99	**73,47**	**81,89**	66,19	**1,00**
Breast	**62,50**	**63,36**	61,31	0,82	52,08	51,00	50,85	0,61	**62,50**	62,58	**61,70**	**0,92**
CNS	59,32	57,52	58,27	0,77	62,71	**59,21**	**59,62**	**0,88**	**66,10**	33,05	50,00	0,55
Colon	**80,33**	**84,61**	**73,72**	0,93	59,02	61,32	62,00	0,90	78,69	80,85	72,44	**0,98**
DLBCL	**84,21**	78,84	**82,46**	0,97	76,32	68,92	70,18	0,94	76,32	**88,00**	52,63	**1**
Leukemia 1	80,28	80,99	74,74	0,98	**98,59**	**98,94**	**98,00**	**1**	87,32	91,82	82,00	**1**
Lung cancer	**89,60**	73,16	67,32	0,99	88,61	87,39	73,59	0,96	89,11	73,50	66,06	**1**
Lymphoma	**100**	**100**	**100**	**1**	90,77	96,08	79,12	**1**	95,38	97,92	88,89	**1**
MLL	90,14	91,36	90,37	0,99	**95,77**	**95,59**	**95,95**	**1**	88,73	88,46	88,01	**1**
Ovarian	92,06	92,84	89,88	0,99	90,87	89,87	90,43	0,95	**98,41**	**98,27**	**98,27**	**1**
Prostate tumor	82,18	82,68	81,99	0,95	61,39	63,38	60,73	0,66	**87,13**	**87,15**	**87,09**	**0,98**
SRBCT	73,17	74,20	77,65	0,99	**95,12**	**96,48**	**93,56**	**1**	91,46	93,38	91,27	**1**

Bold indicates the best values in the table

Table 5 Classification accuracy using selected genes by filter-based gene selection methods with a fixed set of 50 selected genes

Classifier	KNN				SVM				NB			
Filter	mRMR	CMIM	Chi-square	Relief-F	mRMR	CMIM	Chi-square	Relief-F	mRMR	CMIM	Chi-square	Relief-F
11-tumors	71,06	**88,99**	74,57	53,83	51,93	57,80	73,97	71,06	70,50	63,60	72,25	67,61
9-tumors	37,42	49,09	35,61	30,61	33,94	28,79	49,09	56,21	**59,39**	32,12	28,94	20,00
Brain tumor 1	77,52	83,14	80,85	74,05	72,68	70,72	**86,47**	85,49	**86,47**	78,56	76,27	77,45
Brain tumor 2	73,33	73,78	73,56	59,56	61,33	51,11	**79,56**	75,33	77,56	69,56	67,33	63,11
Breast	62,42	51,95	62,53	61,53	53,05	62,42	61,53	52,05	**64,68**	56,05	51,11	62,47
CNS	58,94	61,97	66,06	64,39	57,42	66,06	67,42	**79,70**	79,39	60,45	48,94	66,06
Colon	80,13	59,10	78,72	80,51	63,85	81,92	85,13	83,46	**86,67**	72,05	57,56	75,51
DLBCL	84,17	76,00	76,33	77,42	74,75	72,33	93,50	92,00	**96,00**	68,33	76,17	75,00
Leukemia 1	80,38	**98,57**	87,24	67,62	59,33	69,14	94,29	94,29	97,14	77,33	88,67	82,95
Lung cancer	89,57	88,56	89,07	70,27	54,83	70,29	93,04	72,76	**93,51**	78,71	75,21	79,71
Lymphoma	100	90,77	95,38	78,46	81,54	75,38	98,46	95,38	**100**	95,38	93,85	95,38
MLL	90,10	95,71	88,67	76,19	66,10	69,33	92,86	**95,71**	**95,71**	78,86	70,38	80,10
Ovarian	92,06	90,89	**98,42**	71,84	58,45	73,40	**97,61**	96,02	97,21	88,49	80,99	93,65
Prostate tumor	82,19	61,19	87,05	70,24	60,19	67,19	90,05	91,00	**93,00**	79,24	70,19	85,10
SRBCT	73,31	95,15	91,32	59,63	68,53	71,99	**100**	98,82	**100**	72,06	69,34	74,41

Bold indicates the best values in the table

5.3 Evaluation of the proposed *i*BABC-CGO using different classifiers

In this part, we run a comprehensive series of simulations to identify the optimal classifier to pair with our Feature Selection (FS) algorithm. By conducting these rigorous simulations, we aim to ensure that the chosen classifier synergizes well with our FS approach, thus maximizing the overall effectiveness of the solution. This process not only underscores our commitment to a methodical research approach but also emphasizes our pursuit to ensure that each component of our solution - from the FS algorithm to the classifier - is optimized for peak performance. However, it is important to bear in mind that once the optimal subset has been determined, the performance of the *i*BABC-CGO on biomedical datasets is assessed using leave-one-out cross-validation (LOOCV).

Therefore, in an effort to find the best classifier for our proposed approach, Table 7, represents the performance results in terms of statistical metrics i.e. Best and mean values for various datasets using the proposed *i*BABC-CGO with different classifiers: KNN, SVM, and NB. When examining the KNN classifier, for datasets such as "11-tumors" and "9-tumors", the *i*BABC-CGO-V variant seems to have an edge in terms of both best and mean values. However, for other datasets like "Brain tumor 1", "Brain tumor 2", and "CNS", the *i*BABC-CGO-S version performs slightly better in terms of best values. In the case of SVM, the *i*BABC-CGO-S consistently outperforms the *i*BABC-CGO-V across all datasets for the best metric values, though for mean metric values, there's a mixed performance between the two versions. For the Naive Bayes classifier, the *i*BABC-CGO-V version holds better best values for several datasets, including "Brain tumor 1" and "CNS". Interestingly, for the "Ovarian" dataset, the performance remains the same across all classifiers and versions, suggesting a consistent result for that particular dataset. It is clear from the statistical results that among KNN, SVM, and NB classifiers, the SVM classifier is the best classifier to be combined with the proposed *i*BABC-CGO-V variant.

Table 8, presents the average accuracy of the proposed *i*BABC-CGO method using various classifiers: KNN, SVM, and NB. Upon a cursory examination, the SVM classifier with the *i*BABC-CGO-V version displays the highest accuracy for several datasets, achieving a perfect score of 100% on multiple occasions. Both the "Lymphoma" and "Ovarian" datasets achieved consistent 100% accuracy across all classifiers and versions, indicating that the model fits them perfectly. For datasets like "11-tumors" and "9-tumors", the KNN classifier using *i*BABC-CGO-S has higher accuracy compared to its *i*BABC-CGO-V counterpart. The Naive Bayes classifier, indicated as "NB", also exhibits competitive performance, especially with the *i*BABC-CGO-S version.

Table 9, showcases the average number of selected features for various datasets using the proposed *i*BABC-CGO method paired with different classifiers: KNN, SVM, and NB. A glance at the data reveals a distinct pattern. The SVM classifier with the *i*BABC-CGO-V version often tends to select a larger number of features compared to the *i*BABC-CGO-S variant, as seen in the "9-tumors" and "Breast" datasets. The minimal differences in feature selection numbers across classifiers for datasets like "Brain tumor 2", "DLBCL", and "Lymphoma" indicate a level of consistency in the importance of certain features across different classification methods. The "Lymphoma" dataset stands out with a consistent selection of 2 features across all classifiers and versions, suggesting that a minimal set of features is significant for classification in this particular dataset.

Table 6 Classification accuracy using selected genes by filter-based gene selection methods with a fixed set of 100 selected genes

Classifier	KNN				SVM				NB			
Filter	mRMR	CMIM	Chi-square	Relief-F	mRMR	CMIM	Chi-square	Relief-F	mRMR	CMIM	Chi-square	Relief-F
11-tumors	71,06	**88,99**	74,57	61,83	66,39	64,20	77,45	83,23	79,16	65,33	77,45	73,41
9-tumors	37,42	49,09	35,61	32,58	35,76	25,45	44,09	59,24	**68,03**	25,45	28,94	28,79
Brain tumor 1	77,52	83,14	80,85	74,05	77,39	74,12	83,14	84,38	**88,76**	76,34	77,58	77,45
Brain tumor 2	73,33	73,78	73,56	51,33	63,11	60,89	**79,56**	75,56	77,56	67,11	67,56	69,33
Breast	62,42	51,95	62,53	64,63	56,11	**66,63**	64,53	51,05	**66,74**	57,26	53,05	63,42
CNS	58,94	61,97	66,06	59,24	54,09	66,06	65,91	81,36	**82,88**	65,91	55,76	66,06
Colon	80,13	59,10	78,72	78,59	60,90	81,79	85,26	81,92	**88,33**	80,26	55,90	80,38
DLBCL	84,17	76,00	76,33	81,58	73,50	77,58	93,33	92,00	**96,00**	85,50	77,33	88,25
Leukemia 1	80,38	**98,57**	87,24	71,71	69,24	76,00	97,14	95,71	97,14	78,86	91,43	80,29
Lung cancer	89,57	88,56	89,07	72,73	59,30	73,79	91,55	84,65	**94,04**	84,61	75,17	79,22
Lymphoma	**100**	90,77	95,38	92,31	84,62	92,31	**100**	95,38	**100**	96,92	92,31	98,46
MLL	90,10	**95,71**	88,67	65,05	64,67	77,81	94,29	94,29	**95,71**	85,90	83,14	81,52
Ovarian	92,06	90,89	**98,42**	72,65	59,24	74,60	**97,61**	96,03	**98,40**	92,46	76,60	96,03
Prostate tumor	82,19	61,19	87,05	71,33	61,19	66,24	89,10	91,00	**93,00**	74,33	65,19	83,10
SRBCT	73,31	95,15	91,32	69,56	67,21	79,19	**100**	**100**	**100**	75,66	79,19	83,16

Bold indicates the best values in the table

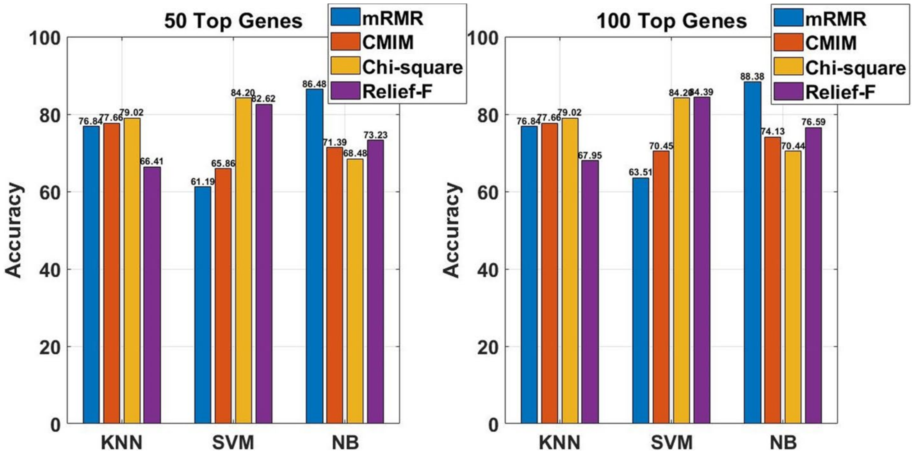


Fig. 5 Comparing the Mean Classification Accuracy of KNN, NB, and SVM Classifiers for Top 50 and Top 100 Genes Selected via Various Filter-Based Feature Selection Algorithms

5.4 Statistical analysis compared to other NIOAs

In the remaining parts of the experiments, we have fixed the SVM classifier (since it provided the best results among other classifiers) to be used in combination with all the metaheuristics that are to be tested. The results for comparison are based on statistical tests in terms of *Best*, *Mean* and *Standard deviation (Std)* fitness values, which represent the minimum, average, and standard deviation, respectively of the 51 independent readings collected from every FS experiment. The results of these parameters are displayed in Table 10, which clearly indicate that the proposed *i*BABC-CGO-V approach produced minimum average fitness values by performing much better than the traditional ABC, *i*ABC, *i*BABC-S, *i*BABC-V, and other binary metaheuristic approaches on 8 out of 15 datasets. In the rest of the datasets, the suggested *i*BABC-CGO-V also performs competitively well against the comparative algorithms, where only the *i*BABC-V algorithm shows better results. Results regarding the Average classification accuracy (*CLACC*) and the reduction of the selected number of features (*SNFs*) values are also shown in Tables 11 and 13 in the next sub-section.

5.4.1 Analysis using average accuracy and number of selected features

To further investigate the efficacy of the suggested approach, a classification accuracy (CLACC) comparison is outlined in Table 11. The highest value of this metric is desired as it signifies the ability of the suggested approach to classify the features accurately. The introduced *i*BABC-CGO-V method attained the best classification accuracy on 43.5% datasets (i.e. 7 out of 15 datasets shown in boldface) as compared to the other tested FS approaches. Whereas, for the rest of the 9 datasets, the *CLACC* results of the suggested method are very competitive in comparison to the other metaheuristics. Moreover, the proposed *i*BABC-CGO-V variant also obtained the highest overall classification accuracy of all datasets as depicted in Fig. 6, which demonstrates the strength of the suggested approach in classifying selected features accurately.

Table 7 Statistical results of the proposed iBABC-CGO using different classifiers (KNN, SVM and NB)

Dataset	**Metric**	KNN		SVM		NB	
		iBABC-CGO-S	**iBABC-CGO-V**	**iBABC-CGO-S**	**iBABC-CGO-V**	**iBABC-CGO-S**	**iBABC-CGO-V**
11-tumors	Best	0,0944	0,1978	**0,0837**	0,0890	0,1152	0,1562
	Mean	0,1027	0,2041	**0,0911**	0,1066	0,1316	0,1729
9-tumors	Best	0,3678	0,4579	**0,2605**	0,3060	0,3526	0,3665
	Mean	0,3916	0,4915	**0,3337**	0,3458	0,3641	0,4304
Brain tumor 1	Best	0,1123	0,1315	0,0821	0,0723	0,0918	**0,0609**
	Mean	0,1144	0,1356	0,1083	**0,0902**	0,0920	0,1176
Brain tumor 2	Best	0,1290	0,1470	0,0919	**0,0735**	0,1107	0,1103
	Mean	0,1328	0,1470	**0,1102**	0,1249	0,1108	0,1434
Breast	Best	0,1781	0,1969	**0,1407**	0,1500	0,1690	0,1500
	Mean	0,1800	0,1969	**0,1689**	0,1725	0,1728	0,1856
CNS	Best	0,1841	0,2138	0,1379	**0,0917**	0,1381	0,1527
	Mean	0,1962	0,2229	0,1626	**0,1137**	0,1385	0,1710
Colon	Best	0,1206	0,1477	0,0903	**0,0887**	0,1064	0,1039
	Mean	0,1217	0,1654	0,1154	**0,1013**	0,1176	0,1600
DLBCL	Best	0,0358	0,0595	**0,0010**	**0,0010**	0,0018	0,0239
	Mean	0,0458	0,0689	0,0180	0,0107	**0,0084**	0,0500
Leukemia 1	Best	0,0133	0,0381	**4,21E-05**	**4,21E-05**	0,0133	0,0127
	Mean	0,0136	0,0381	0,0178	0,0254	**0,0134**	0,0305
Lung cancer	Best	0,0541	0,0535	0,0405	**0,0359**	0,0452	0,0402
	Mean	0,0575	0,0562	0,0513	**0,0371**	0,0461	0,0446
Lymphoma	Best	0,0009	0,0005	**4,97E-05**	**4,97E-05**	0,0006	**4,97E-05**
	Mean	0,0011	0,0086	0,0056	0,0111	**0,0007**	0,0111
MLL	Best	0,0127	0,0254	0,0127	**0,0004**	0,0133	0,0127
	Mean	0,0152	0,0254	0,0127	**0,0081**	0,0183	0,0152
Ovarian	Best	**1,98E-05**	**1,98E-05**	**1,98E-05**	**1,98E-05**	**1,98E-05**	**1,98E-05**
	Mean	**1,98E-05**	**1,98E-05**	**1,98E-05**	**1,98E-05**	**1,98E-05**	**1,98E-05**
Prostate tumor	Best	0,0633	0,0624	0,0535	**0,0268**	0,0542	0,0535
	Mean	0,0702	0,0767	0,0624	**0,0517**	0,0542	0,0553
SRBCT	Best	0,0134	0,0660	0,0014	**0,0009**	0,0014	0,0222
	Mean	0,0138	0,0683	0,0052	**0,0016**	0,0023	0,0420

Bold indicates the best values in the table

Table 13 presents the comparative results of the suggested method in terms of the average number of selected features (SNFs) with the other algorithms. The minimum value of this performance parameter is desired. As can be seen, the suggested approach provides better results in terms of the minimum SNFs for nearly 31% datasets (i.e. 5 out of 15 datasets) against the compared algorithms and for the rest of the datasets, the standard binary version *i*BABC-V performed better than all tested metaheuristics. It should be noted that except *i*BABC-CGO-V and *i*BABC-V, no other comparative algorithm performed well in providing good SNFs results. However, for all the minimum SNFs outcomes provided by

Table 8 Average accuracy of the proposed iBABC-CGO using different classifiers (KNN, SVM and NB)

Dataset	KNN		SVM		NB	
	iBABC-CGO-S	**iBABC-CGO-V**	**iBABC-CGO-S**	**iBABC-CGO-V**	**iBABC-CGO-S**	**iBABC-CGO-V**
11-tumors	88,67	77,34	90,75	**100**	85,90	80,81
9-tumors	56,61	45,42	81,19	**96,92**	59,66	52,88
Brain tumor 1	87,42	84,94	91,01	**100**	89,89	86,97
Brain tumor 2	85,31	83,67	89,8	**93,88**	87,76	85,31
Breast	80,00	78,13	84,38	**90,14**	80,83	79,79
CNS	78,31	75,25	84,75	**89,83**	84,75	81,02
Colon	86,89	82,62	92,36	**96,16**	87,21	85,57
DLBCL	96,05	92,37	**100**	**100**	99,21	95,26
Leukemia 1	98,59	95,77	**100**	**100**	98,59	96,90
Lung cancer	93,66	93,76	95,54	**96,04**	94,95	95,05
Lymphoma	**100,00**	99,08	**100**	**100**	**100,00**	**100,00**
MLL	98,31	97,18	98,59	**100**	98,03	98,31
Ovarian	**100,00**	**100,00**	**100**	**100**	**100,00**	**100,00**
Prostate tumor	92,28	91,49	94,06	**97,03**	94,06	93,86
SRBCT	98,78	92,44	**100**	**100**	**100,00**	95,37

Bold indicates the best values in the table

Table 9 Average selected features of the proposed iBABC-CGO using different classifiers (KNN, SVM and NB)

Dataset	KNN		SVM		NB	
	iBABC-CGO-S	**iBABC-CGO-V**	**iBABC-CGO-S**	**iBABC-CGO-V**	**iBABC-CGO-S**	**iBABC-CGO-V**
11-tumors	16,8	**13,4**	**18,8**	21,2	15	15,8
9-tumors	10,6	10,2	**9**	24	11,2	9,8
Brain tumor 1	4,4	4,2	**4,8**	8,8	**3,8**	4,4
Brain tumor 2	3,2	**2,8**	**3,2**	3,4	3	3,4
Breast	4,2	3,6	**4,8**	24,6	4,6	4,4
CNS	3	3,2	**2,6**	7.6	3,4	**2,6**
Colon	2,8	2,6	**2,2**	**2,2**	3	2,6
DLBCL	2,6	2,4	2,4	**2**	2,2	2,2
Leukemia 1	**2**	2,4	**2,4**	2,4	2,2	2,4
Lung cancer	8,4	**7,2**	**7,2**	7,6	**7,2**	7,6
Lymphoma	**2**	**2**	**2**	**2**	**2**	**2**
MLL	2,4	**2**	**2**	2,2	2,4	**2**
Ovarian	2,8	2,6	**2,2**	3	3	2,8
Prostate tumor	2,4	**2**	2,8	2,4	2,8	2,8
SRBCT	**4,6**	5	**4,6**	4,8	5,2	5

Bold indicates the best values in the table

the *i*BABC-V are not supported by good accuracy values, as it manages to provide good accuracy in only 18.7% datasets. Therefore, the overall findings confirm that the proposed *i*BABC-CGO-V method has the best ability to maintain an adequate exploration-exploitation ratio during the optimization search process, which resulted in providing good average CLACC and competitive average SNFs outcomes. Moreover, the proposed *i*BABC-CGO-S variant scored the minimum number of selected features overall all algorithms on the average of 15 datasets, as depicted in Fig. 7, which demonstrates the capacity of the suggested approach in reducing the number of selected features.

Table 14 contains the Wilcoxon rank sum (WRS) test, a non-parametric statistical test to examine whether the suggested method's outcomes are statistically distinguishable from those of other algorithms (Rey and Neuhäuser 2011). WRS test produces a p-value to compare the significance levels of the two methods. P-values below 0.05 imply the existence of a statistically significant difference at the 5% confidence level between *i*BABC-CGO and the other compared method.

The p-values in Table 14 indicate that the suggested *i*BABC-CGO-V technique produces significantly different outcomes than most existing algorithms on most gene expression datasets with 95% confidence level except for the binary *i*BABC-V. In order to further strengthen the evidence of the superiority of our suggested *i*BABC-CGO-V method over the compared algorithms, we conducted statistical tests such as the Friedman test (whose lower rank is desired), Friedman align test (higher rank is desirable) and Quade test (lower rank is considered best) on the average CLACC results to rank the performance of each FS algorithm. The collected outcomes as shown in Table 15, clearly indicate that the suggested *i*BABC-CGO-V obtained the best ranks among tested algorithms for all the tests followed by the *i*BABC-V approach.

5.4.2 Box plot and radar plot analysis

The boxplot analysis is the best way to represent the data distribution characteristics of collected results and to find out data anomalies such as skewness and outliers. Boxplots show the data distributions in the form of different quartiles namely the lower (lowest point/edge of the whisker) and upper (highest point/edge of the whisker) quartiles, which represent the minimum and maximum values of the data distribution. The lower and higher quartiles are shown by the corners of the rectangles. A small boxplot rectangle represents strong data concordance. Figures 8 and 14 show the box plots generated from the results of different algorithms for different medical datasets. The boxplots of the proposed *i*BABC-CGO-V, for most datasets are very narrow as compared to other algorithm distributions. Indeed, the proposed *i*BABC-CGO-V method is superior to the other algorithms on the bulk of the investigated datasets.

Figure 9 presents the radar chart that ranks the algorithms based on their average best and average mean fitness results. Levels near the centre of the radar graph represent higher best fitness and average fitness values. Therefore, a resilient algorithm has a narrow area, which is the proposed *i*BABC-CGO-V approach at first, followed by the *i*BABC-V algorithm. The performance of the tested FS methods is compared in Tables 12, 13 and the radar plot in Fig. 9, from which it can be deduced that the suggested method is superior to the established methods.

Table 10 Comparison of *i*BABC-CGO versus other metaheuristic algorithms in terms of fitness values using SVM classifier

Dataset	BASO	BEO	BGA	BHGSO	ABC	iABC	iBABC-S	iBABC-V	iBABC-CGO-S	iBABC-CGO-V
11-tumors										
Best	0,1204	0,0941	0,0996	0,1515	0,1152	0,0944	0,1203	0,1098	0,0837	**0,0008**
Mean	0,1646	0,1302	0,1343	0,1811	0,1442	0,1276	0,1589	0,1398	0,0911	**0,0118**
Std	0,0233	0,0223	0,0231	0,0202	0,0193	0,0232	0,0253	0,0191	0,0047	**0,0061**
9-tumors										
Best	0,3981	0,3826	0,3819	0,4434	0,4122	0,4136	0,3822	0,3520	0,2605	**0,0285**
Mean	0,4629	0,4495	0,4407	0,4876	0,4575	0,4876	0,4572	0,4249	0,3337	**0,0394**
Std	0,0297	0,0463	0,0393	0,0262	0,0333	0,0335	0,0355	0,0362	0,0449	**0,0061**
Brain tumor 1										
Best	0,1329	0,1118	0,1121	0,1330	0,1222	0,1121	0,1430	0,1126	0,0821	**0,0723**
Mean	0,1648	0,1332	0,1275	0,1652	0,1397	0,1405	0,1635	0,1383	0,1083	**0,0902**
Std	0,0134	0,0132	**0,0121**	0,0129	0,0124	0,0172	0,0124	0,0126	0,0167	0,0129
Brain tumor 2										
Best	0,1293	0,1663	0,1476	0,1476	0,1659	0,1474	0,1474	0,1473	0,0919	**0,0735**
Mean	0,1790	0,2281	0,2104	0,2032	0,2249	0,2078	0,1731	0,1993	**0,1102**	0,1249
Std	0,0222	0,0274	0,0385	0,0361	0,0265	0,0404	**0,0182**	0,0224	0,0184	0,0302
Breast										
Best	0,1784	0,1783	0,1690	0,1785	0,3565	0,2066	0,1785	0,1782	0,1407	**0,1143**
Mean	0,2122	0,2028	0,1948	0,2882	0,3743	0,2715	0,2058	0,1995	0,1689	**0,1726**
Std	0,0214	0,0149	0,0107	0,0395	**0,0089**	0,0321	0,0157	0,0139	0,0162	0,0479
CNS										
Best	0,1993	0,2448	0,1689	0,1844	0,1997	0,1688	0,1996	0,2298	0,1379	**0,0917**
Mean	0,2350	0,2770	0,2415	0,2481	0,2656	0,2291	0,2356	0,2884	0,1626	**0,1137**
Std	**0,0166**	0,0226	0,0356	0,0253	0,0293	0,0314	0,0190	0,0264	0,0231	0,0206

Table 10 (continued)

Dataset	BASO	BEO	BGA	BHGSO	ABC	iABC	iBABC-S	iBABC-V	iBABC-CGO-S	iBABC-CGO-V
Colon										
Best	0,1206	0,2961	0,1224	0,1225	0,2812	0,1210	0,1210	0,2806	0,0903	**0,0887**
Mean	0,1648	0,3140	0,1557	0,1703	0,3190	0,1547	0,1616	0,3202	0,1154	**0,1013**
Std	0,0204	0,0106	0,0193	0,0264	0,0222	0,0264	0,0237	0,0187	0,0177	**0,0071**
DLBCL										
Best	0,0367	0,1077	0,0358	0,0366	0,1197	0,0366	0,0252	0,1074	**0,0010**	**0,0010**
Mean	0,0642	0,1318	0,0585	0,0658	0,1490	0,0522	0,0574	0,1272	0,0180	**0,0107**
Std	0,0167	0,0137	0,0183	0,0197	0,0175	0,0095	0,0171	0,0120	0,0109	**0,0054**
Leukemia 1										
Best	0,0257	0,0131	0,0135	0,0259	0,0009	0,0133	0,0140	0,0135	**4,21E-05**	**4,21E-05**
Mean	0,0384	0,0244	0,0190	0,0443	0,0293	0,0217	0,0385	0,0312	**0,0178**	0,0254
Std	0,0106	0,0101	**0,0061**	0,0135	0,0155	0,0076	0,0120	0,0104	0,0113	0,0155
Lung cancer										
Best	0,0541	0,0717	0,0497	0,0498	0,0676	0,0542	0,0542	0,0765	0,0405	**0,0359**
Mean	0,0675	0,1058	0,0670	0,0729	0,0904	0,0649	0,0668	0,1096	0,0513	**0,0371**
Std	0,0074	0,0204	0,0100	0,0104	0,0107	0,0058	0,0077	0,0156	0,0060	**0,0020**
Lymphoma										
Best	0,0007	0,0010	0,0012	0,0009	0,0022	0,0010	0,0010	0,0021	**4,97E-05**	**4,97E-05**
Mean	0,0063	0,0172	**0,0036**	0,0054	0,0207	0,0046	0,0051	0,0193	0,0056	0,0111
Std	0,0065	0,0081	**0,0046**	0,0059	0,0086	0,0054	0,0058	0,0088	0,0076	0,0062
MLL										
Best	0,0259	0,0260	0,0132	0,0386	0,0260	0,0258	0,0509	0,0260	0,0127	**0,0004**
Mean	0,0630	0,0571	0,0425	0,0671	0,0525	0,0476	0,0621	0,0571	0,0127	**0,0081**
Std	0,0188	0,0148	0,0150	0,0163	0,0130	0,0153	0,0117	0,0160	**1,94E-18**	0,0069

Table 10 (continued)

Dataset	BASO	BEO	BGA	BHGSO	ABC	iABC	iBABC-S	iBABC-V	iBABC-CGO-S	iBABC-CGO-V
Ovarian										
Best	0,0145	0,0506	0,0040	0,0182	0,0431	0,0075	0,0182	0,0466	**1,98E-05**	**1,98E-05**
Mean	0,0275	0,0764	0,0118	0,0264	0,0616	0,0167	0,0260	0,0731	**1,98E-05**	**1,98E-05**
Std	0,0058	0,0137	0,0043	0,0045	0,0115	0,0055	0,0048	0,0128	**0**	**0**
Prostate tumor										
Best	0,0721	0,1077	0,0631	0,0900	0,2059	0,0891	0,0721	0,1161	0,0535	**0,0268**
Mean	0,0999	0,1689	0,0853	0,1206	0,2622	0,1164	0,1015	0,1691	0,0624	**0,0517**
Std	0,0173	0,0288	0,0116	0,0186	0,0246	0,0155	0,0166	0,0240	**0,0063**	0,0159
SRBCT										
Best	0,0258	0,0242	0,0146	0,0363	0,0254	0,0149	0,0367	0,0153	0,0014	**0,0009**
Mean	0,0805	0,0651	0,0415	0,0885	0,0632	0,0408	0,0887	0,0530	0,0052	**0,0016**
Std	0,0314	0,0257	0,0180	0,0281	0,0236	0,0208	0,0319	0,0261	0,0049	**0,0006**

Bold indicates the best values in the table

Table 11 Comparison of *i*BABC-CGO versus other metaheuristic algorithms for average classification accuracy using SVM classifier

Dataset	BASO	BEO	BGA	BHGSO	ABC	iABC	iBABC-S	iBABC-V	iBABC-CGO-S	iBABC-CGO-V
11-tumors	81,77	85,59	85,14	79,93	84,04	85,88	82,41	84,52	90,75	**100**
9-tumors	48,67	50,18	51,15	45,94	49,27	45,94	49,33	52,91	81,19	**96,92**
Brain tumor 1	81,78	85,31	85,96	81,74	84,59	84,51	81,94	84,75	91,01	**100**
Brain tumor 2	80,17	74,71	76,68	77,48	75,07	76,97	80,83	77,92	89,8	**93,88**
Breast	76,45	77,49	78,39	68,01	58,44	69,87	77,16	77,86	84,38	**90,14**
CNS	73,97	69,31	73,24	72,52	70,58	74,64	73,91	68,04	84,75	**89,83**
Colon	82,03	65,46	83,02	81,44	64,87	83,14	82,44	64,75	92,36	**96,16**
DLBCL	93,00	85,48	93,61	92,81	83,55	94,31	93,75	86,00	**100**	**100**
Leukemia 1	95,82	97,38	97,99	95,17	96,83	97,69	95,82	96,63	**100**	**100**
Lung cancer	92,56	88,30	92,61	91,96	90,01	92,84	92,64	87,87	95,54	**96,04**
Lymphoma	99,45	98,24	99,78	99,56	97,86	99,67	99,62	98,02	**100**	**100**
MLL	93,06	93,71	95,32	92,61	94,22	94,77	93,16	93,71	98,59	**100**
Ovarian	96,98	91,55	98,74	97,11	93,20	98,19	97,15	91,92	**100**	**100**
Prostate tumor	88,97	81,29	90,59	86,67	70,93	87,13	88,79	81,26	94,06	**97,03**
SRBCT	91,33	93,07	95,69	90,51	93,29	95,78	90,46	94,43	**100**	**100**

Bold indicates the best values in the table

5.4.3 Convergence analysis

Additionally, Figs. 10 and 15 demonstrate that for practically all sixteen datasets, the convergence values of *i*BABC-CGO-V considerably increase after a few iterations. This tendency suggests that *i*BABC-CGO-V requires identifying only a few genes from high-dimensional data to improve classification performance. As can be seen from Figs. 10 and 15, *i*BABC-CGO-V obtained a better convergence rate followed by the *i*BABC-V algorithm.

5.5 Statistical analysis compared to other state-of-art approaches

In this section, we perform a rigorous comparison of the performance of the proposed method with other state-of-the-art approaches. The datasets cover a variety of biological conditions and reflect the diverse applications of these algorithms. Results showing accuracy percentages and their variability provide an overview of the robustness and consistency of each method. With these results, we aim to highlight not only a method's accuracy, but also its reliability over multiple iterations.

In the presented Table 16, various algorithms are compared for their performance on several datasets, with the iBABC algorithm standing out in particular. The iBABC-CGO-S variant demonstrates commendable accuracy, particularly when considering levels such as 90.75% for dataset 11 tumors and 91.01% for brain tumor 1. However, the iBABC-CGO-V variant stands out for its impressive performance, achieving 100% accuracy on datasets 11 tumors and brain tumor 1. While the iBABC algorithm shines in many scenarios, it's pertinent to note that other methods also have their moments of excellence. For example, the BAOAC-SA algorithm achieved almost 97% on the 11-tumors and Brain tumor 1 datasets, while BChOA-C-KNN presented levels in excess of 95% in the datasets on which it was evaluated. However, even among these high-performance algorithms, the iBABC-CGO-V variant stands out for its unrivalled accuracy across multiple datasets. This analysis therefore underlines the superiority of the iBABC algorithm, particularly its CGO-V variant, while acknowledging the strengths of the other methods.

In Table 17, which evaluates the performance of various algorithms on different datasets (namely CNS, Colon, DLBCL, Leukemia 1, and Lung Cancer), we can observe a continuation of the trend noted in the previous table. The iBABC algorithm remains competitive, with the iBABC-CGO-S variant achieving 100% accuracy on the DLBCL and Leukemia 1 datasets, and the iBABC-CGO-V variant also achieving the same 100% accuracy on these two datasets, but with slightly lower variability. However, it is essential to note that other algorithms also show good results. The rMRMR-MGWO algorithm, for example, achieves a remarkable 99.38% on the CNS dataset. The TLBOSA-SVM algorithm performs extraordinarily well, achieving a near-perfect 99.87% on the lung cancer dataset and 99.01% on the colon dataset. Another remarkable observation is the performance of CFC-FBBA, which boasts 100% accuracy on the DLBCL and Leukemia 1 datasets. On the other hand, algorithms such as BChOA-C-KNN and BChOA-KNN present impeccable results with 100% accuracy on the Leukemia 1 dataset. The IG-MBKH algorithm also deserves special mention, as it has consistently achieved results above 96% in all datasets on which it has been tested. While iBABC variants continue to achieve admirable results on different datasets, several algorithms show comparable performance in specific cases.

Table 12 Comparison of *i*BABC-CGO versus other metaheuristic algorithms in terms of Precision, Recall and ROC values using SVM classifier

Dataset	BASO	BEO	BGA	BHGSO	ABC	iABC	iBABC-S	iBABC-V	iBABC-CGO-S	iBABC-CGO-V
11-tumors										
Precision	0,7068	0,7470	0,6878	0,6774	0,7278	0,7327	0,7272	0,7267	0,9098	**1**
Recall	0,6143	0,7218	0,6299	0,5885	0,7010	0,6439	0,6296	0,6868	0,8989	**1**
ROC	0,8728	0,8506	0,8165	0,8301	0,8419	0,8047	0,8456	0,8690	0,8956	**1**
9-tumors										
Precision	0,4695	0,4867	0,4028	0,3810	0,4156	0,3741	0,4043	0,4596	0,8864	**0,9858**
Recall	0,4039	0,4427	0,4171	0,3862	0,4125	0,3718	0,3727	0,4292	0,8071	**0,9326**
ROC	0,8259	0,8757	0,8424	0,8659	0,8418	0,8740	0,8092	0,8387	0,8098	**0,9175**
Brain tumor 1										
Precision	0,6522	0,7194	0,6866	0,6010	0,6674	0,6156	0,6197	0,6640	**0,7611**	0,7593
Recall	0,5149	0,6804	0,5939	0,5299	0,6824	0,5193	0,4996	0,6541	0,7160	**0,7238**
ROC	0,8073	0,8479	0,8461	0,8598	0,8421	0,8912	0,8556	0,8603	0,8829	**0,9567**
Brain tumor 2										
Precision	0,8107	0,7742	0,8102	0,8027	0,7736	0,8018	0,7908	0,8146	0,8474	**0,8816**
Recall	0,7652	0,7503	0,7676	0,7619	0,7465	0,7696	0,7500	0,7740	0,8296	**0,8452**
ROC	0,8303	**0,9098**	0,8300	0,8817	0,7975	0,8299	0,8878	0,8537	0,9037	0,9055
Breast										
Precision	0,7455	0,7738	0,7363	0,6486	0,9500	0,6848	0,7513	0,7529	0,9072	**1**
Recall	0,7689	0,7689	0,7933	0,6289	0,1111	0,6578	0,7600	0,7467	**0,8356**	0,8044
ROC	0,7668	0,7844	0,7702	0,6635	0,5526	0,6936	0,7682	0,7645	0,7923	**0,7983**
CNS										
Precision	0,6532	0,6099	0,7460	0,6264	0,5686	0,7219	0,6295	0,5740	0,7556	**0,8393**
Recall	0,5750	0,5800	0,4850	0,5100	0,7000	0,2600	0,4950	0,7100	**0,8000**	0,7100
ROC	0,7067	0,6810	0,6989	0,6755	0,7064	0,6031	0,6719	0,7165	**0,7878**	0,7687

Table 12 (continued)

Dataset	BASO	BEO	BGA	BHGSO	ABC	iABC	iBABC-S	iBABC-V	iBABC-CGO-S	iBABC-CGO-V
Colon										
Precision	0,8267	0,8589	0,8321	0,8302	0,8581	0,8140	0,8376	0,8390	**0,9282**	0,8785
Recall	0,8949	0,6590	0,9051	0,8769	0,6308	0,9205	0,8897	0,6897	0,9179	**0,9436**
ROC	0,7793	0,7181	0,7889	0,7771	0,7199	0,7716	0,7903	0,7108	**0,8487**	0,8319
DLBCL										
Precision	0,8552	0,7309	0,9229	0,8406	0,7123	0,9175	0,8476	0,7207	0,9337	**0,9632**
Recall	0,8053	0,8053	0,7947	0,7947	0,8316	0,7737	0,7737	0,8368	**0,9684**	0,8842
ROC	0,8781	0,8526	0,8860	0,8711	0,8579	0,8746	0,8614	0,8640	**0,9667**	0,9298
Leukemia 1										
Precision	0,9221	0,9527	0,9525	0,9208	0,9508	0,9579	0,9209	0,9678	**9,82E-01**	9,78E-01
Recall	0,9674	0,9522	0,9761	0,9674	0,9609	0,9761	0,9783	0,9609	**0,9957**	0,9913
ROC	0,9077	0,9321	0,9420	0,9057	0,9344	0,9480	0,9111	0,9504	**0,9713**	0,9670
Lung cancer										
Precision	0,8867	0,8319	0,8184	0,8314	0,7878	0,8101	0,8520	0,7887	0,9420	**0,9551**
Recall	0,7402	0,8727	0,7192	0,7445	0,8309	0,7086	0,7485	0,8314	0,8886	**0,9054**
ROC	0,8559	0,7991	0,8657	0,8653	0,8264	0,8811	0,8069	0,8740	**0,9033**	0,8976
Lymphoma										
Precision	0,9926	0,9873	0,9909	0,9798	0,9753	0,9821	0,9851	0,9758	**1**	**1**
Recall	0,9948	0,9562	0,9687	0,9873	0,9486	0,9785	0,9813	0,9420	**1**	**1**
ROC	0,8455	0,8210	0,8277	0,8860	0,8616	0,8277	0,8650	0,8902	0,9148	**0,9156**
MLL										
Precision	0,9108	0,9155	0,9350	0,9081	0,9053	0,9434	0,8919	0,9060	0,9683	**0,9830**
Recall	0,9076	0,9154	0,9339	0,8983	0,9041	0,9404	0,8846	0,9059	0,9664	**0,9838**
ROC	0,8843	0,8197	0,8855	0,8585	0,9040	0,8960	0,8912	0,8631	9,01E-01	**0,9275**

Table 12 (continued)

Dataset	BASO	BEO	BGA	BHGSO	ABC	iABC	iBABC-S	iBABC-V	iBABC-CGO-S	iBABC-CGO-V
Ovarian										
Precision	0,9557	0,9390	0,9759	0,9654	0,9481	0,9736	0,9591	0,9364	9,88E-01	**1**
Recall	0,9963	0,9556	0,9957	0,9938	0,9574	0,9963	0,9969	0,9660	**1**	**1**
ROC	0,9565	0,9217	0,9756	0,9647	0,9315	0,9737	0,9601	0,9236	0,98654321	**1**
Prostate tumor										
Precision	0,8503	0,8543	0,8670	0,8380	0,8164	0,8490	0,8280	0,8582	0,9042	**0,9497**
Recall	0,8939	0,7918	0,8714	0,8510	0,6612	0,9020	0,8571	0,7776	**0,9673**	0,9265
ROC	0,8719	0,8315	0,8723	0,8476	0,7585	0,8751	0,8440	0,8282	0,9337	**0,9381**
SRBCT										
Precision	0,8420	0,8850	0,9042	0,8345	0,8434	0,8976	0,8385	0,8859	**1**	0,9573
Recall	0,8329	0,8701	0,8856	0,8173	0,8334	0,8928	0,8409	0,8714	**1**	0,9566
ROC	0,8302	0,8489	0,8438	0,8633	0,8444	0,8636	0,8478	0,8219	0,9161	**0,9167**

Bold indicates the best values in the table

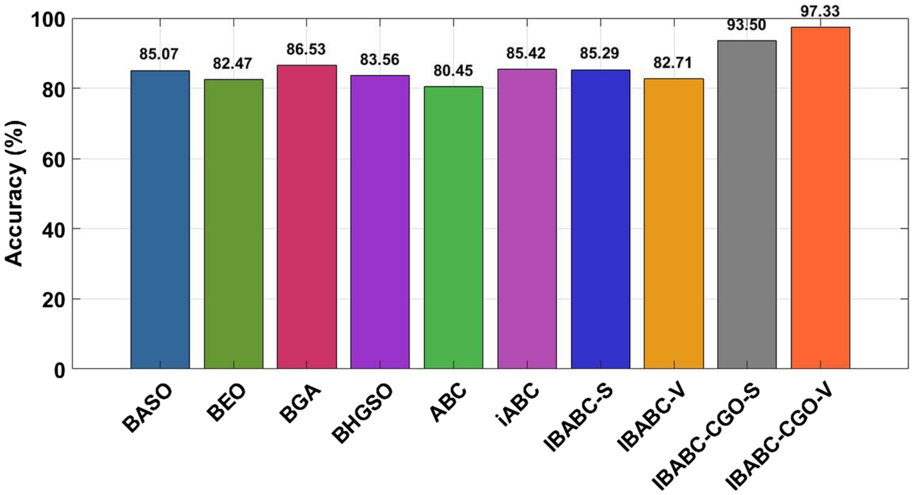


Fig. 6 Average accuracy on 15 datasets

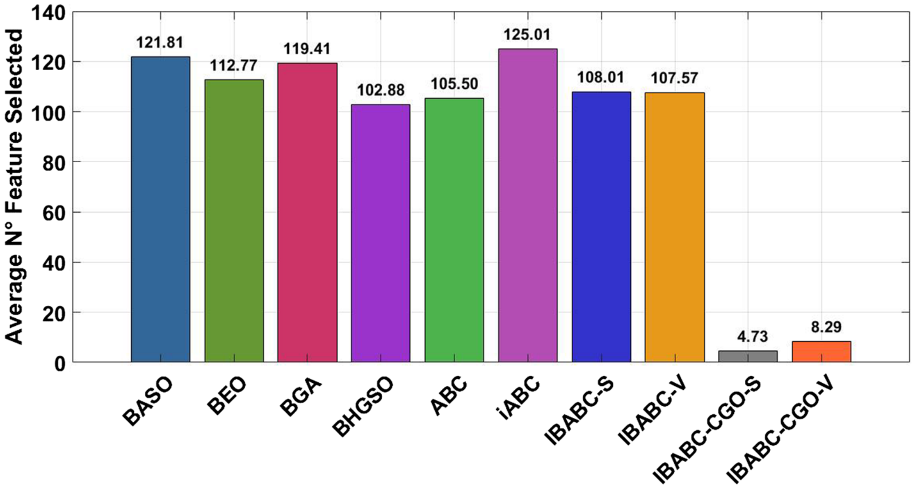


Fig. 7 Average number of features selected on 15 datasets

Table 18, evaluates the performance of the algorithms on the Lymphoma, MLL, Ovary, Prostate Tumor and SRBCT datasets. Many algorithms in the hybrid category, such as IG-MBKH, rMRMR-MGWO and CFC-FBBA, show consistently high performance across multiple datasets, reaching 100% accuracy in several cases. For example, the IG-MBKH method achieves perfect scores for the MLL and Ovarian datasets. Interestingly, the performance of the iBABC algorithm remains excellent, with both its variants (iBABC-CGO-S and iBABC-CGO-V) achieving impressive results. Both variants consistently achieve 100% accuracy on datasets such as Lymphoma, Ovarian and SRBCT, with minor variations in

Table 13 Comparison of *i*BABC-CGO versus other optimization algorithms in terms of the average *SNFs*

Dataset	**BASO**	**BEO**	**BGA**	**BHGSO**	**ABC**	**iABC**	**iBABC-S**	**iBABC-V**	**iBABC-CGO-S**	**iBABC-CGO-V**
11-tumors	165,3	23,0	184,5	102,1	39,0	141,0	189,3	205,3	**18,8**	21,2
9-tumors	153,3	83,8	176,9	35,3	68,8	189,0	65,3	187,2	**9**	24
Brain tumor 1	91,6	130,0	118,7	158,9	109,4	119,2	27,2	47,5	**4,8**	8,8
Brain tumor 2	206,2	73,0	192,1	44,8	81,5	163,5	191,0	45,7	**3,2**	3,4
Breast	162,4	64,1	202,3	35,2	103,5	90,6	81,6	114,3	**4,8**	24,6
CNS	24,2	48,9	64,3	55,1	164,3	146,4	30,5	188,9	**2,6**	7.6
Colon	179,0	77,5	20,0	148,0	104,0	52,0	150,3	34,1	**2,2**	**2,2**
DLBCL	157,5	148,3	58,0	31,1	123,0	129,8	28,8	70,6	2,4	**2**
Leukemia 1	153,7	23,8	162,7	143,8	110,5	57,8	82,5	89,1	**2,4**	2,4
Lung cancer	132,6	201,3	58,3	177,5	166,0	179,8	178,3	194,4	**7,2**	7,6
Lymphoma	84,0	193,4	62,7	213,1	85,9	200,9	67,2	123,0	**2**	**2**
MLL	158,4	126,4	60,4	50,8	23,1	41,3	18,7	155,4	**2**	2,2
Ovarian	120,1	141,9	165,3	130,6	173,1	93,3	105,8	109,6	**2,2**	3
Prostate tumor	17,5	168,7	78,6	51,5	114,4	159,2	210,2	23,5	2,8	**2,4**
SRBCT	21,6	187,7	186,4	165,4	116,0	111,4	193,5	24,9	**4,6**	4,8

Bold indicates the best values in the table

Table 14 *p*-values of the Wilcoxon signed rank test of *i*BABC-CGO-V versus other metaheuristic (insignificant values with p≥0.05 are highlighted in bold)

	BASO		**BEO**		**BGA**		**BHGSO**		**ABC**		**iABC**		**iBABC-S**		**iBABC-V**		**iBABC-CGO-S**	
	p	**win**	**p**	**win**	**p**	**win**	**p**	**win**	**p**	**win**	**p**	**win**	**p**	**win**	**p**	**win**	**p**	**win**
11-tumors	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	2,62E-07	+
9-tumors	1,61E-06	+	1,60E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	7,93E-04	+
Brain tumor 1	1,61E-06	+	1,60E-06	+	1,60E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	2,62E-07	+
Brain tumor 2	1,60E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,60E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	2,94E-07	-
Breast	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,60E-06	+	1,60E-06	+	1,61E-06	+	1,61E-06	+	3,72E-07	+
CNS	1,61E-06	+	1,60E-06	+	1,60E-06	+	1,60E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	2,62E-07	+
Colon	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	2,62E-07	+
DLBCL	1,60E-06	+	1,61E-06	+	1,60E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	2,62E-07	+
Leukemia 1	1,61E-06	+	1,61E-06	+	1,60E-06	+	1,60E-06	+	1,60E-06	+	1,61E-06	+	1,60E-06	+	1,60E-06	+	2,62E-07	-
Lung cancer	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,60E-06	+	1,61E-06	+	2,62E-07	+
Lym-phoma	1,61E-06	+	1,60E-06	+	1,60E-06	-	1,60E-06	+	1,60E-06	+	1,60E-06	+	1,60E-06	+	1,60E-06	+	5,27E-07	+
MLL	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,60E-06	+	1,61E-06	+	1,60E-06	+	1,60E-06	+	5,27E-07	+
Ovarian	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,60E-06	+	1,61E-06	+	1,61E-06	+	1,60E-06	+	1,61E-06	+	**1**	=
Prostate tumor	1,61E-06	+	1,61E-06	+	1,60E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	2,94E-07	+
SRBCT	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	1,61E-06	+	2,62E-07	+
Summary	15+/0=/0-		15+/0=/0-		14+/0=/1-		15+/0=/0-		15+/0=/0-		15+/0=/0-		15+/0=/0-		15+/0=/0-		12+/1=/2-	

Table 15 Statistical results of the Friedman test, Friedman align test and Quade test

Metric	BASO	BEO	BGA	BHGSO	ABC	iABC	iBABC-S	iBABC-V	iBABC-CGO-S	iBABC-CGO-V
Friedman value	6,05	6,94	4,97	7,00	8,31	5,66	6,42	6,65	1,69	**1,31**
Friedman rank	5	8	3	9	10	4	6	7	2	**1**
Friedman align value	6,20	6,49	6,24	6,68	6,23	7,40	5,23	6,63	2,27	**1,63**
Friedman align rank	4	7	6	9	5	10	3	8	2	**1**
Quad value	80,21	92,65	62,41	91,40	110,32	75,39	83,26	85,55	20,41	**6,17**
Quad rank	5	9	3	8	10	4	6	7	2	**1**

Bold indicates the best values in the table

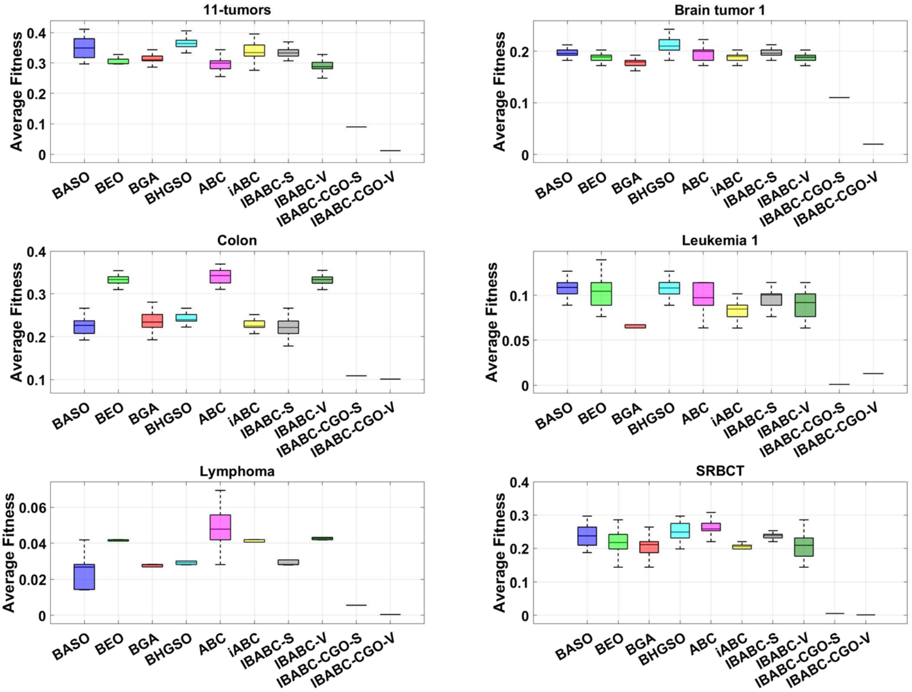


Fig. 8 Boxplots analysis of algorithms across six datasets

variability. The TLBOSA-SVM algorithm achieves a remarkable 99.13% on the prostate tumor dataset, and the BCOOT-CSA and BAOAC-SA methods both stand out for their uniform 100% accuracy across multiple datasets with minimal variability. On the SRBCT dataset, numerous algorithms, including BChOA-C-KNN, IG-MBKH, rMRMR-MGWO, IDGA-F-SVM, BCO-KNN, iBABC-CGO-S and iBABC-CGO-V, all achieved perfect 100% accuracy, despite variations in the associated variability values. In the filter category, the NB-mRMR method continues to display the robust performance of previous data sets. Both variants (100 and 50) achieved 100% accuracy for the Lymphoma and SRBCT datasets, making it a consistent contender. This comparative table reaffirms the effectiveness of hybrid algorithms, particularly the iBABC variants.

As also can be seen from Fig. 11, algorithms classified as hybrid methods frequently achieve high accuracy rates on a wide range of data sets. In the hybrid category, iBABC-CGO-S and iBABC-CGO-V variants consistently deliver first-rate performance. Their near-perfect or perfect accuracy scores on multiple datasets underline their robustness and efficiency. The NB-mRMR filter-based method also produces competitive accuracy rates. This shows that while hybrid methods can be powerful, well-optimized traditional methods can still be of great use. Beyond accuracy, the variability values provided with many algorithms give a better understanding of the situation. Algorithms with lower variability can be more consistent and therefore more reliable, particularly in critical applications such as medical diagnostics.

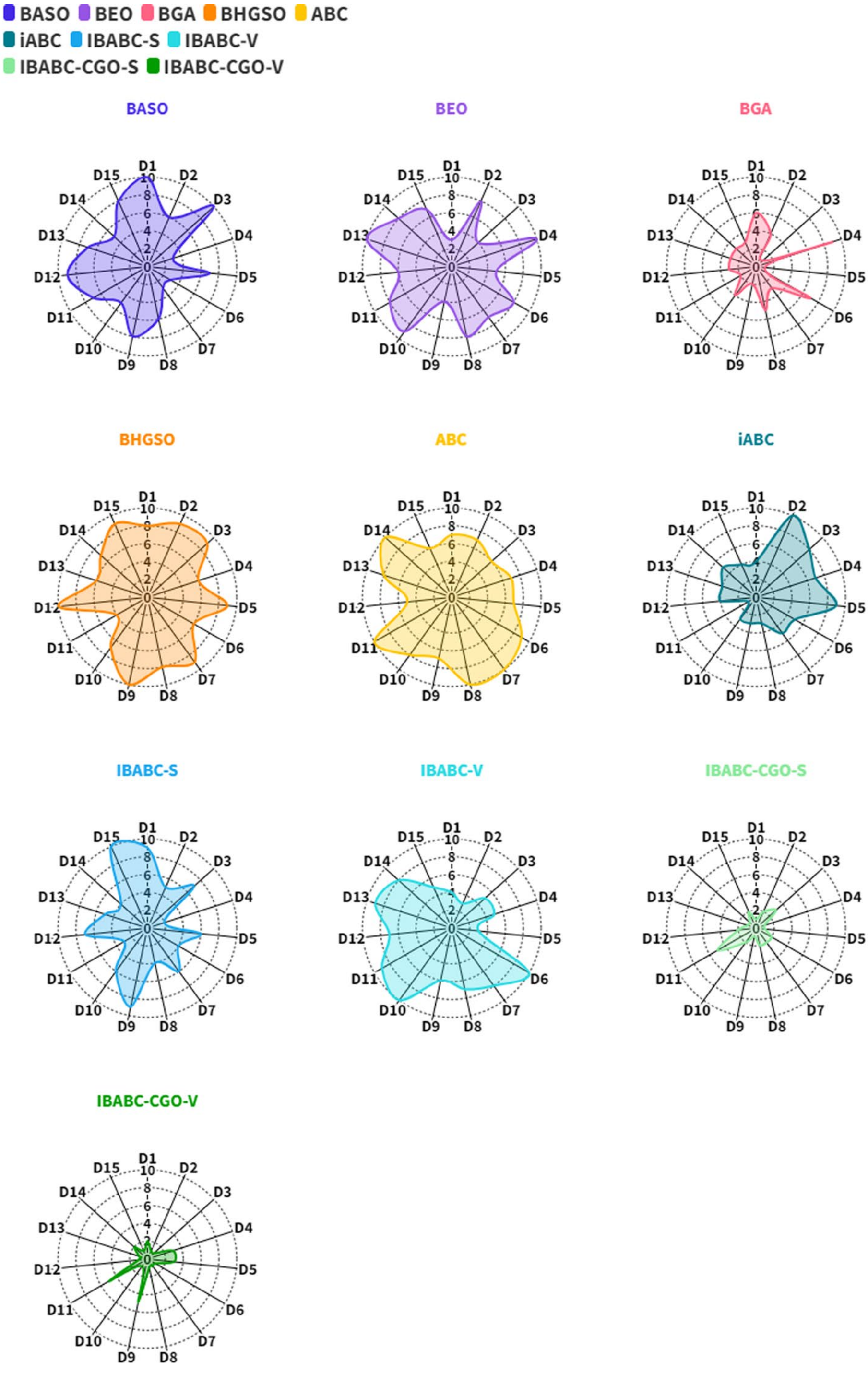


Fig. 9 The radar plot curves of the proposed *i*BABC-CGO and the comparative algorithms obtained with 15 medical datasets

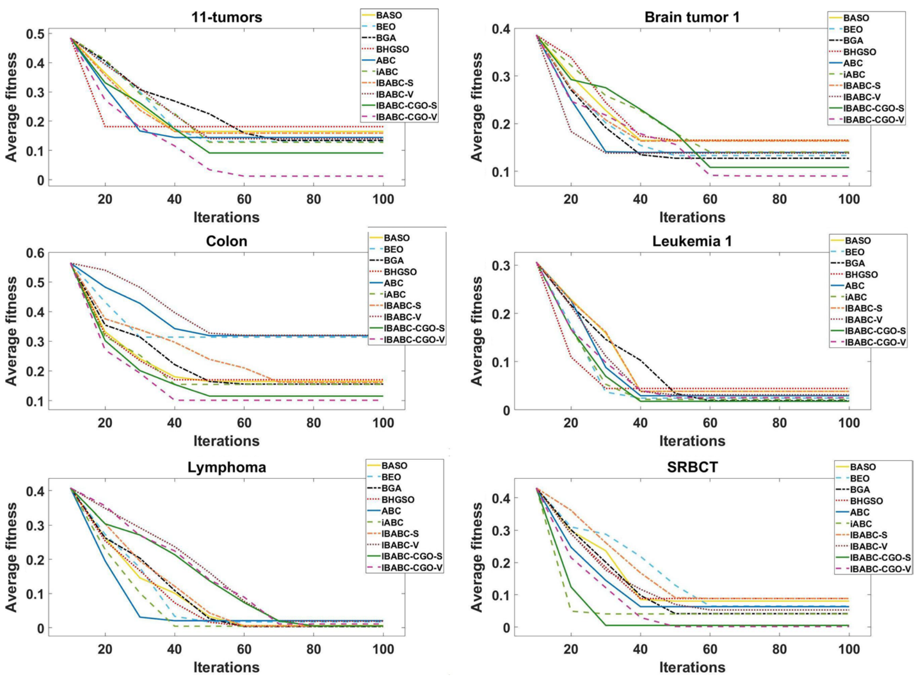


Fig. 10 Convergence curves of algorithms on six datasets

5.6 Overall assessment

Overall, the simulation results in the previous subsections demonstrate that *i*BABC-CGO-V is an effective wrapper-based feature selection method capable of solving a wide range of feature selection problems. Feature selection optimization differs from the continuous search space in which search agents can freely move. A thorough exploration is required to change the variable from 0 to 1 and vice versa. The algorithm must also avoid local optima, which occurs when the variables of a binary problem are changed infrequently compared to the number of iterations. The adaptive mechanism of the binary island ABC algorithm accelerates convergence. It also plays a crucial role in establishing an adequate exploitation-exploration balance, so that feature selection problems do not have many local solutions. This fundamentally distinguishes *i*BABC-CGO-V from the other algorithms in this study in terms of performance.

We have seen that the *i*BABC-CGO-V offers the highest accuracy and the fewest number of genes. Having said that, it is also important to note that *i*BABC-CGO-V demands more time to compute than the many other algorithms in this study. The suggested method uses the SVM classifier, and switching to a different classifier could result in longer execution times. As a result, attention should be taken when using a different classifier. We can conclude from the experimental analysis and comparisons that the V-shaped version of *i*BABC-CGO has the advantage of superior outcomes and high performance compared to other algorithms. The reasons for this high performance are attributed to the given factors; the island binary ABC provides the benefits of mobility, which can strike a better exploration-exploitation balance; the integration of the CGO algorithm further escapes local

Table 16 Evaluating the performance of the proposed method in comparison to existing approaches (Part1)

Type	Dataset algorithm	11-tumors	9-tumors	Brain tumor 1	Brain tumor 2	Breast
Hybrid-based	BChOA-C-KNN	95.14 (22.6)	–	95.85 (13.22)	–	–
	BChOA-KNN	93.75 (17.8)	–	95.77 (10.33)	–	–
	IG-MBKH	–	–	–	–	–
	rMRMR-MGWO	–	–	–	–	–
	IWSS²-MB (1NN)	–	–	–	–	–
	TLBOSA-SVM	92.23 (13)	73.51 (11)	96.98 (12)	–	–
	IDGA-F-SVM	–	–	–	–	**93.0 (7.6)**
	VLPSO-LS-KNN	82.81 (367.4)	56.78 (61.9)	75.54 (102.1)	73.25 (61.4)	–
	BCO-KNN	89.62 (24.1)	92.22 (28.3)	96.30 (20.5)	**100 (9)**	–
	PS-NSGA-KNN	83.94 (338.3)	58.30 (194.8)	73.81 (57.8)	73.07 (76.7)	–
	CDNC-SVM	–	–	–	–	–
	PSO-ensemble	–	–	–	–	86.36 (45)
	BCOOT-CSA	95.54 (35.2)	–	95.42 (12.66)	–	95.54 (15)
	BAOAC-SA	96.94 (24)	–	96.21 (7.5)	–	96.188 (10.8)
	CFC-FBBA	–	–	97.82 (10)	–	–
	iBABC-CGO-S	90.75 (18.8)	81.19 (9)	91.01 (4.8)	89.8 (3.2)	84.38 (4.8)
	iBABC-CGO-V	**100 (21.2)**	**96.92 (24)**	**100 (8.8)**	93.88 (3.4)	90.14 (24.6)
Filter-based	NB-mRMR-100	79.15 (100)	68.03 (100)	88.75 (100)	77.55 (100)	66.73 (100)
	NB-mRMR-50	70.50 (50)	59.39 (50)	86.47 (50)	77.55 (50)	64.68(50)

Bold indicates the best values in the table

optima and improves convergence rate. These advantages allow for avoiding the local optimum and accelerating convergence to the global optimum. Besides, when the solution is in an unsatisfactory search space region, the binary transfer function based on the V shape causes it to move. This feature, therefore, suggests that the algorithm operates effectively for the gene selection problem.

Table 17 Evaluating the performance of the proposed method in comparison to existing approaches (Part2)

Type	Dataset algorithm	CNS	Colon	DLBCL	Leukemia 1	Lung cancer
Hybrid-based	BChOA-C-KNN	–	–	–	100 (3.1)	–
	BChOA-KNN	–	–	–	100 (3.4)	–
	IG-MBKH	–	96.47 (17.1)	–	100.00 (4.20)	96.12(23.8)
	rMRMR-MGWO	**99.38 (17.46)**	95.86 (9.8)	–	100 (5.06)	97.91(15.8)
	IWSS^2-MB (1NN)	–	71.0 (5.4)	93.4 (9.1)	95.7 (7.3)	–
	TLBOSA-SVM	–	**99.01 (11)**	99.52 (11)	95.31 (12)	**99.87(10)**
	IDGA-F-SVM	–	–	98.8 (9.7)	100 (15)	–
	VLPSO-LS-KNN	–	–	96.13 (59.9)	93.75 (59.3)	90.17 (242.9)
	BCO-KNN	–	–	100 (3.1)	100 (3)	99.34 (32.1)
	PS-NSGA-KNN	–	–	87.02 (9.4)	91.98 (16.2)	88.48 (107.6)
	CDNC-SVM	–	88.89 (13.45)	–	91.16 (21.98)	91.82 (28.21)
	PSO-ensemble		92.86 (41)	–	100 (26)	–
	BCOOT-CSA	93.22 (7)	94.75 (8.75)	–	–	–
	BAOAC-SA	94.60 (5.2)	95.68 (7.66)	–	–	–
	CFC-FBBA	–	98.83 (1.80)	100.00 (5.7)	100 (3)	–
	iBABC-CGO-S	84.75 (2.6)	92.36 (2.2)	100 (2.4)	**100 (2.4)**	95.54 (7.2)
	iBABC-CGO-V	89.83 (7.6)	96.16 (2.2)	**100 (2)**	**100 (2.4)**	96.04 (7.6)
Filter-based	NB-mRMR-100	82.87 (100)	88.33 (100)	96 (100)	97.14 (100)	94.03 (100)
	NB-mRMR-50	79.39 (50)	86.66 (50)	96 (50)	97.14 (50)	93.51 (50)

Bold indicates the best values in the table

6 The biological interpretations of the selected genes by the proposed iBABC-CGO method

In this section, we delve into the biological significance of the best subset of genes that have been identified using our proposed method. For each binary dataset, which includes Breast, CNS, Colon, DLBCL, Leukemia, Ovarian, and Prostate, the gene names and indices are shown in Table 19. Their specific biological relevance can be further consulted in Tables 20, 21, 22, 23, and 24.

It should be mentioned that the interpretation of genes from the prostate cancer and DLBCL datasets is not included, given the absence of specific gene names in these datasets.

To further comprehend the biological implications of the selected genes, we relied on two reputable online resources and comprehensive databases of human genes, namely GeneCards (https://www.genecards.org/), and the National Center for Biotechnology Information (NCBI) (https://pubmed.ncbi.nlm.nih.gov/). These two databases present searchable and comprehensive genetic analysis data that provides concise information on all known and predicted human genes in the genome, proteome, transcription,

Table 18 Evaluating the performance of the proposed method in comparison to existing approaches (Part3)

Type	Dataset algorithm	Lymphoma	MLL	Ovarian	Prostate tumor	SRBCT
Hybrid-based	BChOA-C-KNN	–	–	–	97.52 (5.75)	**100 (4.4)**
	BChOA-KNN	–	–	–	97.49 (6)	100 (6.20)
	IG-MBKH	–	99.72 (11.1)	100 (3.4)	–	100 (6.30)
	rMRMR-MGWO	–	100 (8.4)	100 (3.56)	–	100 (37.5)
	IWSS^2-MB (1NN)	–	–	–	90.40 (9)	93.80 (7.3)
	TLBOSA-SVM	–	–	–	99.13 (8)	99.91 (5)
	IDGA-F-SVM	–	–	–	96.3 (14)	100 (18)
	VLPSO-LS-KNN	–	–	–	92.58 (56.4)	99.75 (71.4)
	BCO-KNN	–	–	–	**100 (7)**	100 (7.4)
	PS-NSGA-KNN	–	–	–	89.44 (65)	96.35 (18.6)
	CDNC-SVM	100 (3)	–		83.91 (22.81)	82.79 (17.43)
	PSO-ensemble	–	–	100 (13)	–	–
	BCOOT-CSA	**100 (2)**	–	100 (2.6)	–	100 (6.34)
	BAOAC-SA	**100 (2)**	100 (3.6)	100 (2.6)	–	100 (5.8)
	CFC-FBBA	100 (4.37)	100 (9.4)	100 (3)	99.42 (5.25)	–
	iBABC-CGO-S	**100 (2)**	98.59 (2)	**100 (2.2)**	94.06 (2.8)	100 (4.6)
	iBABC-CGO-V	**100 (2)**	**100 (2.2)**	100 (3)	97.03 (2.4)	100 (4.8)
Filter-based	NB-mRMR-100	100 (100)	95.71 (100)	98.4 (100)	93 (100)	100 (100)
	NB-mRMR-50	100 (50)	95.71 (50)	97.20 (50)	93 (50)	100 (50)

Bold indicates the best values in the table

genetics and function. The insights gained from these platforms confirm that our method is capable of identifying cancer-relevant genes in each respective dataset.

Furthermore, to better visualize the expression patterns of genes across various tumor samples, we employed heat maps Figs. 12 and 13. These graphical representations facilitate a deeper understanding of the data distribution and highlight differential gene expression across samples. The process involves the grouping of genes that share similar expression profiles into clusters. What's particularly noteworthy is that upon closer examination, a substantial proportion of these genes within these clusters exhibit a synchronized down-regulation in their expression levels. This coordinated down-regulation of gene expression suggests a strong likelihood of these genes being subject to shared regulatory mechanisms or being functionally linked. This result highlights not only the importance of understanding the collective behavior of genes, but also the potential importance of these co-regulated genes in specific biological processes or pathways.

The comprehensive gene analysis for the 28 genes listed in the Table 20, was carried out to ascertain their potential relation with breast cancer. These genes were studied for their respective pathways and functional associations using authoritative genomic databases.

Our analysis indicates that several of these genes are implicated in key cellular processes, fundamental for maintaining cell integrity and function. Some genes are particularly noteworthy:

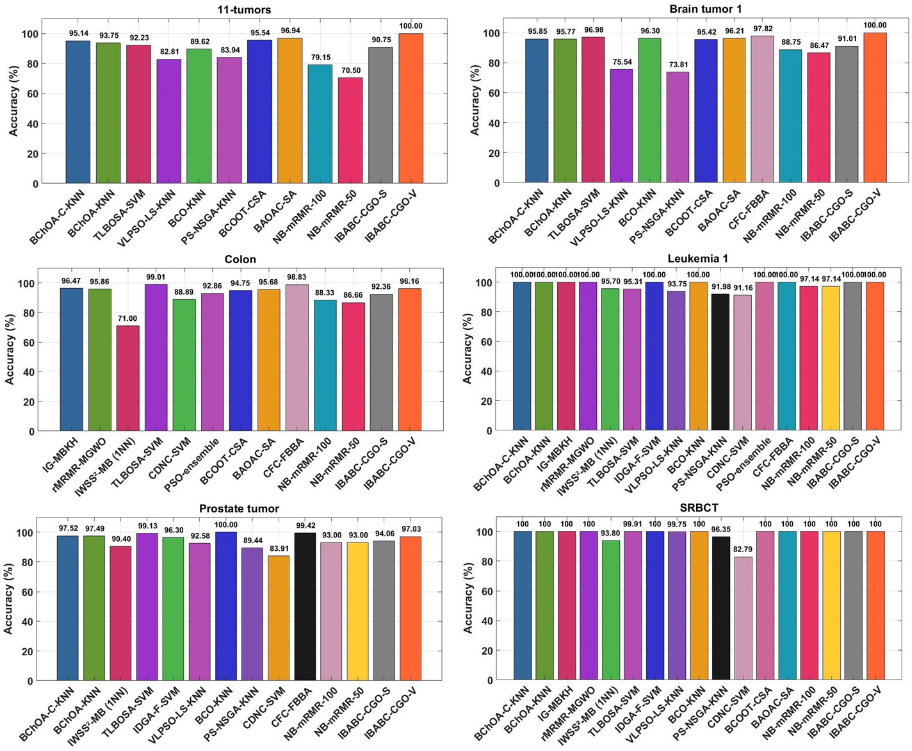


Fig. 11 Accuracy results of the various state-of-the-art hybrid methods and proposed iBABC algorithms for different datasets

-MDM4: Known to inhibit the p53 tumor suppressor protein, its overexpression has been linked with various human cancers. Its potential role in breast cancer could be explored in the context of p53 pathway dysregulation.

-CD44: This cell-surface glycoprotein is involved in cell-cell interactions, cell adhesion, and migration. It's implicated in many types of cancers, including breast cancer.

-CD69 Molecule: While CD69 is primarily known for its role in immune regulation, the association of this gene with diseases like Coccidioidomycosis and Eosinophilic Pneumonia suggests an inflammatory component that might be relevant in the tumor microenvironment of breast cancer.

-HIG1 Hypoxia Inducible Domain Family Member 2A: Hypoxia, or low oxygen levels, is a common feature in solid tumors like breast cancer. Genes associated with hypoxia can contribute to the survival and proliferation of tumor cells.

Based on their identified functions and associations, it is plausible that these genes can influence cellular activities such as adhesion, migration, differentiation, proliferation, and apoptosis processes that, when disrupted, can lead to tumorigenesis. Notably, the involvement of certain genes in pathways associated with cancer development makes them promising candidates as tumor-specific biomarkers.

For the CNS Cancer Dataset presented in Table 21, a deep-dive was executed into a selection of genes to determine their potential significance in CNS cancer. Notable genes and their relevance include:

Table 19 The best subset of selected genes from the gene selection method iBABC-CGO-V for binary datasets

Dataset	Index of genes	Gene names
Breast	651, 1488, 2085, 2370, 3280, 3671, 3939, 4358, 4439,5076, 6264, 6373, 6737, 7583, 8908, 9855, 10548, 10562,10567, 10595, 10887, 12479, 12507, 14497, 17968, 18191, 19950, 20382, 23766	NM_002393, NM_001781, NM_002575, Contig50520_RC, Contig46075_RC, Contig37204_RC, Contig36703_RC, Contig28311_RC, Contig37051_RC, NM_004304, AL117571, NM_003712, Contig48716_RC, X07695, NM_014166, NM_004859, NM_014391, NM_005647, NM_006378, AK000001, AL080058, NM_006597, NM_005878, Contig44611_RC, Contig28383_RC, NM_000036, NM_017687, NM_000308, X05299
CNS	318, 346, 842, 1003, 1052, 1429, 1690, 1695, 1804, 2402, 2472, 2494, 2693, 2715, 2801, 2991, 2994, 3006, 3222, 3483, 4147, 4171, 5635, 5810, 5885	D26561_cds2_at, D29954_at, HG2416-HT2512_at, HG4458-HT4727_at, J00306_at, L25876_at, M12529_at, M13149_at, M21389_at, M96684_at, S71129_at, S76067_at, U08989_at, U09584_at, U14973_at, U28727_at, U28833_at, U29725_at, U43923_at, U60415_at, X14675_at, X15875_at, Z84497_s_at, HG2987-HT3136_s_at, HG4264-HT4534_s_at
Colon	188, 1421, 1464, 1669	T63133, R48303, M18216
DLBCL	1451,2226	–
Leukemia	3521, 4845, 6330	U62962_at, X95654_at, X52282_s_at
Ovarian	1677, 1822, 2235	MZ244.66041, MZ288.82415, MZ434.68588
Prostate	3570, 6416, 7446, 8764	–

CDP-Diacylglycerol Synthase 2: It functions downstream of many G protein-coupled receptors and tyrosine kinases, which are crucial in various cellular signalling pathways. Any aberrations here can influence cell growth and proliferation, common in CNS cancers. It could be explored as a potential biomarker for signalling anomalies in CNS tumors.

Non-SMC Condensin II Complex Subunit D3: Its primary function in mitotic chromosome assembly suggests its importance in cell division. Malfunctions could lead to uncontrolled proliferation, a hallmark of cancer. Given its association with developmental diseases, this gene might be indicative of tumors with proliferative behavior.

Somatostatin: This hormone is involved in various physiological processes, including neurotransmission. Disruption in its expression might influence tumor growth, especially if CNS tumors show deregulation of neuroendocrine pathways.

Cyclin Dependent Kinase Inhibitor 3: Cyclin-dependent kinases play pivotal roles in cell cycle progression. Abnormalities in these pathways can lead to unchecked cell growth, typical of cancers. If overexpressed or silenced in CNS tumors, it may be a strong candidate biomarker for aggressive growth.

Apolipoprotein E: While its primary function relates to lipid metabolism, any genetic mutations might affect cell membrane structure or function, possibly aiding tumor cell survival or migration in CNS.

Table 20 The best subset of selected genes for Breast Cancer Dataset

Index of Genes	Gene Names	Gene Description	Specification
651	NM_002393	MDM4 Regulator Of P53	This gene encodes a nuclear protein that contains a p53 binding domain at the N-terminus and a RING finger domain at the C-terminus, and shows structural similarity to p53-binding protein MDM2. Both proteins bind the p53 tumor suppressor protein and inhibit its activity, and have been shown to be overexpressed in a variety of human cancers
1488	NM_001781	CD69 Molecule	This gene encodes a member of the calcium dependent lectin superfamily of type II transmembrane receptors. Diseases associated with CD69 include Coccidioidomycosis and Eosinophilic Pneumonia
2085	NM_002575	Serpin Family B Member 2	Predicted to enable serine-type endopeptidase inhibitor activity. Diseases associated with SERPINB2 include Gingivitis and Pre-Eclampsia
2370	Contig50520_RC	N-Alpha-Acetyltransferase 50, NatE Catalytic Subunit	Enables H4 histone acetyltransferase activity; peptide alpha-N-acetyltransferase activity; and peptidyl-lysine acetyltransferase activity. Diseases associated with NAA50 include Ogden Syndrome and Neurogenic Arthropathy
3280	Contig46075_RC	Cadherin Related 23	This gene is a member of the cadherin superfamily, whose genes encode calcium dependent cell-cell adhesion glycoproteins. The encoded protein is thought to be involved in stereocilia organization and hair bundle formation. The gene is located in a region containing the human deafness loci DFNB12 and USH1D. Usher syndrome 1D and nonsyndromic autosomal recessive deafness DFNB12 are caused by allelic mutations of this cadherin-like gene. Upregulation of this gene may also be associated with breast cancer
3671	Contig37204_RC	MBD3-Like Protein 5	MBD3L5 (Methyl-CpG Binding Domain Protein 3 Like 5) is a Protein Coding gene. Diseases associated with MBD3L5 include Ehlers-Danlos Syndrome, Spondylodysplastic Type, 1 and Ehlers-Danlos Syndrome, Musculocontractural Type, 1
3939	Contig36703_RC	Unc-5 Netrin Receptor C	This gene product belongs to the UNC-5 family of netrin receptors. Netrins are secreted proteins that direct axon extension and cell migration during neural development. Diseases associated with UNC5C include Alzheimer Disease, Familial, 1 and Immunodeficiency 43
4358	Contig28311_RC	HIG1 Hypoxia Inducible Domain Family Member 2A	The protein encoded by this gene is a subunit of the cytochrome c oxidase complex (complex IV), which is the terminal enzyme in the mitochondrial respiratory chain. Diseases associated with HIGD2A include Mitochondrial Complex Iv Deficiency, Nuclear Type 1

Table 20 (continued)

Index of Genes	Gene Names	Gene Description	Specification
4439	Contig37051_RC	—	—
5076	NM_004304	ALK Receptor Tyrosine Kinase	This gene encodes a receptor tyrosine kinase, which belongs to the insulin receptor superfamily. It plays an important role in the development of the brain and exerts its effects on specific neurons in the nervous system. This gene has been found to be rearranged, mutated, or amplified in a series of tumours including anaplastic large cell lymphomas, neuroblastoma, and non-small cell lung cancer
6264	AL117571	—	—
6373	NM_003712	Phospholipid Phosphatase 2	The protein encoded by this gene is a member of the phosphatidic acid phosphatase (PAP) family. Diseases associated with PLPP2 include Myxosarcoma
6737	Contig48716_RC	Long Intergenic Non-Protein Coding RNA 963	RNA Gene, and is affiliated with the lncRNA class. Diseases associated with LINC00963 include Prostate Disease and Renal Cell Carcinoma, Nonpapillary
7583	X07695	Keratin 4	The protein encoded by this gene is a member of the keratin gene family. Mutations in these genes have been associated with White Sponge Nevus, characterized by oral, esophageal, and anal leukoplakia
8908	NM_014166	Mediator Complex Subunit 4	This gene encodes a component of the Mediator complex. The Mediator complex interacts with DNA-binding gene-specific transcription factors to modulate transcription by RNA polymerase II. Diseases associated with MED4 include Achondrogenesis, Type Ib and Epiphyseal Dysplasia, Multiple, 4
9855	NM_004859	Clathrin Heavy Chain	Clathrin is a major protein component of the cytoplasmic face of intracellular organelles, called coated vesicles and coated pits. These specialized organelles are involved in the intracellular trafficking of receptors and endocytosis of a variety of macromolecules. Diseases associated with CLTC include Intellectual Developmental Disorder, Autosomal Dominant 56 and Autosomal Dominant Non-Syndromic Intellectual Disability
10548	NM_014391	Ankyrin Repeat Domain 1	The protein encoded by this gene is localized to the nucleus of endothelial cells and is induced by IL-1 and TNF-alpha stimulation. Diseases associated with ANKRD1 include Familial Isolated Dilated Cardiomyopathy and Dilated Cardiomyopathy

Table 20 (continued)

Index of Genes	Gene Names	Gene Description	Specification
10562	NM_005647	Transducin Beta Like 1 X-Linked	The protein encoded by this gene has sequence similarity with members of the WD40 repeat-containing protein family. The WD40 group is a large family of proteins, which appear to have a regulatory function. Diseases associated with TBL1X include Hypothyroidism, Congenital, Nongoitrous, 8 and Acute Gonococcal Prostatitis
10567	NM_006378	Semaphorin 4D	Enables identical protein binding activity; semaphorin receptor binding activity; and transmembrane signaling receptor activity. Diseases associated with SEMA4D include Cholangitis, Primary Sclerosing and Kallmann Syndrome
10595	AK000001	Multivesicular Body Subunit 12B	The protein encoded by this gene is a component of the ESCRT-I complex, a heterotetramer, which mediates the sorting of ubiquitinated cargo protein from the plasma membrane to the endosomal vesicle. ESCRT-I complex plays an essential role in HIV budding and endosomal protein sorting. Diseases associated with MVB12B include Nail-Patella Syndrome and Chronic Dacryocystitis
10887	AL080058	Monooxygenase DBH Like 1	Predicted to enable copper ion binding activity and dopamine beta-monooxygenase activity. Predicted to be involved in dopamine catabolic process; norepinephrine biosynthetic process; and octopamine biosynthetic process. Diseases associated with MOXD1 include Neuroaspergillosis
12479	NM_006597	Heat Shock Protein Family A (Hsp70) Member 8	This gene encodes a member of the heat shock protein 70 family, which contains both heat-inducible and constitutively expressed members. Diseases associated with HSPA8 include Auditory System Disease and Borna Disease
12507	NM_005878	CD44 Molecule (Indian Blood Group)	The protein encoded by this gene is a cell-surface glycoprotein involved in cell-cell interactions, cell adhesion and migration. This protein participates in a wide variety of cellular functions including lymphocyte activation, recirculation and homing, hematopoiesis, and tumor metastasis. CD44 has been implicated in various types of cancers, including breast cancer
14497	Contig44611_RC	Protein Tyrosine Phosphatase Non-Receptor Type 11	The protein encoded by this gene is a member of the protein tyrosine phosphatase (PTP) family. PTPs are known to be signaling molecules that regulate a variety of cellular processes including cell growth, differentiation, mitotic cycle, and oncogenic transformation. Diseases associated with PTPN11 include Noonan Syndrome 1 and Juvenile Myelomonocytic Leukemia

Table 20 (continued)

Index of Genes	Gene Names	Gene Description	Specification
17968	Contig28383_RC	HORMA Domain Containing 2	Predicted to be involved in meiotic sister chromatid cohesion. Diseases associated with HORMAD2 include Autoimmune Glomerulonephritis and Nephrotic Syndrome, Type 7
18191	NM_000036	Adenosine Monophosphate Deaminase 1	Adenosine monophosphate deaminase 1 catalyzes the deamination of AMP to IMP in skeletal muscle and plays an important role in the purine nucleotide cycle. Diseases associated with AMPD1 include Myopathy Due To Myoadenylate Deaminase Deficiency and Adenosine Monophosphate Deaminase Deficiency
19950	NM_017687	Zinc Finger Protein 180	Zinc finger proteins have been shown to interact with nucleic acids and to have diverse functions. Diseases associated with ZNF180 include Renal Cell Carcinoma, Nonpapillary
20382	NM_000308	Cathepsin A	This gene encodes a member of the peptidase S10 family of serine carboxypeptidases. Diseases associated with CTSA include Galactosialidosis and Cathepsin A-Related Arteriopathy-Strokes-Leukoencephalopathy
23766	X05299	Centromere Protein B	This gene product is a highly conserved protein that facilitates centromere formation. It is a DNA-binding protein that is derived from transposases of the pogo DNA transposon family. Diseases associated with CENPB include Crest Syndrome and Raynaud Disease

Table 21 The best subset of selected genes for CNS Cancer Dataset

Index of Genes	Gene Names	Gene Description	Specification
318	D26561_cds2_at	CDP-Diacylglycerol Synthase 2	Breakdown products of phosphoinositides are ubiquitous second messengers that function downstream of many G protein-coupled receptors and tyrosine kinases regulating cell growth, calcium metabolism, and protein kinase C activity
346	D29954_at	Non-SMC Condensin II Complex Subunit D3	Condensin complexes I and II play essential roles in mitotic chromosome assembly and segregation. Diseases associated with NCAPD3 include Microcephaly 22, Primary, Autosomal Recessive and Primary Autosomal Recessive Microcephaly
842	HG2416-HT2512_at	–	–
1003	HG4458-HT4727_at	–	–
1052	J00306_at	Somatostatin	The hormone somatostatin has active 14 aa and 28 aa forms that are produced by alternate cleavage of the single preproprotein encoded by this gene. Diseases associated with SST include Orofaciodigital Syndrome Viii and Placenta Disease
1429	L25876_at	Cyclin Dependent Kinase Inhibitor 3	The protein encoded by this gene belongs to the dual specificity protein phosphatase family. Diseases associated with CDKN3 include Hepatocellular Carcinoma and Cowden Syndrome
1690	M12529_at	Apolipoprotein E	The protein encoded by this gene is a major apoprotein of the chylomicron. Diseases associated with APOE include Lipoprotein Glomerulopathy and Hyperlipoproteinemia, Type Iii
1695	M13149_at	Histidine Rich Glycoprotein	This histidine-rich glycoprotein contains two cystatin-like domains and is located in plasma and platelets. Diseases associated with HRG include Thrombophilia Due To Histidine-Rich Glycoprotein Deficiency and Thrombophilia Due To Hrg Deficiency
1804	M21389_at	Keratin 5	The protein encoded by this gene is a member of the keratin gene family. Diseases associated with KRT5 include Epidermolysis Bullosa Simplex 2F, With Mottled Pigmentation and Epidermolysis Bullosa Simplex 2E, With Migratory Circinate Erythema

Table 21 (continued)

Index of Genes	Gene Names	Gene Description	Specification
2402	M96684_at	Purine Rich Element Binding Protein A	This gene product is a sequence-specific, single-stranded DNA-binding protein. Diseases associated with PURA include Neurodevelopmental Disorder With Neonatal Respiratory Insufficiency, Hypotonia, And Feeding Difficulties and Pura-Related Neurodevelopmental Disorders
2472	S71129_at	Acetylcholinesterase (Cartwright Blood Group)	Acetylcholinesterase hydrolyzes the neurotransmitter, acetylcholine at neuromuscular junctions and brain cholinergic synapses, and thus terminates signal transmission. Diseases associated with ACHE include Yt Blood Group Antigen and Colonic Pseudo-Obstruction
2494	S76067_at	Cyclic Nucleotide Gated Channel Subunit Alpha 2	The protein encoded by this gene represents the alpha subunit of a cyclic nucleotide-gated olfactory channel. Diseases associated with CNGA2 include Anosmia, Isolated Congenital and Retinitis Pigmentosa 45
2693	U08989_at	Solute Carrier Family 1 Member 1	This gene encodes a member of the high-affinity glutamate transporters that play an essential role in transporting glutamate across plasma membranes. Diseases associated with SLC1A1 include Dicarboxylic Aminoaciduria and Schizophrenia 18
2715	U09584_at	Transmembrane Protein 115	Enables identical protein binding activity. Involved in retrograde vesicle-mediated transport, Golgi to endoplasmic reticulum. Diseases associated with TMEM115 include Short-Rib Thoracic Dysplasia 2 With Or Without Polydactyly and Louse-Borne Relapsing Fever
2801	U14973_at	Ribosomal Protein S29	Ribosomes, the organelles that catalyze protein synthesis, consist of a small 40 S subunit and a large 60 S subunit. Together these subunits are composed of 4 RNA species and approximately 80 structurally distinct proteins. Diseases associated with RPS29 include Diamond-Blackfan Anemia 13 and Diamond-Blackfan Anemia
2991	U28727_at	Pappalysin 1	This gene encodes a secreted metalloproteinase which cleaves insulin-like growth factor binding proteins (IGFBPs). Diseases associated with PAPPA include Orofaciodigital Syndrome Viii and Placenta Disease

Table 21 (continued)

Index of Genes	Gene Names	Gene Description	Specification
2994	U28833_at	Regulator Of Calcineurin 1	The protein encoded by this gene interacts with calcineurin A and inhibits calcineurin-dependent signaling pathways, possibly affecting central nervous system development. Diseases associated with RCAN1 include Down Syndrome and Ulcer Of Lower Limbs
3006	U29725_at	Mitogen-Activated Protein Kinase 7	The protein encoded by this gene is a member of the MAP kinase family. MAP kinases act as an integration point for multiple biochemical signals, and are involved in a wide variety of cellular processes such as proliferation, differentiation, transcription regulation and development. Diseases associated with MAPK7 include Scoliosis, Isolated 1 and Multiple Endocrine Neoplasia, Type Iia
3222	U43923_at	SPT4 Homolog, DSIF Elongation Factor Subunit	This gene encodes the small subunit of DRB (5,6-dichloro-1-beta-d-ribofuranosylbenzimidazole) sensitivity-inducing factor (DSIF) complex, which regulates mRNA processing and transcription elongation by RNA polymerase II. Diseases associated with SUPT4H1 include Spinocerebellar Ataxia 36 and Cockayne Syndrome B
3483	U60415_at	Basic Helix-Loop-Helix ARNT Like 1	The protein encoded by this gene is a basic helix-loop-helix protein that forms a heterodimer with CLOCK. Diseases associated with BMAL1 include Delayed Sleep Phase Disorder and Rem Sleep Behavior Disorder
4147	X14675_at	BCR Activator Of RhoGEF And GTPase	A reciprocal translocation between chromosomes 22 and 9 produces the Philadelphia chromosome, which is often found in patients with chronic myelogenous leukemia. Diseases associated with BCR include Leukemia, Chronic Myeloid and B-Lymphoblastic Leukemia/Lymphoma With T(9;22)(Q34.1;Q11.2)
4171	X15875_at	Activating Transcription Factor 2	This gene encodes a transcription factor that is a member of the leucine zipper family of DNA binding proteins. Diseases associated with ATF2 include Retinoblastoma and Cardiomyopathy, Familial Hypertrophic, 25
5635	Z84497_s_at	Bromodomain Containing 2	This gene encodes a transcriptional regulator that belongs to the BET (bromodomains and extra terminal domain) family of proteins. Diseases associated with BRD2 include Epilepsy, Myoclonic Juvenile and Nut Midline Carcinoma

Table 21 (continued)

Index of Genes	Gene Names	Gene Description	Specification
5810	HG2987-HT3136_s_at	–	–
5885	HG4264-HT4534_s_at	–	–

Table 22 The best subset of selected genes for Colon Cancer Dataset

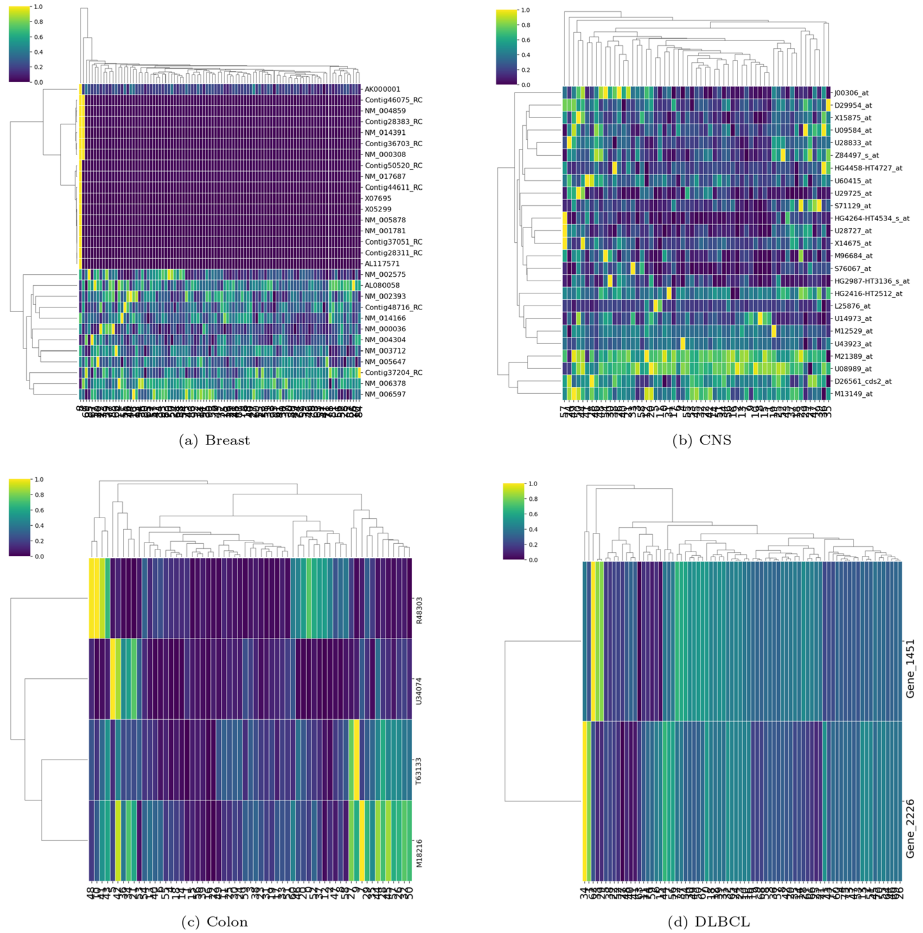
Index of genes	Gene names	Gene description	Specification
188	T63133	Thymosin Beta 10	Predicted to enable actin monomer binding activity. Diseases associated with TMSB10 include Diamond-Blackfan Anemia 5 and Thyroid Gland Cancer
1421	R48303	Dermatopontin	Dermatopontin is an extracellular matrix protein with possible functions in cell-matrix interactions and matrix assembly. Diseases associated with DPT include Leiomyoma and Febrile Seizures, Familial, 4
1464	M18216	CEA Cell Adhesion Molecule 6	This gene encodes a protein that belongs to the carcinoembryonic antigen (CEA) family whose members are glycosyl phosphatidyl inositol (GPI) anchored cell surface glycoproteins. Diseases associated with CEACAM6 include Cystic Fibrosis and Gastrointestinal Carcinoma

Table 23 The best subset of selected genes for Leukemia Cancer Dataset

Index of genes	Gene names	Gene description	Specification
3521	U62962_at	EukaryoticTranslation Initiation Factor 3 Subunit E	Enables protein N-terminus binding activity. Contributes to translation initiation factor activity. Diseases associated with EIF3E include Breast Cancer
4845	X95654_at	Synaptonemal Complex Protein 1	Enables double-stranded DNA binding activity. Involved in protein homotetramerization. Diseases associated with SYCP1 include Male Infertility and Azoospermia
6330	X52282_s_at	Natriuretic Peptide Receptor 3	This gene encodes one of three natriuretic peptide receptors. Natriutetic peptides are small peptides which regulate blood volume and pressure, pulmonary hypertension, and cardiac function as well as some metabolic and growth processes. Diseases associated with NPR3 include Boudin-Mortier Syndrome and Ischemic Colitis

Table 24 The best subset of selected genes for Ovarian Cancer Dataset

Index of genes	Gene names	Gene description	Specification
1677	MZ244.66041	ADAM Metallopeptidase With Thrombospondin Type 1 Motif 18	This gene encodes a member of the ADAMTS protein family. Diseases associated with ADAMTS18 include Microcornea, Myopic Chorioretinal Atrophy, and Telecanthus and Telecanthus
1822	MZ288.82415	APC Regulator Of WNT Signaling Pathway	This gene encodes a tumor suppressor protein that acts as an antagonist of the Wnt signaling pathway. Diseases associated with APC include Familial Adenomatous Polyposis 1 and Gastric Adenocarcinoma And Proximal Polyposis Of The Stomach
2235	MZ434.68588	Collagen Triple Helix Repeat Containing 1	This locus encodes a protein that may play a role in the cellular response to arterial injury through involvement in vascular remodeling. Diseases associated with CTHRC1 include Barrett Esophagus and Adenocarcinoma



(a) Breast

(b) CNS

(c) Colon

(d) DLBCL

Fig. 12 A heatmap representation of the gene expression patterns (Part 1)

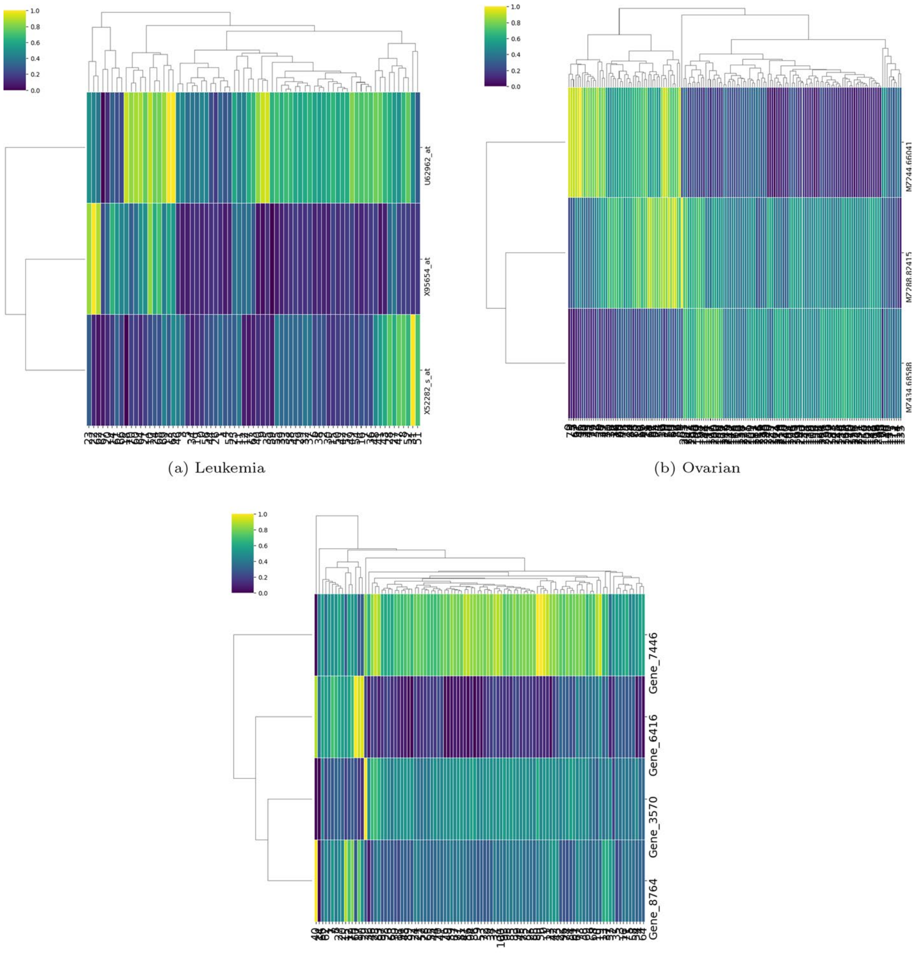
Histidine Rich Glycoprotein: Given its presence in plasma and platelets, any associations with vascular or angiogenesis-related changes in CNS tumors might make it a relevant biomarker.

Purine Rich Element Binding Protein A: DNA-binding proteins have regulatory functions. Disruptions can affect genes downstream, potentially driving tumorigenesis. It may be a marker for transcriptional anomalies in CNS tumors.

Solute Carrier Family 1 Member 1: Involved in neurotransmission, changes in its expression might relate to neurologically active tumors or those affecting neurotransmission in CNS cancer.

Bromodomain Containing 2: As a transcriptional regulator, its abnormal activity can lead to broad changes in gene expression profiles. If found to be consistently deregulated in CNS tumors, it may be a viable biomarker for diagnostic or prognostic purposes.

Several of these genes, based on their function and associated diseases, have clear implications in critical cellular processes and diseases. Some of these processes include cell growth, chromosome assembly, protein synthesis, neurotransmission, and DNA



(a) Leukemia

(b) Ovarian

(c) Prostate

Fig. 13 A heatmap representation of the gene expression patterns (Part 2)

binding. Understanding these genes' pathways and interactions can shed light on their role in CNS cancers.

For the Colon Cancer Dataset, the implications of each gene presented in Table 22 and their potential as diagnostic or prognostic biomarkers are explored below:

-Thymosin Beta 10: Thymosins play key roles in cell migration, differentiation, and proliferation, processes that are crucial during tumorigenesis. Thymosin Beta 10 is implicated in actin monomer binding, suggesting a potential role in the cell's cytoskeletal dynamics. Its association with thyroid gland cancer implies it might be involved in other malignancies as well.

-Dermatopontin: Extracellular matrix (ECM) proteins are fundamental in tissue architecture and integrity. Dysregulation of ECM components, like Dermatopontin, can influence cell-cell and cell-matrix interactions, potentially facilitating tumor growth, invasion, and metastasis in the colon.

-CEA Cell Adhesion Molecule 6: Carcinoembryonic antigen (CEA) family members, including CEACAM6, have been previously associated with various cancers. They play roles in cell adhesion, a fundamental process that when altered can contribute to the invasiveness of cancer cells and their potential to metastasize. Given its association with gastrointestinal carcinoma, CEACAM6 might serve as a diagnostic or prognostic biomarker in colon cancer.

As with any potential biomarkers, it is essential to validate their association with colon cancer in extensive clinical studies. This involves comparing their expression or mutation status in a broad range of colon cancer samples and correlating these findings with clinical outcomes.

For the Leukemia Cancer Dataset, each gene's potential implications in leukemia presenetd in Table 23 and their possible utility as diagnostic or prognostic biomarkers is depicted below:

-Eukaryotic Translation Initiation Factor 3 Subunit E: Translation initiation factors are vital for protein synthesis. Any aberrations in their function can lead to deregulated protein synthesis, which might contribute to malignant transformation and tumor progression. EIF3E's association with breast cancer implies that its dysregulation might be important in other cancers as well.

-Natriuretic Peptide Receptor 3: Natriuretic peptides play roles in regulating blood volume, blood pressure, and certain metabolic processes. Changes in the function of their receptors, like NPR3, might lead to altered cellular responses, which could potentially influence the leukemic microenvironment or the behavior of leukemia cells.

For the Ovarian Cancer Dataset, understanding the role of each gene and its implications in ovarian cancer presented in Table 24 can shed light on potential diagnostic or prognostic biomarkers.

-ADAM Metallopeptidase With Thrombospondin Type 1 Motif 18: ADAMTS family members are known for their roles in tissue remodeling and cell-matrix interactions. Given that cancer often involves tissue remodeling, changes in the expression of this gene may influence ovarian tumor growth, invasion, or metastasis. If ADAMTS18 is found to be consistently deregulated (either overexpressed or underexpressed) in ovarian tumors compared to normal tissue, it could serve as a potential biomarker for disease onset, progression, or even therapeutic targeting.

-APC Regulator Of WNT Signaling Pathway: The Wnt signaling pathway is pivotal in numerous cellular processes, including cell growth, differentiation, and migration. The APC protein acts as a tumor suppressor, and its dysfunction can lead to aberrant activation of the Wnt pathway, promoting tumorigenesis. Its role is well-established in colorectal cancers, but any changes in ovarian tumors could indicate similar pathway dysregulation.

-Collagen Triple Helix Repeat Containing 1: While primarily involved in vascular remodeling, the process is also crucial in tumor growth, as tumors require new blood vessels to support their rapid growth-a process called angiogenesis. Alterations in CTHRC1 might be linked to ovarian tumor vascularization or metastatic potential.

For effective clinical translation, it would be essential to study these genes in a larger cohort of ovarian cancer patients, assessing their expression or mutation status and correlating with clinical outcomes. This will determine their true potential as biomarkers for ovarian cancer diagnosis, prognosis, or therapeutic intervention.

7 Pros and Cons of the proposed method

This section describes the advantages and disadvantages of the suggested *i*BABC-CGO-V method. In addition, suggestions for improving this strategy are also highlighted. Numerous research works demonstrate that metaheuristic algorithms (MHs) are excellent optimization approaches, although they may have downsides such as premature convergence, an imbalanced exploration-exploitation ratio, and an inability to escape from the local optimal region. In order to create an efficient FS approach using an MH algorithm to improve classification performance, we researched contemporary algorithms and their features. The *i*ABC is a relatively latest modified version of the ABC metaheuristic algorithm that has been shown in the literature to be effective at resolving practical optimization issues. It utilises the structured population mechanism by incorporating the island model for improving the exploration of the original ABC. Due to its straightforward approach and few parameters, the *i*ABC is simple to implement. It demonstrated better performance on CEC 2015 functions in (Awadallah et al. 2020) when compared to a number of other MH methods.

However, besides its advantages, the *i*ABC still struggles with slow convergence, lack of direct application to the discrete optimization problems, unbalanced exploration-exploitation ratio, and the possibility of getting trapped in local optima. Therefore, we integrated two mechanisms to the original *i*ABC, which are highlighted as the following:

- We adapted the S and V transfer functions to convert the continuous version of the *i*ABC to its equivalent binary version so that it fits the discrete FS problem. Besides the binary conversion, the transfer functions also reportedly help in improving exploration and exploitation abilities.
- CGO method is incorporated in the migration phase of the original *i*ABC to further improve exploration, rate of convergence, and escape local optima.

These advantages of integration of binary transfer functions, CGO method, and use of SVM classifier in *i*BABC-CGO-V approach also come with certain disadvantages, which are listed along with their future proposed resolutions below:

- However, to select the minimum *NFs* for classifying features accurately, the proposed approach still requires more improvement as it provides second-best results. This may be due to the use of a weighted approach and manual adjustment of weights for the lowest classification error rate and the fewest features in the fitness calculation. This problem can be resolved by employing the multi-objective approach, which considers both objectives viz. classification error and selection of the fewest features simultaneously.
- *i*BABC-CGO-V requires more time to compute than many compared algorithms in this study due to the use and integration of binary transfer functions, the CGO method, and the use of an island approach. Careful use and integration with the other metaheuristic approach are the resolutions to reduce the compute time of the suggested approach. The convergence rate can be further enhanced by using chaotic maps or opposition-based learning methods in the initialization phase.

However, as per the NFL theory, no universally effective optimization algorithm exists for all optimization problems for all datasets. As a result, the authors conclude that the proposed *i*BABC-CGO-V, like the other MH methods, follows the same rule; nonetheless, it surpasses several other contemporary and well-known algorithms.

8 Conclusion

This paper presents a unique Hybrid island Binary Artificial Bee Colony with Chaos Game Optimization (*i*BABC-CGO) method as a wrapper-based technique for feature selection problem. This new approach merges the simplicity and flexibility of the ABC algorithm, the island concept that ensures diversity during the search process of solutions, the advantage of the CGO algorithm ensuring better exploration-exploitation ratio and escaping local optima, and a binary model to better address the discrete feature selection optimization problem using microarray data. In this work, the efficacy of the proposed technique was validated using 15 real-world high-dimensional datasets. According to our research, the suggested *i*BABC-CGO method beats the competing algorithms across most datasets. Furthermore, the *i*BABC-CGO successfully identified a subset of highly discriminative characteristics that effectively characterised the target ideas among competitors. Overall, in feature selection operations on 15 datasets, the *i*BABC-CGO frequently attains the best accuracy on most individual datasets and the highest overall average accuracy as well. The proposed method obtained the best results among tested metaheuristics in terms of highest classification accuracy, highest overall average accuracy, competitive number of selected features, and largest area under the ROC curves. At last, the biological interpretations of the selected genes by the proposed method are provided, which delve into the biological significance of the best subset of genes that have been identified by our method. Future research can utilize *i*BABC-CGO-V for various technical and clinical applications, including electromyography pattern recognition, optimized deep neural networks, and power quality. The performance of *i*BABC-CGO can also be improved by using several improved initialization procedures, the selection of an alternative classifier, and the use of a multi-objective approach for optimizing classification accuracy and the fewest number of features selected objectives simultaneously.

Additional simulations

See Figs. 14 and 15.

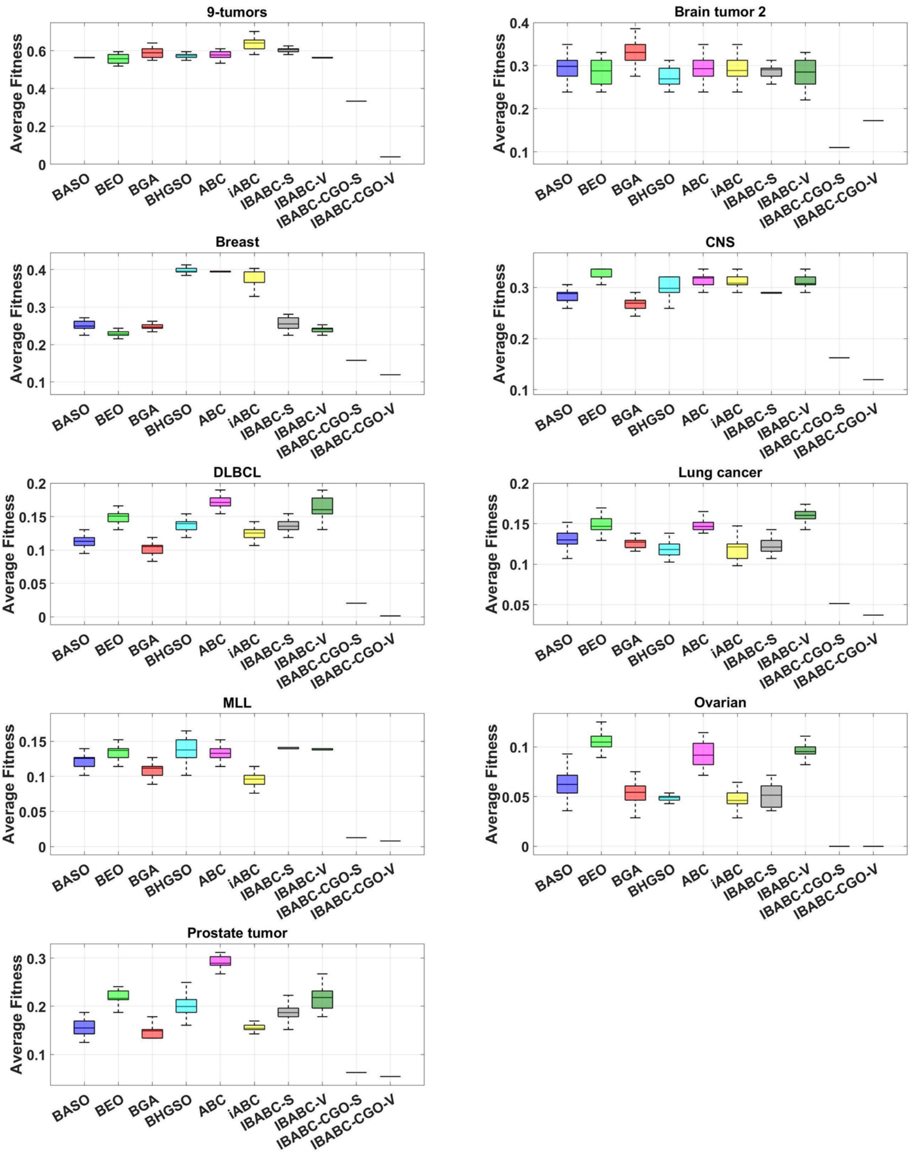


Fig. 14 Boxplots analysis of algorithms across 9 remaining datasets

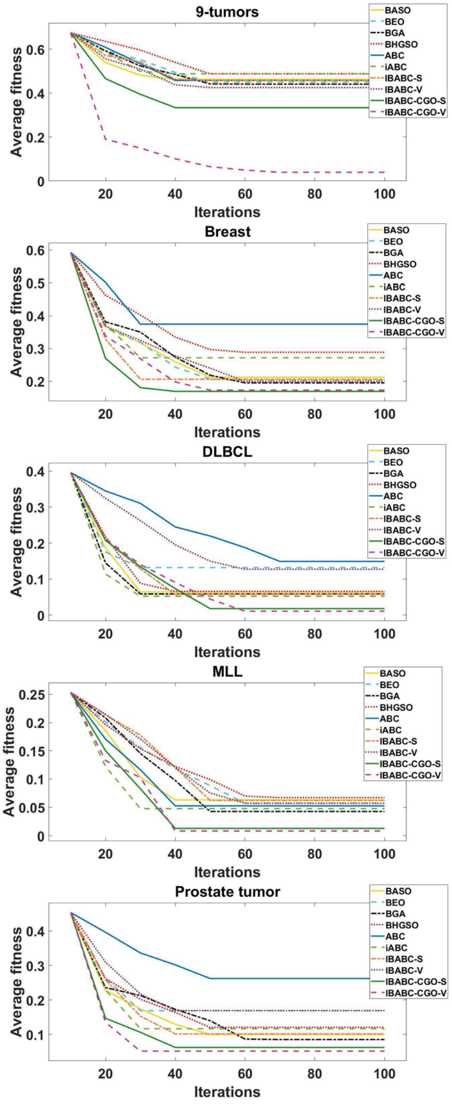


Fig. 15 Convergence analysis of algorithms across 9 remaining datasets

Declarations

Conflicts of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

Abadlia H, Smairi N, Ghedira K (2017) Particle swarm optimization based on dynamic island model, in: 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 709–716

Abiodun EO, Alabdulatif A, Abiodun OI, Alawida M, Alabdulatif A, Alkhawaldeh RS (2021) A systematic review of emerging feature selection optimization methods for optimal text classification: the present state and prospective opportunities. Neural Comput Appl 33:15091–15118

Abu Khurma R, Aljarah I, Sharieh A, Abd Elaziz M, Damaševičius R, Krilavičius T (2022) A review of the modification strategies of the nature inspired algorithms for feature selection problem. Mathematics 10:464

Abualigah L, Elaziz MA, Khasawneh AM, Alshinwan M, Ibrahim RA, Al-Qaness MA, Mirjalili S, Sumari P, Gandomi AH (2022) Meta-heuristic optimization algorithms for solving real-world mechanical engineering design problems: a comprehensive survey, applications, comparative analysis, and results. Neural Comput Appl 34:4081–4110

Agrawal P, Abutarboush HF, Ganesh T, Mohamed AW (2021) Metaheuristic algorithms on feature selection: a survey of one decade of research (2009–2019). IEEE Access 9:26766–26791

Ahmed MS, Shahjaman M, Rana MM, Mollah MNH et al (2017) Robustification of naïve bayes classifier and its application for microarray gene expression data analysis. BioMed Res Int 2017:3020627

Al-Betar MA, Awadallah MA (2018) Island bat algorithm for optimization. Expert Syst Appl 107:126–145

Al-Betar MA, Awadallah MA, Khader AT, Abdalkareem ZA (2015) Island-based harmony search for optimization problems. Expert Syst Appl 42:2026–2035

Alomari OA, Khader AT, Al-Betar MA, Abualigah LM (2017) Mrmr ba: a hybrid gene selection algorithm for cancer classification. J Theor Appl Inf Technol 95:2610–2618

Alomari OA, Makhadmeh SN, Al-Betar MA, Alyasseri ZAA, Doush IA, Abasi AK, Awadallah MA, Zitar RA (2021) Gene selection for microarray data classification based on gray wolf optimizer enhanced with triz-inspired operators. Knowl Based Syst 223:107034

Alrefai N, Ibrahim O (2022) Optimized feature selection method using particle swarm intelligence with ensemble learning for cancer classification based on microarray datasets. Neural Comput Appl 34:13513–13528

Alshamlan H, Badr G, Alohali Y (2019) Microarray gene selection and cancer classification method using artificial bee colony and SVM algorithms (ABC-SVM), in: Abawajy JH, Othman M, Ghazali R, Deris MM, Mahdin H, Herawan T (Eds), Proceedings of the International Conference on Data Engineering 2015 (DaEng-2015), Lecture Notes in Electrical Engineering, Springer, Singapore, 2019, pp. 575–584. https://doi.org/10.1007/978-981-13-1799-6-59

Al-Tashi Q, Rais H, Jadid S (2018) Feature selection method based on grey wolf optimization for coronary artery disease classification. In: Saeed F, Gazem N, Patnaik S, Balaid ASS, Mohammed F (eds) International conference of reliable information and communication technology. Springer, Berlin, pp 257–266

Al-Thanoon NA, Algamal ZY, Qasim OS (2021) Feature selection based on a crow search algorithm for big data classification. Chemom Intell Lab Syst 212:104288

Apolloni J, Leguizamón G, Alba E (2016) Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments. Appl Soft Comput 38:922–932

Araujo L, Merelo JJ (2010) Diversity through multiculturality: Assessing migrant choice policies in an island model. IEEE Trans Evol Comput 15:456–469

Awadallah MA, Al-Betar MA, Bolaji AL, Doush IA, Hammouri AI, Mafarja M (2020) Island artificial bee colony for global optimization. Soft Comput 24:13461–13487

Aziz RM (2022) Cuckoo search-based optimization for cancer classification: a new hybrid approach. J Comput Biol 29:565–584

Babatunde OH, Armstrong L, Leng J, Diepeveen D (2014) A genetic algorithm-based feature selection. Int J Electron Commun Comput Eng 5:2278–4209

Cantú-Paz E et al (1998) A survey of parallel genetic algorithms. Calculateurs paralleles, reseaux et systems repartis 10:141–171

Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electric Eng 40:16–28

Chen G, Chen J (2015) A novel wrapper method for feature selection and its applications. Neurocomputing 159:219–226

Chen Y, Miao D, Wang R (2010) A rough set approach to feature selection based on ant colony optimization. Pattern Recognit Lett 31:226–233

Chen Z, Xuan P, Heidari AA, Liu L, Wu C, Chen H, Escorcia-Gutierrez J, Mansour RF (2023) An artificial bee bare-bone hunger games search for global optimization and high-dimensional feature selection. Iscience 26:106679

Coleto-Alcudia V, Vega-Rodríguez MA (2020) Artificial bee colony algorithm based on dominance (ABCD) for a hybrid gene selection method. Knowl Based Syst 205:106323

Corcoran AL, Wainwright RL (1994) A parallel island model genetic algorithm for the multiprocessor scheduling problem, in: Proceedings of the 1994 ACM symposium on Applied computing, pp. 483–487

Črepinšek M, Liu S-H, Mernik M (2013) Exploration and exploitation in evolutionary algorithms: a survey. ACM Comput Surv (CSUR) 45:1–33

da Silveira LA, Soncco-Álvarez JL, de Lima TA, Ayala-Rincón M (2019) Parallel Island Model Genetic Algorithms applied in NP-Hard problems. IEEE Congress on Evolutionary Computation (CEC) 2019:3262–3269

Dash M, Liu H (1997) Feature selection for classification. Intell Data Anal 1:131–156

Dashtban M, Balafar M (2017) Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts. Genomics 109:91–107

Del Ser J, Osaba E, Molina D, Yang X-S, Salcedo-Sanz S, Camacho D, Das S, Suganthan PN, Coello CAC, Herrera F (2019) Bio-inspired computation: where we stand and what's next. Swarm Evolut Comput 48:220–250

H. Dhrif, L. G. S. Giraldo, M. Kubat, S. Wuchty, A stable hybrid method for feature subset selection using particle swarm optimization with local search, in: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 19, Association for Computing Machinery, Prague, Czech Republic, 2019, pp. 13–21. https://doi.org/10.1145/3321707.3321816

Duarte G, Lemonge A, Goliatt L (2017) A dynamic migration policy to the Island Model. IEEE Congress on Evolutionary Computation (CEC) 2017:1135–1142

Emary E, Zawbaa HM, Hassanien AE (2016) Binary ant lion approaches for feature selection. Neurocomputing 213:54–65. https://doi.org/10.1016/j.neucom.2016.03.101

Erguzel TT, Tas C, Cebi M (2015) A wrapper-based approach for feature selection and classification of major depressive disorder-bipolar disorders. Comput Biol Med 64:127–137

Esfandiari A, Farivar F, Khaloozadeh H (2023) Fractional-order binary bat algorithm for feature selection on high-dimensional microarray data. J Ambient Intell Humaniz Comput 14:7453–7467

Fernandez F, Tomassini M, Vanneschi L (2003) An empirical study of multipopulation genetic programming. Genet Progr Evol Mach 4:21–51

Fleuret F (2004) Fast binary feature selection with conditional mutual information. J Mach Learn Res 5:1531–1555

Friedman J, Hastie T, Tibshirani R et al (2001) The elements of statistical learning, vol 10. Springer, New York

Fushiki T (2011) Estimation of prediction error by using K-fold cross-validation. Stat Comput 21:137–146

Gao Y, Zhou Y, Luo Q (2020) An efficient binary equilibrium optimizer algorithm for feature selection. IEEE Access 8:140936–140963

Garro BA, Vazquez RA, Rodriguez K (2014) Classification of DNA microarrays using Artificial Bee Colony (ABC) algorithm. In: Tan Y, Shi Y, Coello CAC (eds) Advances in swarm intelligence. Lecture Notes in Computer Science. Springer, Cham, pp 207–214. https://doi.org/10.1007/978-3-319-11857-4-24

Ghosh M, Begum S, Sarkar R, Chakraborty D, Maulik U (2019) Recursive memetic algorithm for gene selection in microarray data. Expert Syst Appl 116:172–185

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182

Hancer E, Xue B, Zhang M (2018) Differential evolution for filter feature selection based on information theory and feature ranking. Knowl Based Syst 140:103–119

Huang C-L (2009) ACO-based hybrid classification system with feature subset selection and model parameters optimization. Neurocomputing 73:438–448

X. Jin, A. Xu, R. Bie, P. Guo, Machine learning techniques and chi-square feature selection for cancer classification using sage gene expression profiles, in: Data Mining for Biomedical Applications: PAKDD 2006 Workshop, BioDM 2006, Singapore, April 9, 2006. Proceedings, Springer, 2006, pp. 106–115

Kabir MM, Shahjahan M, Murase K (2012) A new hybrid ant colony optimization algorithm for feature selection. Expert Syst Appl 39:3747–3763

Karaboga D, Gorkemli B, Ozturk C, Karaboga N (2014) A comprehensive survey: artificial bee colony (ABC) algorithm and applications. Artif Intell Rev 42:21–57

Kushida JI, Hara A, Takahama T, Kido A (2013) Island-based differential evolution with varying subpopulation size, in: 2013 IEEE 6th International Workshop on Computational Intelligence and Applications (IWCIA), pp. 119–124

Lal TN, Chapelle O, Weston J, Elisseeff A (2006) Embedded Methods. In: Guyon I, Nikravesh M, Gunn S, Zadeh LA (eds) Feature extraction: foundations and applications, studies in fuzziness and soft computing. Springer, Berlin, Heidelberg, pp 137–165

Lipowski A, Lipowska D (2012) Roulette-wheel selection via stochastic acceptance. Physica A 391:2193–2196

Liu W, Chen H, Chen L (2013) An ant colony optimization based algorithm for identifying gene regulatory elements. Comput Biol Med 43:922–932

Lu H, Chen J, Yan K, Jin Q, Xue Y, Gao Z (2017) A hybrid feature selection algorithm for gene expression data classification. Neurocomputing 256:56–62

Maleki N, Zeinali Y, Niaki STA (2021) A k-nn method for lung cancer prognosis with the use of a genetic algorithm for feature selection. Expert Syst Appl 164:113981

Masoudi-Sobhanzadeh Y, Motieghader H, Omidi Y, Masoudi-Nejad A (2021) A machine learning method based on the genetic and world competitive contests algorithms for selecting genes or features in biological applications. Sci Rep 11:1–19

Meiri R, Zahavi J (2006) Using simulated annealing to optimize the feature selection problem in marketing applications. Euro J Oper Res 171:842–858

Mernik M, Liu S-H, Karaboga D, Črepinšek M (2015) On clarifying misconceptions when comparing variants of the artificial bee colony algorithm by offering a new implementation. Inf Sci 291:115–127

Mirjalili S, Lewis A (2013) S-shaped versus v-shaped transfer functions for binary particle swarm optimization. Swarm Evol Comput 9:1–14

Mora AM, García-Sánchez P, Merelo JJ, Castillo PA (2013) Pareto-based multi-colony multi-objective ant colony optimization algorithms: an island model proposal. Soft Comput 17:1175–1207

Mukherjee S, Classifying microarray data using support vector machines, in: A practical approach to microarray data analysis, Springer, Cham. pp. 166–185

Neggaz N, Houssein EH, Hussain K (2020) An efficient henry gas solubility optimization for feature selection. Expert Syst Appl 152:113364

Oduntan IO, Toulouse M, Baumgartner R, Bowman C, Somorjai R, Crainic TG (2008) A multilevel tabu search algorithm for the feature selection problem in biomedical data. Comput Math Appl 55:1019–1033

Oliveira LS, Sabourin R, Bortolozzi F, Suen CY (2003) A methodology for feature selection using multiobjective genetic algorithms for handwritten digit string recognition. Int J Pattern Recognit Artif Intell 17:903–929

Oliveira AL, Braga PL, Lima RM, Cornélio ML (2010) Ga-based method for feature selection and parameters optimization for machine learning regression applied to software effort estimation. Info Softw Technol 52:1155–1166

Palomo-Romero JM, Salas-Morera L, García-Hernández L (2017) An island model genetic algorithm for unequal area facility layout problems. Expert Syst Appl 68:151–162

Pashaei E (2022) Mutation-based binary aquila optimizer for gene selection in cancer classification. Comput Biol Chem 101:107767

Pashaei E, Pashaei E (2021) Gene selection using hybrid dragonfly black hole algorithm: a case study on RNA-seq covid-19 data. Anal Biochem 627:114242

Pashaei E, Pashaei E (2022) An efficient binary chimp optimization algorithm for feature selection in biomedical data classification. Neural Comput Appl 34:6427–6451

Pashaei E, Pashaei E (2022) Hybrid binary arithmetic optimization algorithm with simulated annealing for feature selection in high-dimensional biomedical data. J Supercomput 78:15598–15637

Pashaei E, Pashaei E (2023) Hybrid binary coot algorithm with simulated annealing for feature selection in high-dimensional microarray data. Neural Comput Appl 35:353–374

Ponmalar A, Dhanakoti V (2022) An intrusion detection approach using ensemble support vector machine based chaos game optimization algorithm in big data platform. Appl Soft Comput 116:108295

Qasim OS, Al-Thanoon NA, Algamal ZY (2020) Feature selection based on chaotic binary black hole algorithm for data classification. Chemom Intell Lab Syst 204:104104

Qi C, Diao J, Qiu L (2019) On estimating model in feature selection with cross-validation. IEEE Access 7:33454–33463

Ramadan A, Kamel S, Hussein MM, Hassan MH (2021) A new application of chaos game optimization algorithm for parameters extraction of three diode photovoltaic model. IEEE Access 9:51582–51594

Rao H, Shi X, Rodrigue AK, Feng J, Xia Y, Elhoseny M, Yuan X, Gu L (2019) Feature selection based on artificial bee colony and gradient boosting decision tree. Appl Soft Comput 74:634–642

Rey D, Neuhäuser M (2011) Wilcoxon-signed-rank test. International encyclopedia of statistical science. Springer, Berlin, pp 1658–1659

Robnik-Šikonja M, Kononenko I (2003) Theoretical and empirical analysis of relieff and rrelieff. Mach Learn 53:23–69

Roshanzamir M, Balafar MA, Razavi SN (2020) A new hierarchical multi group particle swarm optimization with different task allocations inspired by holonic multi agent systems. Expert Syst Appl 149:113292

Rostami M, Forouzandeh S, Berahmand K, Soltani M, Shahsavari M, Oussalah M (2022) Gene selection for microarray data classification via multi-objective graph theoretic-based method. Artif Intell Med 123:102228

Ruciński M, Izzo D, Biscani F (2010) On the impact of the migration topology on the island model. Parallel Comput 36:555–571

Saha SK, Sarkar S, Mitra P (2009) Feature selection techniques for maximum entropy based biomedical named entity recognition. J Biomed Inform 42:905–911

Sánchez-Maroño N, Alonso-Betanzos A, Tombilla-Sanromán M (2007) Filter methods for feature selection—a comparative study. In: Yin H, Tino P, Corchado E, Byrne W, Yao X (eds) Intelligent data engineering and automated learning—IDEAL 2007, lecture notes in computer science. Springer, Berlin, Heidelberg, pp 178–187

Sharma A, Rani R (2019) C-hmoshssa: gene selection for cancer classification using multi-objective meta-heuristic and machine learning methods. Comput Methods Progr Biomed 178:219–235

Shukla AK, Singh P, Vardhan M (2019) A new hybrid wrapper tlbo and sa with svm approach for gene expression data. Inf Sci 503:238–254

Shukla AK, Tripathi D, Reddy BR, Chandramohan D (2020) A study on metaheuristics approaches for gene selection in microarray data: algorithms, applications and open challenges. Evolut Intell 13:309–329. https://doi.org/10.1007/s12065-019-00306-6

Skolicki Z, De Jong K (2005) The influence of migration sizes and intervals on island models, in: Proceedings of the 7th annual conference on Genetic and evolutionary computation, pp. 1295–1302

Talatahari S, Azizi M (2020) Optimization of constrained mathematical and engineering design problems using chaos game optimization. Comput Indust Eng 145:106560

Talatahari S, Azizi M (2021) Chaos game optimization: a novel metaheuristic algorithm. Artif Intell Rev 54:917–1004

Talbi E-G (2002) A taxonomy of hybrid metaheuristics. J Heuristics 8:541–564

Tomassini M (2006) Spatially structured evolutionary algorithms: artificial evolution in space and time. Springer, Cham

Too J, Rahim Abdullah A (2020) Binary atom search optimisation approaches for feature selection. Connect Sci 32:406–430

Tran B, Xue B, Zhang M (2018) Variable-length particle swarm optimization for feature selection on high-dimensional classification. IEEE Trans Evol Comput 23:473–487

Turgut MS, Turgut OE, Eliiyi DT (2020) Island-based Crow search algorithm for solving optimal control problems. Appl Soft Comput 90:106170

Vieira SM, Sousa J, Runkler TA (2009) Multi-criteria ant feature selection using fuzzy classifiers. In: Caello CAC, Dehuri S, Ghosh S (eds) Swarm intelligence for multi-objective problems in data mining. Springer, Berlin, pp 19–36

Wang Y (2010) A sociopsychological perspective on collective intelligence in metaheuristic computing. Int J Appl Metaheuristic Comput (IJAMC) 1:110–128

Wang X, Yang J, Teng X, Xia W, Jensen R (2007) Feature selection based on rough sets and particle swarm optimization. Pattern Recognit Lett 28:459–471

Wang A, An N, Yang J, Chen G, Li L, Alterovitz G (2017) Wrapper-based gene selection with Markov blanket. Comput Biol Med 81:11–23

Wang H, Jing X, Niu B (2017) A discrete bacterial algorithm for feature selection in classification of microarray gene expression cancer data. Knowl Based Syst 126:8–19

Wang Y-Y, Zhang H, Qiu C-H, Xia S-R (2018) A novel feature selection method based on extreme learning machine and fractional-order Darwinian PSO. Comput Intell Neurosci 2018:1–8

Whitley D, Rana S, Heckendorn RB (1997) Island model genetic algorithms and linearly separable problems. AISB international workshop on evolutionary computing. Springer, Cham, pp 109–125

Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. IEEE Trans Evolut Comput 1:67–82

Wu G, Mallipeddi R, Suganthan PN (2019) Ensemble strategies for population-based optimization algorithms—a survey. Swarm Evol Comput 44:695–711

Xue B, Zhang M, Browne WN (2012) Particle swarm optimization for feature selection in classification: a multi-objective approach. IEEE Trans Cybern 43:1656–1671

Xue B, Zhang M, Browne WN, Yao X (2015) A survey on evolutionary computation approaches to feature selection. IEEE Trans Evolut Comput 20:606–626

Yan C, Ma J, Luo H, Patel A (2019) Hybrid binary coral reefs optimization algorithm with simulated annealing for feature selection in high-dimensional biomedical datasets. Chemom Intell Lab Syst 184:102–111

Yang K, Cai Z, Li J, Lin G (2006) A stable gene selection in microarray data analysis. BMC Bioinform 7:228

Yang X-S, Deb S, Fong S (2014) Metaheuristic algorithms: optimal balance of intensification and diversification. Appl Math Inf Sci 8:977

Yaqoob A, Aziz RM, Verma NK, Lalwani P, Makrariya A, Kumar P (2023) A review on nature-inspired algorithms for cancer disease prediction and classification. Mathematics 11:1081

Yusta SC (2009) Different metaheuristic strategies to solve the feature selection problem. Pattern Recognit Lett 30:525–534

Zhang G, Hou J, Wang J, Yan C, Luo J (2020) Feature selection for microarray data classification using hybrid information gain and a modified binary krill herd algorithm. Interdiscip Sci: Comput Life Sci 12:288–301

Zhou Y, Zhang W, Kang J, Zhang X, Wang X (2021) A problem-specific non-dominated sorting genetic algorithm for supervised feature selection. Inf Sci 547:841–859

Zhu G, Kwong S (2010) Gbest-guided artificial bee colony algorithm for numerical function optimization. Appl Math Comput 217:3166–3173

Zorarpacı E, Özel SA (2016) A hybrid approach of differential evolution and artificial bee colony for feature selection. Expert Syst Appl 62:91–103

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Maha Nssibi[1,2] · Ghaith Manita[1,3] · Amit Chhabra[4] · Seyedali Mirjalili[5,6] · Ouajdi Korbaa[7]

✉ Amit Chhabra
amit.cse@gndu.ac.in

Maha Nssibi
maha.nssibi@ensi-uma.tn

Ghaith Manita
gaith.manita@esen.tn

Seyedali Mirjalili
ali.mirjalili@gmail.com

Ouajdi Korbaa
Ouajdi.Korbaa@centraliens-lille.org

[1] Laboratory MARS, LR17ES05, ISITCom, University of Sousse, Sousse, Tunisia

[2] ENSI, University of Manouba, Manouba, Tunisia

[3] ESEN, University of Manouba, Manouba, Tunisia

[4] Department of Computer Engineering and Technology, Guru Nanak Dev University, Amritsar, India

[5] Centre for Artificial Intelligence Research and Optimisation, Torrens University Australia, Brisbane, Australia

[6] YFL (Yonsei Frontier Lab), Yonsei University, Seoul, Korea

[7] ISITCom, University of Sousse, Sousse, Tunisia