RAIRO-Oper. Res. 60 (2026) 1053–1080
https://doi.org/10.1051/ro/2026021

RAIRO Operations Research
www.rairo-ro.org

HYBRID GENE SELECTION AND CLASSIFICATION OF CANCER MICROARRAY DATA USING AN IMPROVED BINARY FIREFLY ALGORITHM

Brahim Sahmadi[1,2,*] and Dalila Boughaci[1]

Abstract. Cancer microarray datasets are distinguished by their high dimensionality and a relatively small sample sizes, which presents significant challenges for accurate cancer classification. Gene selection therefore becomes essential to eliminate irrelevant genes and improve classification accuracy. This paper presents a hybrid approach combining filter and wrapper techniques for gene selection, integrating an improved binary firefly algorithm and the support vector machine classifier. The objective is to select the most cancer-related genes to decrease computation time and enhance classification model performance. Three filter methods (Information Gain Ratio, ReliefF, and Correlation-based Feature Selection) are used in ensemble with the enhanced binary firefly algorithm. The firefly algorithm's exploration and exploitation capabilities are improved through opposition-based learning during initialization and movement of the fireflies. Additionally, a mutation step is added to improve the diversity of solutions. To validate our approach, we conducted an experimental study on twelve public benchmark datasets and compared it to several recent gene selection methods used for cancer gene expression data classification. The results reveal that the suggested methodology enhances classifier performance while reducing data volume by finding a limited group of genes with strong predictive power for cancer classification.

Mathematics Subject Classification. 68T20, 90C27, 68T05, 90C59, 62F30.

Received January 30, 2025. Accepted February 6, 2026.

1. Introduction

DNA microarray technology allows multiple genes to be measured simultaneously in a biological sample. The obtained gene expression data can be applied in various medical contexts, including the categorization of tissue samples into normal or malignant. However, automatic analysis of such datasets can be enormously inefficient, as the data produced from gene expressions contain a high level of noise and have a high dimensionality. These datasets are defined by a substantial number of genes within a limited number of samples, leading to major challenges such as the "curse of dimensionality"–a problem that arises when there is a significant number of features in a limited number of training instances – [1].

In a microarray dataset, most genes may be redundant, noise-related, or irrelevant to the classification process. Several studies have revealed that a majority of genes whose expression is measured in microarray experiments

Keywords. Gene selection, cancer microarray dataset, classification, binary firefly algorithm, support vector machine.

[1] LRIA/Computer Science Department, University of Sciences and Technology Houari Boumediene (USTHB), Algiers, Algeria.
[2] LMP2M Laboratory, University Yahia Fares of Medea, Medea, Algeria.
*Corresponding author: bsahmadi@usthb.dz

lack relevance in many classification problems [2]. In this context, the selection of relevant genes among all the genes tested in the microarray data is often considered a necessary pre-treatment step. It is important for data analysis because it can reduce the volume of the microarray data, lower the risk of overfitting or underfitting, and decrease the computation time required for the classification task.

Gene selection consists of choosing from a large set of genes, a small, relevant subset of genes relative to the problem under study. It can accelerate the process of data mining; enhance the predictive accuracy of classification algorithms and simplify the interpretation of the resulting models [3]. The methodologies employed for assessing a subset of genes in selection algorithms are traditionally categorized into three types: filter, wrapper, and embedded methods [4].

Filter methods select genes independently of the classifier that will be utilized in the next step of classification. Typically, these methods rank genes based on specific criteria of merit and then choose the highest-ranking genes. F-score, mutual information and information gain illustrate examples of popular filters. Recent studies have explored the use of multiple filtering techniques on DNA microarray data. For instance, Ghosh *et al.* [5] investigates chi-square, ReliefF and mutual information filters, while [6] suggests a combined method that includes several filtering techniques, such as: information gain, minimum redundancy and maximum relevance (mRMR), ReliefF, chi-square, and Spearman correlation. In Sun *et al.* [7], the authors propose a filtering strategy based on the Area Under the Curve (AUC) and gene complementarity to identify highly discriminative genes.

Wrapper methods operate in a closed-loop system, using feedback from the classifier to assist in identifying the optimal subset of genes. These methods generally aim to maximize classification accuracy with the chosen gene subset. For example, Xu *et al.* [8] uses Support Vector Machine Recursive Feature Elimination (SVM-RFE) to identify relevant genes. Other approaches include surrogate-assisted particle swarm optimization [9]. A memory-guided cuckoo search is used in Alzaqebah *et al.* [10]. Particle swarm optimization algorithm is also used in Song *et al.* [11] with a dynamic feature clustering. In Chatra *et al.* [12], a binary bat algorithm is employed for feature selection and extreme learning machine for classification. Evolutionary computation continues to gain traction for gene selection in high-dimensional biomedical data. A binary multi-objective differential evolution based self-learning (MOFS-BDE) was proposed in Zhang *et al.* [13] to simultaneously optimize classifier accuracy and dimensionality reduction. Similarly, Hu *et al.* [14] introduces a multi-objective fuzzy feature selection method based on PSO to handle feature selection problems involving fuzzy cost.

Embedded methods integrate the procedure of identifying an optimal gene subset into the training process. Unlike wrapper methods, they utilize the classifier not only to assess a potential gene subset but also to direct the gene selection process [4]. In Guo *et al.* [15], an L1-regularized method is used to eliminate redundant genes and identify key biomarkers. The approach in Kang *et al.* [16] named "relaxed Lasso-GenSVM", uses relaxed Lasso for gene selection and a generalized multi-class SVM (GenSVM) for classification. Similarly, Alharthi *et al.* [17] introduces penalized logistic regression using the adaptive elastic net (AEN) to improve gene selection accuracy.

Recently, hybrid gene selection approaches –which combine filter and wrapper techniques– have attracted considerable interest owing to their capacity to combine the advantage of the computational efficiency of filter techniques and the superior performance of wrapper algorithms. For instance, the minimum redundancy maximum relevance (mRMR) filter method has been enhanced through integration with wrapper-based optimizers such as the Manta Ray Foraging Optimization Algorithm [18], the Binary Arithmetic Optimization Algorithm [19], and Binary Differential Evolution [20]. In Dashtban and Balafar [1], two filter methods (Laplacian and Fisher score) are utilized alongside an integer-coded genetic algorithm. Other notable combinations include PSO with ensemble learning [21], mutual information with Monte Carlo Tree Search [22], and SVM-RFE with the Gini index [23]. A two-stage hybrid framework using ensemble filtering and the Binary African Vultures Optimization Algorithm is proposed in Li *et al.* [24]. In Xie *et al.* [25], a feature selection algorithm centered on a multidimensional graph is implemented for the classification of microarray datasets. While [26] presents a combined filter–GA strategy for cancer classification, relevance-complementarity ratio and graph-based analysis are integrated in Chamlal *et al.* [27], and Bir-Jmel *et al.* [28] proposes two hybrid strategies–FBPSO-SVM and

RBPSO-1NN–combining Fisher score and ReliefF with binary PSO. A comprehensive review of different feature selection methods used in the context of microarray data can be found in [29,30], [11] and [31].

Filter feature selection methods are fast but often produce less accurate predictions, whereas wrapper methods yield better performance but require more computing resources and time, particularly in the context of high-dimensional microarray data. To address this trade-off, we propose a hybrid gene selection methodology that combines both approaches to efficiently search for small subsets of highly predictive genes for cancer classification. In our method, an ensemble of three filter techniques–Information Gain Ratio, ReliefF, and Correlation-based Feature Selection (CFS)–is first applied to rapidly reduce the dimensionality, using multiple complementary criteria. Then, a wrapper strategy based on an Enhanced Binary Firefly Algorithm (EBFA) is employed. Unlike traditional Firefly variants [32,33], our EBFA introduces two major enhancements: Opposition-Based Learning (OBL) during both initialization and updates of solutions to improve global exploration, and a dynamic mutation operator to maintain diversity and escape local optima. A Support Vector Machine (SVM) is used as the classifier to evaluate the fitness of gene subsets. We evaluate the suggested approach using twelve cancer microarray datasets of various sizes, spanning both binary and multi-class classification problems. Results are compared across different filter methods and with previous methods employed for feature selection in the classification of microarray data.

The main contributions of this study are as follows:

- To address the limitations of standalone filter or wrapper approaches by combining them in a two-phase hybrid strategy for selecting relevant genes for cancer classification tasks,
- To improve the exploration-exploitation balance of the Firefly Algorithm using OBL and mutation, enabling more effective feature space navigation and faster convergence,
- To evaluate the proposed method on twelve benchmark cancer microarray datasets (both bi-class and multi-class) with varying dimension, and compare it with other state-of-the-art feature selection methods,
- To demonstrate the ability of the method to identify compact, biologically meaningful gene subsets while maintaining high classification performance.

This paper is structured as follows: Section 2 covers DNA microarrays, support vector machines, the filtering methods used in this study, the concept of opposition-based learning, and the Firefly algorithm. Section 3 develops the methodology used for selecting genes and classifying cancer microarray data. Section 4 gives the findings and analyzes them in extensive detail. In the end, Section 5 summarizes the results and gives some recommendations for further study.

2. Preliminaries

This section is devoted to generalities about DNA microarray data and their classification. We also present the Support Vector Machine (SVM) classifier, the firefly algorithm, the opposition-based learning concept and the three filtering methods used in the proposed hybrid approach.

2.1. Microarray dataset

In recent years, DNA microarray technology has been exceptionally well developed and is attracting significant interest from the scientific community. It is used in several applications in the field of genetics such as DNA sequencing, gene transcription studies, DNA-protein interaction studies and medical diagnosis. A DNA chip is a small surface (glass, silicon or plastic) subdivided into thousands of cells on which are fixed DNA molecules, classified and arranged in the form of a grid. The utilization of DNA chips enables the concurrent assessment of the expression levels of a significant quantity of genes (tens of thousands) in a biological sample under specific experimental conditions through the process of nucleic acid hybridization. These gene expression profiles make it possible to obtain an overall view of the cell and to capture anomalies. They can also serve as input for a data analysis system, such as in the diagnosis of cancer [34].

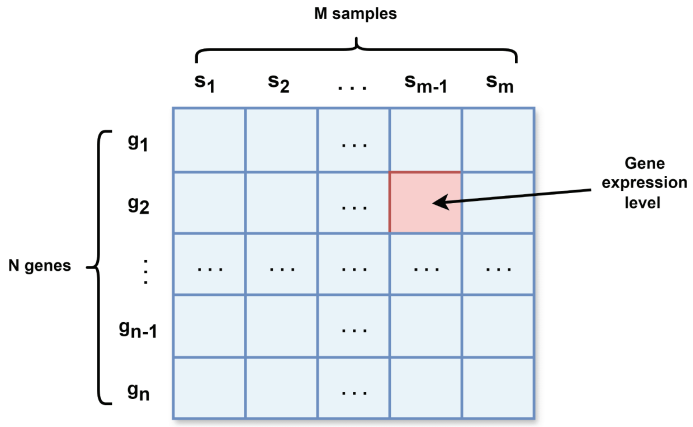

FIGURE 1. Structure of DNA microarray data matrix.

The analysis of DNA microarray data assists scientists in examining the activity of different genes in an organism and supports biologists understand the mechanisms of certain pathologies, particularly in the study of cancer. With known developments in microarray technology, researchers will be able to distinguish and classify cancer types based on gene expression values in tumor cells [29]. In addition, it will be possible to track the activity of a drug, such as an anticancer agent, on a group of patients, as well as to analyze the impact of a treatment on gene expression.

A microarray dataset contains data produced by the DNA chips, standardized and presented in the form of a matrix, with rows indicating genes and columns denoting instances, and every single cell carrying the expression level of a specific gene in one of the instances (see Fig. 1) [1]. In this study, we use twelve microarray datasets concern several types of cancer.

2.2. Support Vector Machine (SVM)

SVM is a commonly used discriminative classifier in data learning, originating from the statistical learning theory developed by Vladimir Vapnik [35]. SVM creates a decision function, also known as a prediction model, from a training dataset that categorizes new input data into corresponding classes. Initially designed for binary classification, SVM can be adapted to handle multiple classes and regression tasks.

SVM is based on two main principles: maximized margin and the integration of a kernel function. The concept of maximum margin involves identifying the hyperplane that separates positive and negative examples with the greatest distance from the separating boundary to the nearest examples, known as support vectors [36]. When dealing with non-linearly separable data, SVM utilizes kernel functions to convert the data into a higher-dimensional space where linear separation is attainable.

In this study, the Radial Basis Function (RBF) kernel is utilized, with its parameters (C and γ) optimized using an iterative search technique. It has been demonstrated by prior studies that the kernel function choice and optimization of its parameters greatly impact the performance of the SVM model obtained [37].

The choice of SVM as a classification algorithm is explained by their ability to work with large data and by the great success of this learning method in different fields of applications and especially in the classification of microarray data, as well as due to the richness of its theoretical foundation.

2.3. Firefly Algorithm

The Firefly optimization algorithm is an innovative bioinspired metaheuristic method, introduced by Xin-She Yang in 2008 [32]. This algorithm draws inspiration from the flashing behavior and mutual attraction of fireflies. Its functioning is predicated on three fundamental principles:

- All fireflies are considered unisex, which implies a mutual attraction independent of sex.
- The attractiveness between fireflies is directly related to their brightness; a firefly of lesser luminosity will gravitate towards a more luminous counterpart when it emits light. However, the attractiveness and brightness decrease as the distance between them increases. When two fireflies have equal brightness, their movement is determined randomly.
- The brightness of a firefly is dependent on the value of the objective function that is being optimized. In scenarios involving maximization, brightness corresponds proportionately to the value of this function.

For two fireflies X_i and X_j, The attractiveness β is defined as a function of the inter-firefly distance r_{ij} by:

$$\beta = \beta_0 \exp\left(-\gamma r_{ij}^2\right) \tag{1}$$

where β_0 denotes the level of attractiveness at a distance of zero and γ is a constant indicating the light absorption coefficient. The distance can be defined as a Cartesian distance by the formula (2).

$$r_{ij} = \|X_i - X_j\| = \sqrt{\sum_{k=1}^{d} \left(X_i^k - X_j^k\right)^2} \tag{2}$$

where X_i^k represents the kth element of the position vector associated with the firefly X_i in a d-dimensional space. When a firefly X_i moves towards a brighter firefly X_j, its position is adjusted according to the following mathematical expression:

$$X_i^{(t+1)} = X_i^t + \beta_0 \exp\left(-\gamma r_{ij}^2\right)\left(X_j^t - X_i^t\right) + \alpha(\text{rand} - 0.5) \tag{3}$$

where X_i^t and X_j^t, respectively, represent the current positions of fireflies X_i and X_j at iteration t, while $X_i^{(t+1)}$ corresponds to the new position of firefly X_i. The inclusion of the term $\alpha(\text{rand} - 1/2)$, where rand is a random number in the range [0,1] and α is the randomization parameter, adds a stochastic component to the position of firefly [33]. Figure 2 presents a flowchart that summarizes the primary steps of firefly algorithm. In this study, we have elected to utilize a binary version of the firefly algorithm as a search procedure in the gene selection process. This choice is driven by the algorithm's efficiency, straightforward implementation, and limited number of parameters. Furthermore, the firefly algorithm, as a population metaheuristic, allows for the efficient exploration of the search space while leveraging promising solutions.

2.4. Opposition Based-Learning

Opposition-Based Learning (OBL) is an optimization strategy that exploits the principle of complementarity to improve the performance of algorithms by considering both a potential solution and its opposite. This method evaluates the original solution and its opposite to select the best one, thus increasing the chances of finding optimal solutions and accelerating convergence. This strategy is particularly useful in complex and high-dimensional search spaces, as it facilitates a more effective examination of the search domain and helps to avoid local optima [38].

In the context of gene selection in microarray data, configurations are typically represented as binary vectors, with each element denoting the presence or absence of a specific gene within the configuration. Let $s = (s_1, s_2, \ldots, s_d)$ be a d-dimensional vector that represents a configuration, where d corresponds to the

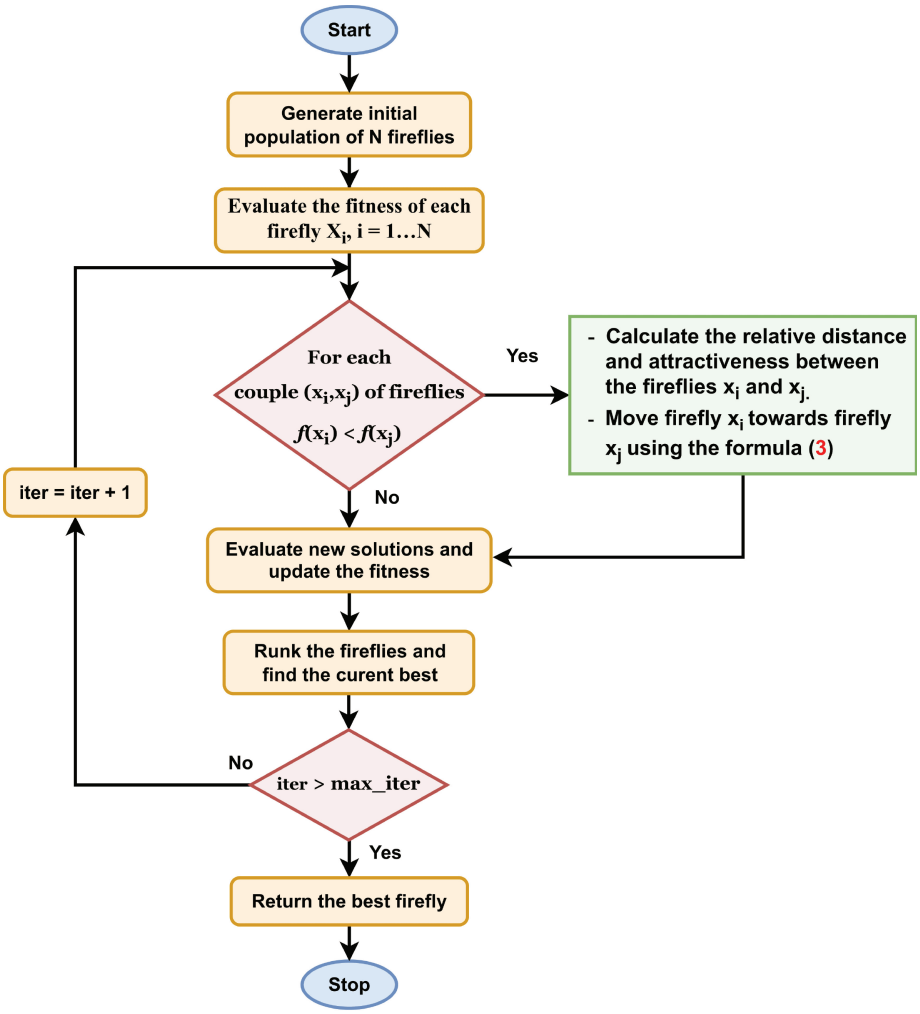

FIGURE 2. Firefly algorithm flowchart.

total number of genes, and $s_j \in \{0,1\}$, for $j = 1, 2, \ldots, d$. The complementary configuration can be articulated using Type-I opposition as a vector $\tilde{s} = (\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_d) \in \{0,1\}^d$ [39] where:

$$\tilde{s}_j = 1 - s_j, \quad j = 1, 2, \ldots, d \tag{4}$$

In binary optimization problems, it has been theoretically proven that a candidate solution s and its inverse $\tilde{s}$ are more likely to be close to the ideal solution compared to the first and second random candidate solutions [38]. Based on this, this work improves the performance of the Firefly Algorithm by including Opposition-Based Learning (OBL) during initialization and population updating.

2.5. Filter methods

In microarray data, a large number of irrelevant genes are present. As a result, gene ranking algorithms based on filtering are frequently utilized for preliminary gene selection. These filtering methods choose genes regardless of the classification algorithm used. They use the general characteristics of the data (usually statistical measures

like correlation, entropy, distance and consistency) to evaluate a single gene or a subset of genes. Filter methods are generally faster than wrapper methods in gene selection; however, they tend to select subsets containing a large number of genes. Consequently, a threshold is imperative for the selection of a more compact gene subset. Various univariate and multivariate filtering techniques are commonly employed for the purpose of feature selection within microarray datasets. These include correlation-based feature selection (CFS), Minimum Redundancy Maximum Relevance (mRMR), ReliefF, information gain and INTERACT. The filtering techniques applied in this study are as follows:

2.5.1. Information gain ratio

Information gain is a univariate filtering method widely used in microarray data analysis to identify informative genes. This technique evaluates the amount of "information" that a particular feature imparts regarding the target class [40]. It quantifies the dependency between a feature and the target variable in a classification task. The feature, with the highest information gain, will be preferred over other features. Information gain is measured by the reduction in entropy obtained by segmenting samples according to a specific feature X. The additional information about the class C given by a feature X (information gain) is computed by:

$$IG(X, C) = H(C) - H(C \mid X) \tag{5}$$

where $H(C)$ represents the entropy of class C, while $H(C \mid X)$ corresponds to the conditional entropy, which quantifies the amount of information needed to determine the values of class C when the distribution of values of feature X is known.

The information gain ratio operates as an enhancement of information gain, normalizing the latter by incorporating the intrinsic information associated with a feature. This normalization corrects the bias inherent in the information gain, which have a tendency to favor features with a greater number of levels. The formulation of the information gain ratio is delineated as follows:

$$IGR(X, C) = \frac{IG(X, C)}{H(X)} \tag{6}$$

where $H(X)$ is the entropy of the feature X, which represents the intrinsic information of the feature.

2.5.2. ReliefF

The ReliefF algorithm is a widely used method in machine learning for feature selection. Developed by Kononenko [41] as an extension of the Relief algorithm proposed by Kira and Rendell [42], ReliefF is specifically tailored to multi-class problems. This algorithm works by assigning weights to features and then selecting those whose weight exceeds a certain threshold. The weight assigned to each feature indicates its effectiveness in clustering data points within the same class and distinguishing them from those in different classes. The algorithm evaluates each feature according to its capability to correctly classify the nearest neighbors of each instance in the dataset.

Compared to the Relief algorithm, the ReliefF algorithm can handle all types of features and multiclass problems, it also considers feature dependencies and it is more robust to large and noisy data.

2.5.3. Correlation-based feature selection (CFS)

The CFS filtering method is a simple multivariate technique that identifies an optimal subset of attributes using a merit metric. This metric considers the correlation between attributes and the target class, as well as the intercorrelations among the attributes themselves [43]. The irrelevant features will be ignored as their correlation with the class will be weak. Redundant attributes will be automatically excluded because they will be highly correlated with other attributes. CFS measures the relevance M_s (Merit) of a subset of attributes S that includes k attributes, taking into account both their relevance and redundancy, using the following evaluation function:

$$M_s = \frac{k\overline{\rho_{cf}}}{\sqrt{k + k(k-1)\overline{\rho_{ff}}}} \tag{7}$$

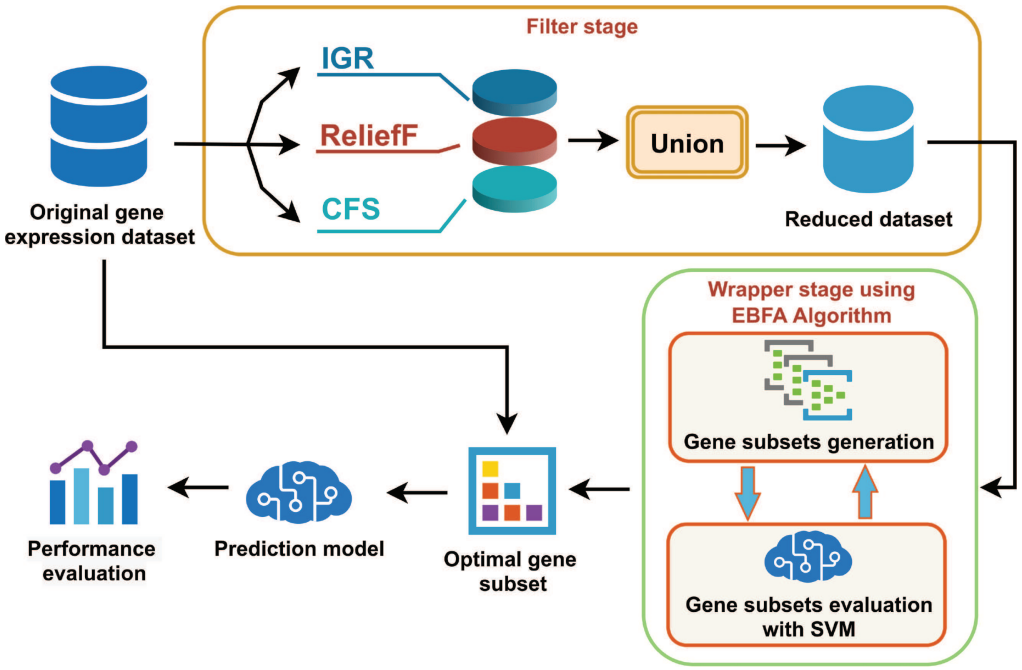

FIGURE 3. The flowchart of the methodology proposed for microarray data classification.

where $\overline{\rho_{cf}}$ represents the average of the correlations between attributes and the class (for $f \in S$), and $\overline{\rho_{ff}}$ denotes the average of the cross-correlations between attributes. We can think that the numerator of this equation is an indicator of the class's prediction by an attribute subset, while the denominator gives information about redundancy between attributes.

3. PROPOSED APPROACH FOR GENE SELECTION AND CANCER MICROARRAY DATA CLASSIFICATION

Gene selection can be viewed as a combinatorial optimization problem, wherein the objective is to identify the ideal set of relevant genes from an initial collection. Theoretically, if the initial dataset has n genes, we must consider testing $2^n - 1$ gene subsets to know the best one, which is practically impossible (NP-hard problem). The proposed gene selection method for microarray datasets integrates both filter and wrapper approaches in a two-stage to obtain a hybrid gene selection method. First, the initial number of genes is reduced by an ensemble of filter methods, and then a wrapper method utilizing an enhanced binary firefly algorithm is employed in order to explore the relevant gene subsets among the genes previously found during the filtering stage. A rigorous performance evaluation of the predictive model, trained on the returned optimal gene subset, is conducted at the end of the protocol. This experimental phase aims to test the effectiveness of the proposed hybrid approach in the microarray data classification task. The flowchart in Figure 3 presents the different steps of the proposed hybrid method.

3.1. Filtering phase

During the first step, a filtering process is applied to the original microarray dataset to take out irrelevant genes. Three filter-based methods are employed: Information Gain Ratio (IGR), ReleifF and Correlation-based feature selection (CFS).

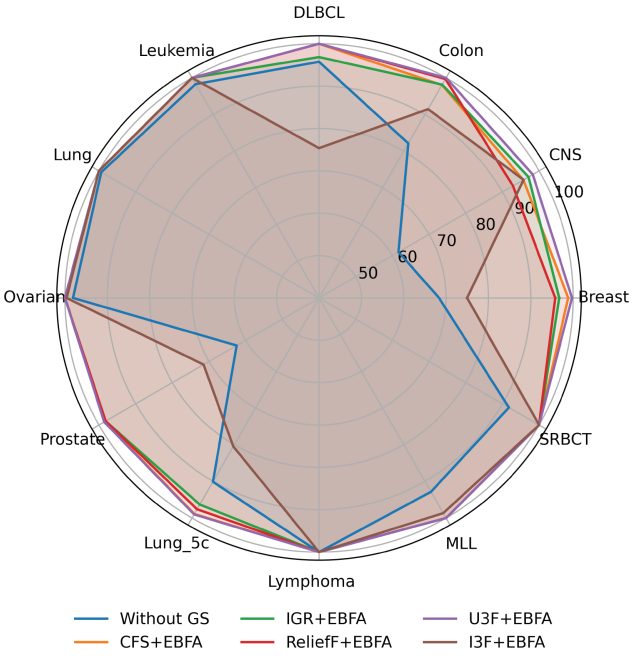

FIGURE 4. Comparison of average accuracy before and after selecting genes using the different proposed methods.

Determining an optimal threshold for excluding non-informative features remains a recurring methodological challenge in filtering techniques. Previous research on similar microarray datasets offers useful empirical guidance. For example, Pashaei and Pashaei [44] used the mRMR method on subsets of 10–150 genes and found that the top 100 genes worked well for classification on all tested datasets. Guyon *et al.* [45] report that 64 genes were enough to make strong classifications with an SVM on the Colon and Leukemia databases. Additional recent research [46] shows that the optimal performance on Colon, Leukemia, Lung, and Prostate datasets was achieved with only 16–24 genes with 16–24 genes. Based on these empirical results, our approach, for both IGR and ReliefF filters, selects the top 100 genes, as in the work [44].

Each of the three filtering methods is independently applied to the dataset to assess gene relevance. Consequently, three distinct gene subsets will be generated. To take advantage of the potential benefits of each method, the resulting subsets are further combined using union and intersection operations. The combined gene sets as well as the individual gene sets are subsequently processed through a wrapper approach based on the Enhanced Binary Firefly Algorithm (EBFA). This experimental design yield to five gene selection variants: IGR+EBFA, ReliefF+EBFA, CFS+EBFA, U3F+EBFA (representing the union of the three filter-based subsets) and I3F+EBFA (representing their intersection).

The filtering step aims to reduce the dimensionality of the search space, which helps to decrease the computational time required to apply the wrapper method in the next gene selection phase. Moreover, this preprocessing step helps to prevent overfitting, a common issue in datasets with limited sample sizes.

3.2. Wrapping phase

In the second step, a wrapper method for gene selection is proposed for processing the reduced dataset obtained from the first step. For this, an enhanced binary version of firefly algorithm (EBFA) is proposed; where a search is conducted in the space of genes by a population of fireflies (each firefly represents a candidate gene subset).

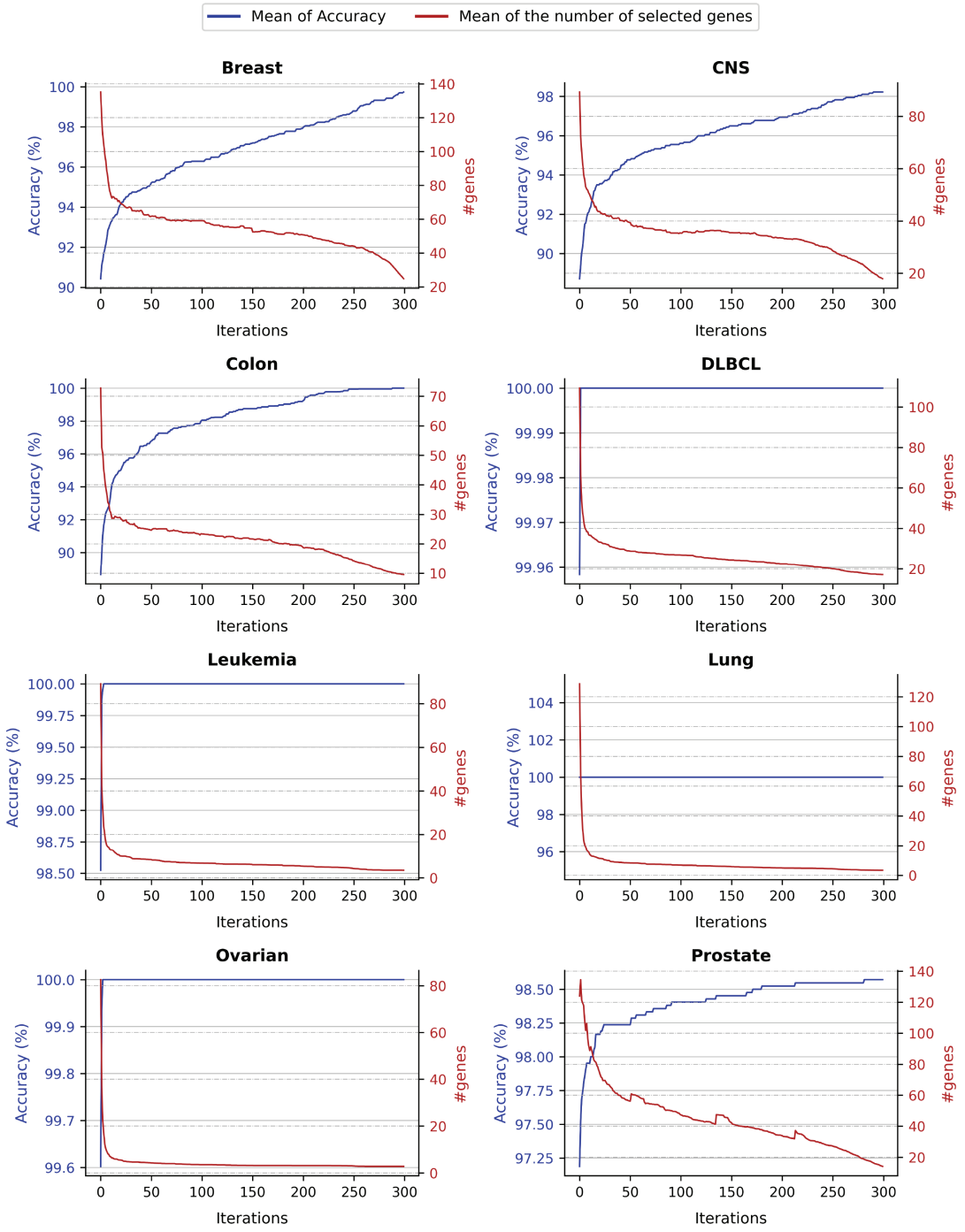

FIGURE 5. Convergence graphs of U3F+EBFA method in all eight binary class microarray datasets, in terms of average classification accuracy and average number of selected genes over 30 independent runs.

In the Firefly algorithm, fireflies move toward the brightest ones representing the best solutions, implementing a robust exploitation mechanism capable of accelerating convergence to local optima. To avoid the premature convergence problem and improve the balance between exploration and exploitation, we integrate two complementary strategies: Opposition-Based Learning (OBL) and an adaptive mutation operator.

OBL is used at two critical stages: when the first population is created as detailed in Section 3.2.1 and at each update of firefly as indicated in Section 3.2.4. In each case, the algorithm looks at both the current candidate and its opposite and picks the one that is better. This creates a double exploration of the search space, promoting greater initial diversity and reducing the risk of trapping in local minima.

Additionally, at the end of each iteration, we apply an adaptive binary mutation. The mutation starts with a probability of P_0 and decays over successive generations. This strategy ensures significant genetic diversity during early stages, while enabling progressive stabilization as the search evolves. The mutation operator can also reactivate stagnated dimensions–those that would otherwise remain unchanged due to the deterministic attractiveness-based movement rule.

These improvements (OBL and decay mutation) provide more adaptive and dynamic control over the trade-off between exploration and exploitation, improving the algorithm's ability to escape local optima while maintaining convergent behavior.

The flow of the algorithm is described as follows: After loading the data and specifying the parameters, the algorithm generates an initial population of fireflies randomly, incorporating an opposition-based learning strategy. The quality of each generated gene subset is evaluated by a function as described in the Section 3.2.2. Following the evaluation phase, an iterative process begins. At each iteration, pairwise comparisons are performed between fireflies. If one firefly is lower in brightness than another, it is attracted to its brighter counterpart based on a relative attractiveness measure calculated using Equation (1). Firefly movement is a two-step process involving an attraction step and a mutation step, as detailed in Section 3.2.4. To accommodate the binary representation, the normalized Hamming distance is employed as a proximity measure for attractiveness calculation (see Sect. 3.2.3). After movement, a population update is performed using an opposition-based selection strategy. This iterative process continues until the termination criterion is satisfied, with the global best solution being updated simultaneously. Algorithm 1 presents the detailed pseudo-code of the EBFA algorithm, describing its procedural steps.

3.2.1. Solutions representation and initialization using OBL method

A binary encoding scheme was employed to represent gene subsets within the search space. Each gene subset (firefly) is encoded as a binary vector whose number of bits equals the number of genes in the dataset obtained after the filtering process applied on the original microarray dataset. A binary value of 1 signifies the inclusion of a gene in the subset, while 0 denotes its exclusion.

An opposition-based initialization strategy is used to generate the initial firefly population. A preliminary population, P_0, is randomly generated with binary strings representing gene subsets. Subsequently, an opposite population, OP_0, is created using the opposition-based learning principle. The final initial population is constructed by selecting the best N fireflies from the union of two populations ($P_0 \cup OP_0$), as ranked by the objective function defined by the formula (8).

3.2.2. Evaluation Function

The goal of the EBFA algorithm, when looking for the ideal-quality gene subset is to enhance the classifier's accuracy while simultaneously minimizing gene subset size, making it a bi-objective optimization problem. To accommodate this, we use a linear weighted method to combine the two objectives into a single fitness function as defined in the formula (8). Where, "acc" represents the accuracy given by the classification algorithm and the parameter α is a weight factor that takes a value between 0 and 1. A higher value of α prioritizes accuracy, potentially at the expense of reducing the number of genes, while a lower value of α emphasizes the size of the feature subset, potentially compromising classification performance.The fitness function is assigned as the brightness value of a firefly in the EBFA algorithm.

Algorithm 1 The Enhanced Binary Firefly Algorithm (EBFA) for gene selection.

Input: Microarray dataset; N (number of fireflies); max_iter (maximum number of iterations); β_0 (attractiveness at $r_{ij} = 0$); γ (absorption coefficient); P_{m0} (initial mutation probability).
Output: Best gene subset selected.
Begin
Generate initial population P_0 of fireflies randomly (a firefly = gene subset) and calculate the opposite of the initial population (OP_0).
Evaluate the fitness of each firefly in ($P_0 \cup OP_0$) using the formula (8) and select the N fittest from them.
Set iter = 0.
while iter < max_iter **do**
  for $i = 1$ to N **do**
    for $j = 1$ to N **do**
      if brightness$_j$ > brightness$_i$ **then**
        Calculate the attractiveness β using formula (1) with normalized Hamming distance r_{ij} as in formula (10).
        Move firefly i towards firefly j by changing the bits of firefly i using formula (11).
        Mutate the bits of firefly i with probability P_m using formula (12).
      end if
    end for
    Evaluate firefly i and its opposite, and choose the best one to update the current population.
  end for
  Sort the current population and identify the best firefly.
  iter = iter + 1.
  Update the mutation probability P_m using formula (13).
end while
Return the global best firefly (the best gene subset).
End

The main objective of the proposed approach is to improve the classification accuracy with fewer genes. Thus, during the evaluation process, if two subsets of genes have the same classification accuracy, the one with fewer genes is retained. For this reason, in our experimental study, more importance was given to classification accuracy (with $\alpha = 0.98$).

$$\text{Fitness} = \alpha \cdot \text{acc} + (1 - \alpha) \cdot \left(1 - \frac{\text{Number of selected genes}}{\text{Total number of genes}}\right) \tag{8}$$

The classification accuracy, denoted by "acc" is determined by the ratio of the number of accurately classified samples to the total number of samples, according to formula (9). In this research, the performance of each gene subset (firefly) is assessed using an SVM classifier and a rigorous 10-fold cross-validation methodology.

$$\text{acc} = \frac{\text{Number of correctly predicted samples}}{\text{Total number of samples}} \tag{9}$$

3.2.3. The attractiveness of fireflies

The normalized Hamming distance is employed to determine the attractiveness among two fireflies X_i and X_j, as per the following expression:

$$r_{ij} = 1 - \frac{\sum_{k=1}^{d} \left(X_i^k \oplus X_j^k\right)}{d} \tag{10}$$

where $\oplus$ represents the XOR operation, and d denotes the dimension of the position vector. The Hamming distance quantizes the number of different bits between the two position vectors of fireflies X_i and X_j. The attractiveness β between the two fireflies X_i and X_j is then determined according to the distance of Hamming r_{ij} by applying formula (1).

3.2.4. Movement of a firefly

The standard Firefly Algorithm is primarily designed for continuous optimization problems. To facilitate the transition from continuous to binary representation, a modified firefly movement mechanism was implemented, similar to the approach used in Zhang *et al.* [47]. The update of each binary component of a firefly X_i when it moves to another firefly X_j, involves an attraction step, governed by the attractiveness parameter β as described by formula (11), and a mutation step, controlled by the mutation probability P_m as given by formula (12). The mutation step allows the algorithm to explore different areas of the research space, thus increasing the diversity of solutions. This adaptation facilitates the application of the algorithm to gene selection in microarray data.

$$X_i^k = \begin{cases} X_j^k & \text{if } X_i^k \neq X_j^k \text{ and rand(0,1)} < \beta \\ X_i^k & \text{otherwise} \end{cases} \tag{11}$$

$$X_i^k = \begin{cases} 1 - X_i^k & \text{if rand(0,1)} < P_m \\ X_i^k & \text{otherwise} \end{cases} \tag{12}$$

The mutation probability parameter, P_m, governs the random behavior of firefly movement. This parameter is dynamically calculated at each iteration of the EBFA algorithm according to the following equation:

$$P_m = P_{m0} \cdot \left(1 - \frac{\text{iter}}{\text{max\_iter}}\right) \tag{13}$$

where P_{m0} is the initial mutation probability taken as an input parameter for the EBFA algorithm. The dynamic adjustment of P_m during the optimization process is governed by the current iteration count (iter) and the maximum iteration count (max_iter). A higher P_m value in the early stages of the algorithm promotes exploration of the search space, while its gradual decrease facilitates convergence towards optimal solutions. In the experimental part, the value of the parameter P_{m0} is determined empirically by varying P_{m0} from 0.1 to 0.6 with a step of 0.1, and the best value found is $P_{m0} = 0.3$ (corresponding to 300 iterations).

At each iteration, the population of fireflies is updated by integrating an opposition-based learning strategy. For each solution (firefly), its opposite is calculated according to the formula (4). The next population is then formed by retaining, the solution offering the best value of the objective function between the present firefly and its opposite. This strategy effectively expands the search space and mitigates the risk of premature convergence to suboptimal solutions.

4. Experiments and results

The proposed methodology was implemented using Python within the Google Colab environment, which offers a convenient, cloud-based Jupyter notebook service.

4.1. Datasets

In order to assess the efficacy of the proposed methodology, we carried out a set of experiments using twelve widely utilized microarray datasets, obtained from sources [48], [49] and [50]. These datasets encompass binary datasets and multi-class datasets. Binary datasets include breast cancer, colon cancer, diffuse large B-cell lymphoma (DLBCL), ovarian cancer, lung cancer, central nervous system (CNS) cancer, leukemia, and prostate cancer. Multi-class datasets include Lung 5c, Lymphoma, MLL, and SRBCT. The used datasets are employed either to distinguish cancerous tissues from normal tissues or for the subtype classification of specific cancers. Table 1 provides the essential properties of the datasets, including the number of samples, genes, classes, and class distribution. For the prediction process using Support Vector Machines (SVMs), it is essential to normalize the datasets. Normalization ensures that features with larger numeric ranges do not overshadow those with smaller ranges and mitigates numerical difficulties during computation. To achieve this, the expression values of each gene are linearly scaled to the [0,1] range using Min-Max normalization.

TABLE 1. The datasets description.

Dataset	# Genes	# Instances	# Classes	Distribution
Breast	24481	97	2	46 benign samples and 51 malignant samples
CNS	7129	60	2	39 failure samples and 21 survivor samples
Colon	2000	62	2	40 normal samples and 22 tumor samples
DLBCL	7129	77	2	58 DLBCL samples and 19 FL samples
Leukemia	7129	72	2	25 AML samples and 47 ALL samples
Lung	12533	181	2	31 MPM samples and 150 ADCA samples
Ovarian	15154	253	2	91 normal samples and 162 cancer samples
Prostate	12600	136	2	59 normal samples and 77 tumor samples
Lung_5c	12600	203	5	17 samples of normal tissue, 139 samples of adenocarcinoma, 20 samples of pulmonary carcinoid, 21 samples of squamous carcinoma and 6 samples of small cell cancer
Lymphoma	4026	62	3	11 samples chronic lymphocytic lymphoma, 42 samples follicular lymphoma and 9 samples diffuse large B-cell lymphoma
MLL	12582	72	3	24 samples ALL, 20 samples MLL and 28 samples AML
SRBCT	2308	83	4	29 samples of Ewing sarcoma (EWS), 11 samples of Burkitt lymphoma (BL), 18 samples of neuroblastoma (NB) and 25 samples of rhabdomyosarcoma (RMS)

TABLE 2. Parameters of EBFA algorithm.

Parameters	Values
Population size (number of fireflies)	30
Maximum number of generations	300
β_0	1
γ (absorption coefficient)	1
Initial mutation probability	0.3
α (weight factor used in objective function)	0.98
Number of runs	30

4.2. EBFA parameter settings

The appropriate parameters of the EBFA algorithm were determined empirically through a series of experiments. The specific parameter settings employed in this study are tabulated in Table 2.

4.3. Evaluation measures

The performance of the suggested methodology was carefully assessed using a list of widely-used classification metrics. Specifically, the accuracy, recall, precision and F1-score rates were calculated for each dataset and the number of selected genes is also monitored. The accuracy rate measures the proportion of correctly classified samples as defined by the formula (9). The recall and precision metrics are expressed in terms of true negatives (TN), true positives (TP), false negatives (FN) and false positives (FP) by the formulas (14) and (15). Formula (16) presents the F1-score, a synthetic measure of model performance. This score is defined as the

TABLE 3. Number of genes after the filtering step.

Dataset	# Initial genes	# Genes after using filter methods				
		IGR	ReliefF	CFS	Union	Intersection
Breast	24481	100	100	138	271	3
CNS	7129	100	100	39	185	6
Colon	2000	100	100	26	151	10
DLBCL	7129	100	100	93	235	8
Leukemia	7129	100	100	81	189	19
Lung	12533	100	100	160	288	16
Ovarian	15154	100	100	35	177	12
Prostate	12600	100	100	75	250	1
Lung_5c	12600	100	100	550	699	2
Lymphoma	4026	100	100	229	344	23
MLL	12582	100	100	149	282	11
SRBCT	2308	100	100	111	218	20

harmonic mean of recall and precision.

$$\text{Recall} = \frac{TP}{TP + FN} \times 100 \tag{14}$$

$$\text{Precision} = \frac{TP}{TP + FP} \times 100 \tag{15}$$

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{16}$$

Given the non-deterministic nature of the EBFA algorithm and to guarantee the consistency of the results, each experiment was repeated thirty times. This approach provides more reliable estimates of performance and quantifies the uncertainty associated with the results.

4.4. Results and discussion

After performing the filtering phase, using the three data filtering techniques (IGR, ReleifF and CFS) and combining them using the union operator on the eight datasets, we obtain the number of genes selected by each technique shown in Table 3. In the second experimental phase, for each evaluation metric (accuracy, recall, precision, and F1-score), the minimum (Min), maximum (Max), mean, and standard deviation (Sd) values were calculated. The experimental results obtained using the five proposed variants (IGR filtering method with the EBFA algorithm (IGR+EBFA), ReliefF filtering method with the EBFA algorithm (ReliefF+EBFA), CFS filtering method with the EBFA algorithm (CFS+EBFA), the union of the three filtering methods followed by EBFA (U3F+EBFA) and the intersection of the three filtering methods followed by EBFA (I3F+EBFA)) are presented in Table 4. The average performance metrics for all datasets are also given for each variant of the proposed approach. The best performances obtained for each metric and each dataset are highlighted in bold.

As shown in the Tables 4 and the Figure 4, the performance of the I3F+EBFA variant was found to be inferior, particularly on the Prostate, DLBCL, Breast and Lung_5c datasets, where the average classification accuracies obtained were 71.48%, 75.36%, 74.97% and 80.49%, respectively. These less satisfactory results motivated the exclusion of this variant from the analyses and comparison presented in Tables 5, 6 and 7.

The graphs presented in Figures 5 and 6 illustrate the evolution of two key metrics (the average accuracy (in blue) and the average number of retained genes (in red)) over 300 iterations of the U3F+EBFA method on the

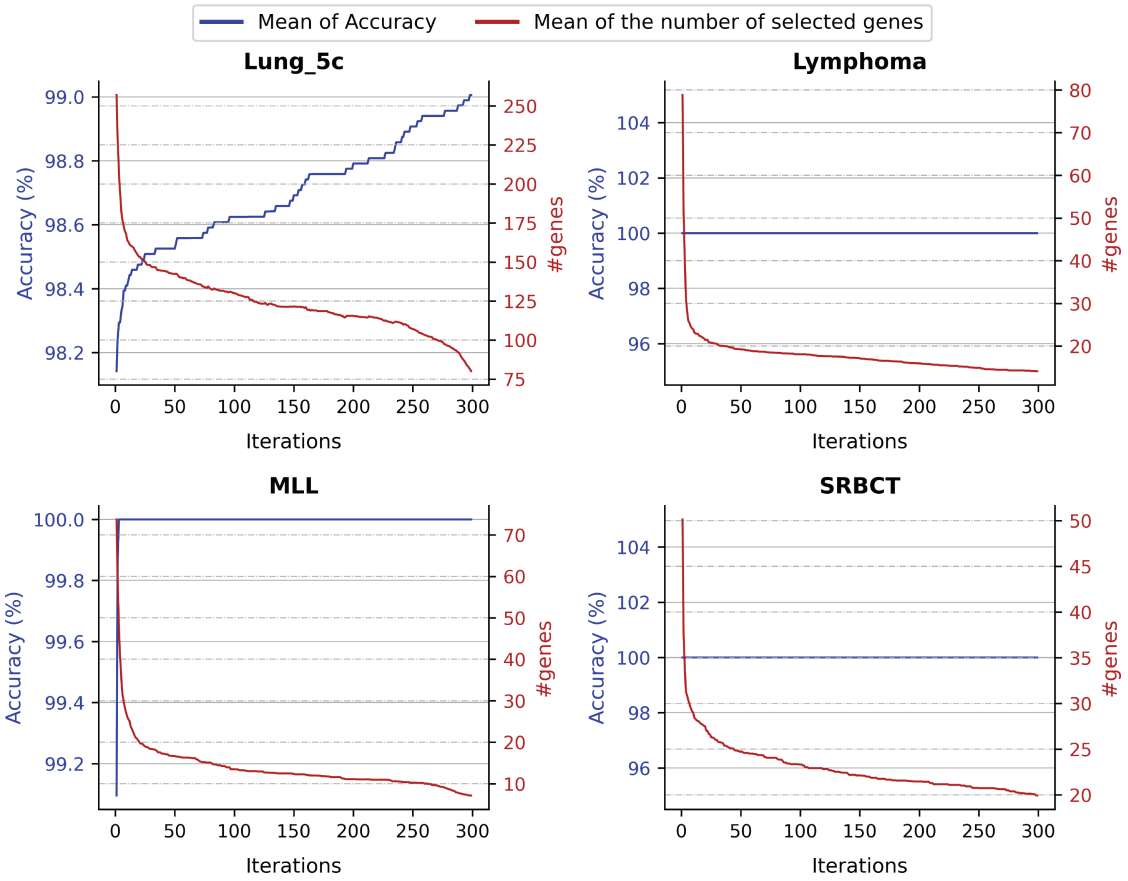

FIGURE 6. Convergence graphs of U3F+EBFA method in all four multi-class microarray datasets, in terms of average classification accuracy and average number of selected genes over 30 independent runs.

twelve test microarray datasets. For the Leukemia, Lung, Ovarian Lymphoma, MLL and SRBCT datasets, the classification accuracy quickly reaches 100% (within the first few iterations), indicating that the EBFA algorithm converges quickly to an optimal solution. On datasets such as Breast, CNS, Colon,Prostate and Lung_5c, the classification accuracy increases progressively over the iterations, suggesting a continuous improvement in model quality with the refinement of the selected gene subsets. The Prostate and Lung_5c datasets show a unique trend where the accuracy increases significantly after an initial period of relative stagnation, suggesting a more complex optimization process. Regarding data dimensionality reduction, for all datasets, we observe a strong reduction in the average number of selected genes from the first iterations. This reflects the efficiency of the algorithm in reducing dimensionality while maintaining or improving classification accuracy. The speed and extent of this reduction vary slightly across datasets. For example, the Breast Prostate and Lung_5c datasets show a continuous reduction in the number of genes over a larger number of iterations compared to the other datasets.

The U3F+EBFA method shows fast and efficient convergence for most datasets, achieving a balance between a substantial decrease in the number of genes and high classification accuracy. The bar chart in Figure 7 shows the average number of genes selected by the U3F+EBFA method on the twelve tested datasets, calculated over 30 independent runs. More complex datasets, such as Breast, CNS, Prostate, DLBCL, Lymphoma, SRBCT and Lung_5c, seem to require a larger number of genes (greater than 10) to maintain high performance, while others,

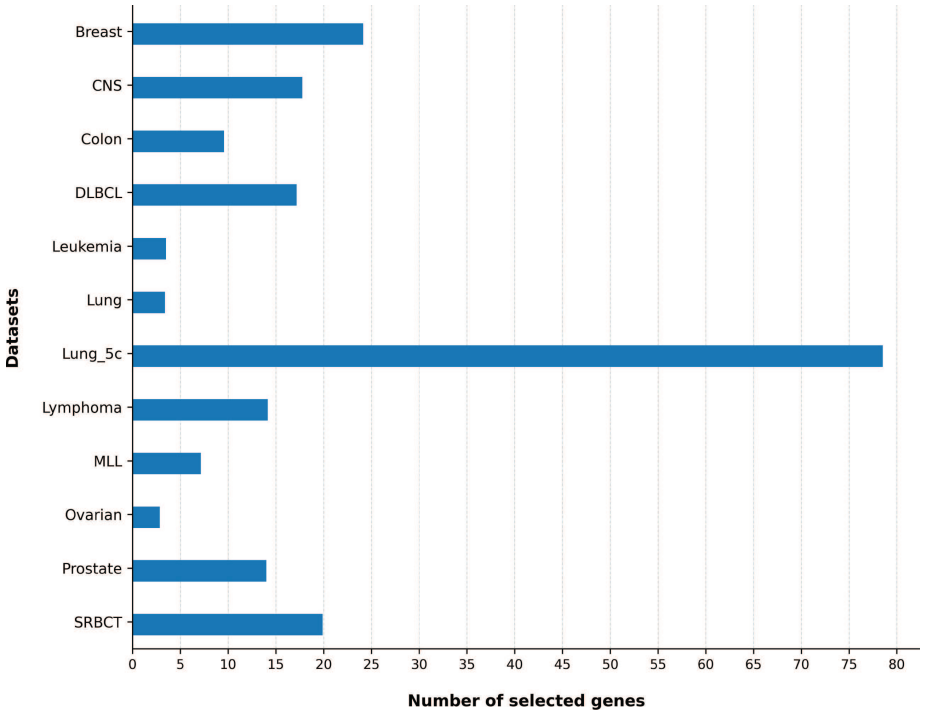

FIGURE 7. Average number of selected genes with U3F+ABFA method on different datasets over 30 independent runs.

such as Leukemia, Ovarian, and Lung, can be classified with a much lower number of genes (less than 4). These results indicate that an important number of genes in microarray datasets are inappropriate or redundant and should be excluded to improve classification accuracy.

4.4.1. A comparison between the four variants of the proposed approach

The results presented in Table 4 reveal a marked superiority of the U3F+EBFA method, which obtains the best performances on all metrics and datasets, frequently reaching perfect scores. On average, this method displays an accuracy of 99.57%, a precision of 99.82%, a recall of 99.20% and an F1-score of 99.39%. The integration of multiple filtering approaches prior to the application of the EBFA algorithm is shown to yield optimal outcomes, thereby underscoring the efficacy of this methodology. The CFS+EBFA method ranks second, with an average precision of 98.88% and an average F1-score of 98.23%. Although slightly inferior to U3F+EBFA, this method demonstrates notable robustness. The other two methods IGR+EBFA and ReliefF+EBFA, with respective average accuracies of 98.38% and 98.31%, show relatively comparable performances. However, these methods show a more pronounced variation in their performances between different datasets, suggesting an increased sensitivity to the specific characteristics of each dataset. The U3F+EBFA method is distinguished by a low variability of its results, as evidenced by the low standard deviations observed for all metrics. This stability indicates a high robustness of the approach, which produces reliable and reproducible results on different datasets. In addition, the CFS+EBFA method also exhibits good consistency, although slightly less marked than that of U3F+EBFA.

On specific datasets like DLBCL, Leukemia, Lung, Ovarian, Lymphoma, MLL and SRBCT, all methods achieve near-perfect or perfect scores, indicating that these datasets are relatively easier to classify regardless

TABLE 4. Experimental results obtained using IGR+EBFA, ReliefF+EBFA, CFS+EBFA, U3F+EBFA and I3F+EBFA methods.

Dataset	Method	Accuracy (%)				Precision (%)				Recall (%)				F1-score (%)			
		Min	Max	Mean	Sd	Min	Max	Mean	Sd	Min	Max	Mean	Sd	Min	Max	Mean	Sd
Breast	IGR+EBFA	92	**100**	96.7	2.06	90.83	**100**	98.02	2.13	88	**100**	95.52	2.83	91.47	**100**	96.34	2.24
	ReliefF+EBFA	91.78	99	95.76	1.61	91.5	98.33	96.59	1.81	89	**100**	94.88	2.53	91.28	99.09	95.37	1.85
	CFS+EBFA	95.89	**100**	98.84	1.13	94.67	**100**	98.87	1.5	95.5	**100**	98.88	1.29	95.64	**100**	98.76	1.19
	U3F+EBFA	**99.00**	**100**	**99.73**	**0.45**	**98.33**	**100**	**99.83**	**0.51**	**97.5**	**100**	**99.65**	**0.8**	**98.57**	**100**	**99.71**	**0.49**
	I3F+EBFA	73.32	76.47	74.36	1.51	73.97	77.55	75.16	1.42	72.63	76.72	74.00	1.77	72.45	76.22	73.72	1.80
CNS	IGR+EBFA	93.33	**100**	97.11	1.57	93.33	**100**	98.69	2	85	**100**	93.17	5.33	88.33	**100**	94.75	3.19
	ReliefF+EBFA	91.67	95	92.89	1.31	86.67	**100**	91.11	3.64	75	85	79.83	3.34	80	90	83.53	2.45
	CFS+EBFA	93.33	98.33	95.72	1.29	95	**100**	98.39	1.93	80	**100**	89.5	3.56	86.67	98	92.11	2.29
	U3F+EBFA	**96.67**	**100**	**98.28**	**0.82**	**96.67**	**100**	**99.81**	**0.75**	**90**	**100**	**95.17**	**2.45**	**93.33**	**100**	**96.66**	**1.6**
	I3F+EBFA	65.00	65.00	65.00	0.00	32.50	32.50	32.50	0.00	50.00	50.00	50.00	0.00	39.37	39.37	39.37	0.00
Colon	IGR+EBFA	96.67	**100**	98.17	1.01	93.33	**100**	98.33	2.1	95	**100**	97	2.49	94.67	**100**	97	1.58
	ReliefF+EBFA	96.67	**100**	99.68	0.88	96.67	**100**	99.67	1.02	95	**100**	99.61	1.21	94.67	**100**	99.56	1.27
	CFS+EBFA	96.67	98.33	98.06	0.63	96.67	**100**	99.44	1.26	95	95	95	**0**	94.67	96.67	96.33	0.76
	U3F+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	I3F+EBFA	85.38	93.46	92.66	2.44	85.85	93.78	93.09	2.11	82.00	93.00	92.02	3.01	82.59	92.73	91.77	2.93
DLBCL	IGR+EBFA	94.82	98.57	96.86	0.77	96.67	**100**	98.22	1.69	85	95	90.83	3.96	88	96.67	92.82	2.15
	ReliefF+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	CFS+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	U3F+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	I3F+EBFA	75.33	75.33	75.33	0.00	37.67	37.67	37.67	0.00	50.00	50.00	50.00	0.00	42.95	42.95	42.95	0.00
Leukemia	IGR+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	ReliefF+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	CFS+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	U3F+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	I3F+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
Lung	IGR+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	ReliefF+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	CFS+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	U3F+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	I3F+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
Ovarian	IGR+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	ReliefF+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	CFS+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	U3F+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	I3F+EBFA	99.61	99.61	99.61	0.00	99.47	99.50	99.49	0.01	99.69	99.70	99.69	0.00	99.58	99.58	99.58	0.00
Prostate	IGR+EBFA	97.14	98.57	98.18	0.41	97.78	**100**	98.85	0.46	96.25	98.75	98.1	0.74	97.41	98.75	98.38	0.38
	ReliefF+EBFA	97.86	98.57	98.17	0.36	98	98.89	98.82	0.23	97.5	98.75	98.13	0.64	98.08	98.75	98.38	0.33
	CFS+EBFA	97.14	**99.29**	98.45	0.73	97.64	98.89	98.77	0.35	96.25	**100**	98.67	1.23	97.41	**99.41**	98.64	0.68
	U3F+EBFA	**98.57**	98.57	**98.57**	**0**	**98.89**	98.89	**98.89**	**0**	**98.75**	98.75	**98.75**	**0**	**98.75**	98.75	**98.75**	**0**
	I3F+EBFA	71.38	71.38	71.38	0.00	77.03	77.03	77.03	0.00	73.52	73.52	73.52	0.00	70.40	70.40	70.40	0.00
Lung_5c	IGR+EBFA	95.56	96.54	96.33	0.28	97.91	98.16	98.11	0.07	90.72	94.05	93.25	1.04	93.50	95.77	95.22	0.72
	ReliefF+EBFA	97.51	98.02	97.61	0.19	93.45	99.44	97.24	2.17	92.52	98.08	95.19	1.52	93.26	97.45	95.87	1.67
	CFS+EBFA	**98.52**	**99.51**	98.89	0.32	**98.92**	**99.86**	99.51	0.35	**95.67**	**99.00**	97.60	0.71	**97.09**	**99.36**	98.32	0.50
	U3F+EBFA	**98.52**	**99.51**	**99.01**	**0.27**	**98.92**	**99.86**	**99.59**	**0.31**	**95.67**	**99.00**	**97.79**	**0.78**	**97.09**	**99.36**	**98.48**	**0.51**
	I3F+EBFA	79.30	85.24	80.49	2.42	42.64	56.88	45.49	5.79	48.00	61.38	50.68	5.44	44.98	58.78	47.74	5.61
Lymphoma	IGR+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	ReliefF+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	CFS+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	U3F+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	I3F+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
MLL	IGR+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	ReliefF+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	CFS+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	U3F+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	I3F+EBFA	98.67	98.67	98.67	0.00	98.67	98.67	98.67	0.00	98.89	98.89	98.89	0.00	98.65	98.65	98.65	0.00
SRBCT	IGR+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	ReliefF+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	CFS+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	U3F+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
	I3F+EBFA	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**0**
Average	IGR+EBFA	97.46	99.47	98.61	0.51	97.49	99.85	99.19	0.70	95.00	98.98	97.32	1.37	96.12	99.27	97.88	0.86
	ReliefF+EBFA	97.96	99.22	98.68	0.36	97.19	99.72	98.62	0.74	95.75	98.49	97.30	0.77	96.44	98.77	97.73	0.63
	CFS+EBFA	98.46	99.62	99.16	0.34	98.58	99.90	99.58	0.45	96.87	99.50	98.30	0.57	97.62	99.45	98.68	0.45
	U3F+EBFA	**99.40**	**99.84**	**99.63**	**0.13**	**99.40**	**99.90**	**99.84**	**0.13**	**98.49**	**99.81**	**99.28**	**0.34**	**98.98**	**99.84**	**99.47**	**0.22**
	I3F+EBFA	87.33	88.76	88.13	0.53	78.98	81.13	79.93	0.78	81.23	83.60	82.40	0.85	79.25	81.56	80.35	0.86

of the filtering method used. For more complex datasets like Breast, Prostate and Lung_5c, the differences are more pronounced, with U3F+EBFA and CFS+EBFA demonstrating superiority over the other approaches.

To more analyze the results, we use box plots to visualize the distribution of classification accuracy across the 30 executions performed on each tested dataset and for each variant of the proposed method. These graphical representations, presented in Figures 8 and 9, provide a visual summary of the results by highlighting the quartiles, the median and the mean (represented by a triangle) of the accuracy. This visualization allows us to more precisely understand the robustness and stability of the different methods in the face of random variations related to the replications. Box plot analysis reveals that the U3F+EBFA method outperforms the remaining methods with regard to robustness and stability. This method is distinguished by higher median performance, lower result dispersion, and fewer outliers. The CFS+EBFA method also performs well, but is slightly less stable. Comparing the IGR+EBFA and ReliefF+EBFA methods to the other two, however, shows less satisfactory results in terms of precision and consistency.

Table 5 provides a comparison of the best results obtained by the four variants of the suggested approach; focusing on the highest accuracy in classification, the number of selected genes and the dimensionality reduction rate for each dataset. The results show that, in terms of gene selection, the U3F+EBFA method performs better than alternative methods. It significantly decreases the size of the data while maintaining high classification accuracy (achieving 98.57% with 8 genes selected on the Prostate dataset, 99.51% with 74 genes selected on the Lung_5c dataset and reaches 100% accuracy for the other datasets with a number of selected genes ranging between 2 and 16). The CFS+EBFA, IGR+EBFA and ReliefF+EBFA methods also show strong performances but generally select a larger number of genes. On the Leukemia, Lung, Ovarian, Lymphoma, MLL and SRBCT datasets, all methods reach 100% accuracy with a very small number of genes selected. For the Colon, Breast, DLBCL, Prostate and Lung_5c datasets, the differences between the methods are less marked, but U3F+EBFA continues to stand out by its ability to choose a significantly reduced number of genes. All of the assessed methods deliver highly satisfactory classification results (with classification accuracies above 95%) with a significant reduction in data dimensionality, ranging from 99.31% to 99.99%.

4.4.2. A comparison of the proposed methods with the use of SVM without gene selection

In this part, we compare the outcomes of the four variants of the specified approach with the results obtained by a pure SVM on the twelve datasets tested. The goal is to show the impact of gene selection on the task of classifying microarray data. Figure 4 presents, in the form of a radar chart, the average of classification accuracies obtained over 30 iterations by an RBF kernel SVM classifier (with optimized parameters) applied to the twelve tested datasets, with and without a prior gene selection phase. Analysis of this chart reveals that the integration of gene selection methods, in particular the U3F+EBFA method, significantly improves the classifier performance on all datasets. The gains in classification accuracy are notable on the Prostate, Breast, CNS, Colon and Lung_5c datasets, where the U3F+EBFA method allowed an increase in accuracy of more than 36% on the Prostate dataset, more than 31% on the Breast dataset, more than 36% on the CNS dataset, more than 17% on the Colon dataset and more than 9% on the Lung_5c dataset. The proposed methods (IGR+EBFA, ReliefF+EBFA, CFS+EBFA and U3F+EBFA) outperform the approach without gene selection, and the U3F+EBFA method stands out for its ability to consistently identify the most discriminatory genes, thus leading to more robust and high performance. The I3F+EBFA method outperforms the approach without gene selection on the majority of datasets, except for DLBCL and Lung_5c. These results show the importance of dimensionality reduction to improve the generalization of classification models in bioinformatics.

4.4.3. A comparison with other methods

We performed comparative study with contemporary state-of-the-art approaches using identical benchmark datasets to further validate the efficacy of our suggested methodology in gene selection. Tables 6 and 7 present the best classification accuracies achieved and the corresponding number of genes selected for each method on the binary class and multi-class datasets respectively. The best performing results are highlighted in bold. The

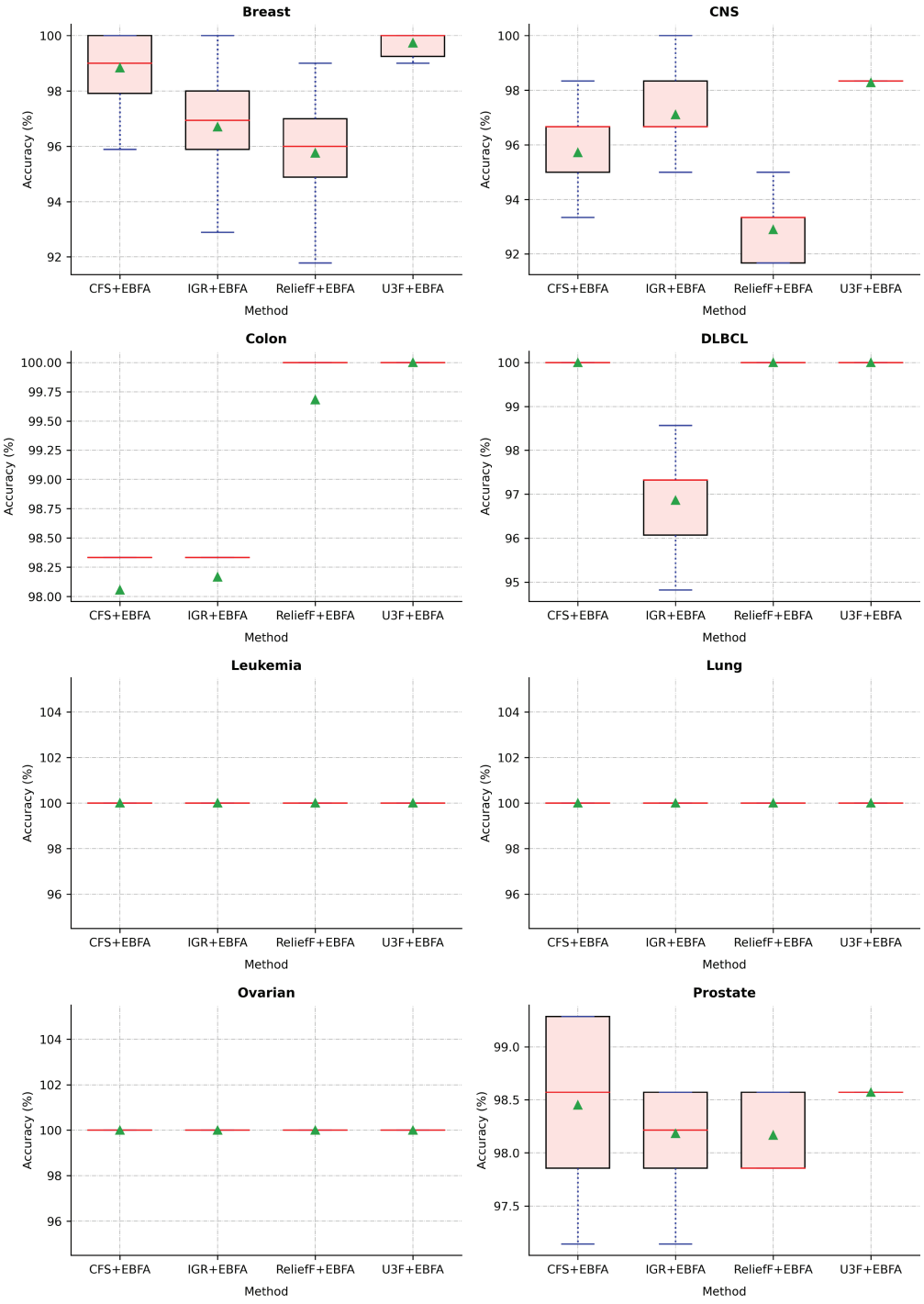

FIGURE 8. Distribution of classification accuracy over 30 runs for each method and for each binary class dataset.

TABLE 5. Best results obtained with the proposed methods.

Dataset	# Initial genes	Method	Best Acc (%)	# Selected genes	Reduction rate (%)
Breast	24481	IGR+EBFA	100.00	15	99.94
		ReliefF+EBFA	99.00	16	99.93
		CFS+EBFA	**100.00**	**14**	**99.94**
		U3F+EBFA	100.00	15	99.94
CNS	7129	IGR+EBFA	100.00	18	99.93
		ReliefF+EBFA	95.00	11	99.96
		CFS+EBFA	98.33	19	99.92
		U3F+EBFA	**100.00**	**12**	**99.95**
Colon	2000	IGR+EBFA	100.00	8	99.97
		ReliefF+EBFA	**100.00**	**7**	**99.97**
		CFS+EBFA	98.33	5	99.98
		U3F+EBFA	**100.00**	**7**	**99.97**
DLBCL	7129	IGR+EBFA	98.57	22	99.91
		ReliefF+EBFA	**100.00**	**11**	**99.96**
		CFS+EBFA	100.00	13	99.95
		U3F+EBFA	100.00	13	99.95
Leukemia	7129	IGR+EBFA	**100.00**	**3**	**99.99**
		ReliefF+EBFA	**100.00**	**3**	**99.99**
		CFS+EBFA	**100.00**	**3**	**99.99**
		U3F+EBFA	**100.00**	**3**	**99.99**
Lung	12533	IGR+EBFA	**100.00**	**2**	**99.99**
		ReliefF+EBFA	**100.00**	**2**	**99.99**
		CFS+EBFA	**100.00**	**2**	**99.99**
		U3F+EBFA	**100.00**	**2**	**99.99**
Ovarian	15154	IGR+EBFA	**100.00**	**2**	**99.99**
		ReliefF+EBFA	100.00	2	99.99
		CFS+EBFA	**100.00**	3	99.99
		U3F+EBFA	**100.00**	**2**	**99.99**
Prostate	12600	IGR+EBFA	98.57	11	99.96
		ReliefF+EBFA	98.57	10	99.96
		CFS+EBFA	**99.29**	**9**	**99.96**
		U3F+EBFA	98.57	8	99.97
Lung_5c	12600	IGR+EBFA	96.54	10	99.92
		ReliefF+EBFA	98.02	12	99.90
		CFS+EBFA	**99.51**	**63**	**99.50**
		U3F+EBFA	99.51	74	99.41
Lymphoma	4026	IGR+EBFA	**100.00**	**4**	**99.90**
		ReliefF+EBFA	**100.00**	**4**	**99.90**
		CFS+EBFA	100.00	12	99.70
		U3F+EBFA	100.00	13	99.68
MLL	12582	IGR+EBFA	**100.00**	**5**	**99.96**
		ReliefF+EBFA	**100.00**	**5**	**99.96**
		CFS+EBFA	100.00	6	99.95
		U3F+EBFA	**100.00**	**5**	**99.96**
SRBCT	2308	IGR+EBFA	**100.00**	**6**	**99.74**
		ReliefF+EBFA	100.00	9	99.61
		CFS+EBFA	**100.00**	**6**	**99.74**
		U3F+EBFA	100.00	16	99.31

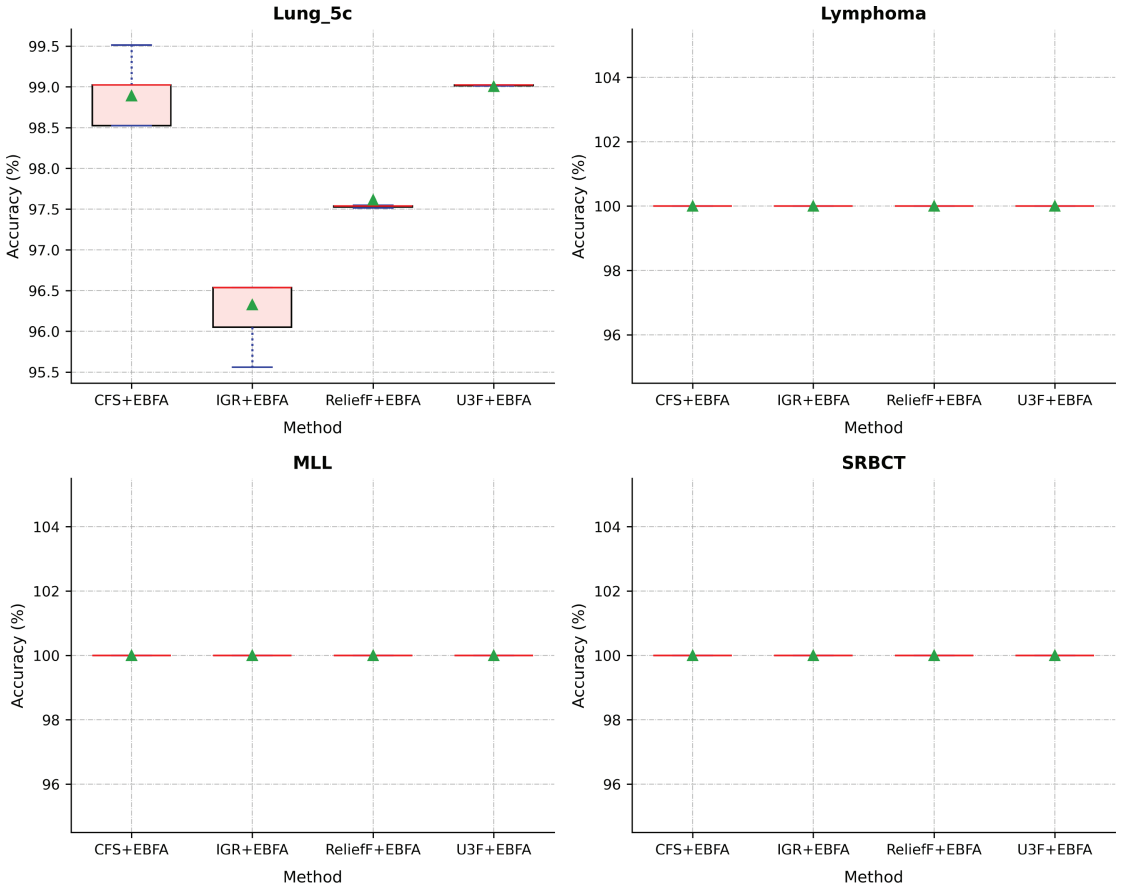

FIGURE 9. Distribution of classification accuracy over 30 runs for each method and for each multi-class dataset.

symbol '–' indicates that no information is available or the fact that the method was not applied to this specific dataset.

The results obtained convincingly prove the superiority of the suggested methodologies, in particular the U3F+EBFA method, generally surpassing the performance of existing approaches with regard to the best classification accuracy, often reaching classification accuracies close to 100% while selecting a significantly reduced number of genes (between 2 and 18 for binary class datasets and between 4 and 74 for muti-class datasets). This demonstrates a strong ability of this approach to identify the most relevant genes for classification. Although some existing methods achieved perfect classification on specific datasets, they often required a higher number of genes. For instance, on the Leukemia and Ovarian datasets, the method cited in Alrefai and Ibrahim [21] required 26 and 13 genes, respectively, to reach 100% accuracy, whereas the proposed method U3F+EBFA achieved the same accuracy with only 3 and 2 genes. Similarly, methods cited in [1,19,23,27,52,54,55] obtained perfect results on at least one dataset from Leukemia, Lung and Ovarian datasets, but with a higher number of selected genes than our proposed methods (which is only 2 or 3 genes). On the Breast and CNS datasets, the method U3F+EBFA demonstrated exceptional performance, achieving 100% accuracy with a small number of genes (15 and 12, respectively). which is very efficient compared to other existing methods that give accuracies between 78% and 98.33%. Even on the Prostate dataset, the proposed methods surpassed existing methods, reaching an accuracy of over 98.57%. For multiclass datasets, the proposed methods outperform other methods

TABLE 6. Comparison between the proposed methods and other approaches binary class datasets (in terms of best classification accuracy % and number of selected genes).

Method	Breast	CNS	Colon	DLBCL	Leukemia	Lung	Ovarian	Prostate
[1]	–	–	–	–	**100.00** (15)	–	–	96.3(14)
[21]	86.36 (45)	85.71 (49)	92.86 (41)	–	**100.00** (26)	–	**100.00** (13)	–
[23]	94 (10)	98.33 (10)	**100.00** (10)	–	–	–	**100.00** (10)	–
[27]	–	78.13 (2)	90.53 (5)	–	**100.00** (6)	**100.00** (5)	**100.00** (36)	–
[24]	–	–	93.46 (3)	–	97.14 (3)	–	–	–
[51]	–	–	96.90(3)	–	98.57(4)	–	**100.00** (4)	–
[19]	96.19 (10.8)	94.60 (5.2)	95.68 (7.66)	–	**100.00** (3.3)	–	**100.00** (2.6)	–
[26]	93.81 (618)	93.33 (182)	–	–	–	–	–	–
[52]	88.17 (10.2)	95.64 (6.7)	94.72 (5.3)	–	99.62 (5.2)	99.16 (5.6)	–	94.18 (7.8)
[53]	–	–	75.00 (4)	–	82.40 (6)	98.00 (3)	95.00 (3)	97.10 (3)
[54]	–	–	87 (19)	99 (4)	**100.00** (6)	–	–	–
[25]	–	78.00 (–)	–	97.9 (3)	98.6 (4)	–	–	–
[20]	–	–	93.58 (7)	96.67 (6)	95.90 (7)	–	–	–
[55]	81.3 (2036)	84.7 (502)	91.9 (127)	–	**100.00** (301)	97.5 (613)	**100.00** (355)	–
IGR+EBFA	**100.00** (15)	**100.00** (18)	**100.00** (8)	98.57 (22)	**100.00 (3)**	**100.00 (2)**	**100.00 (2)**	98.57 (11)
ReliefF+EBFA	99.00(16)	95.00 (11)	**100.00 (7)**	**100.00 (11)**	**100.00 (3)**	**100.00 (2)**	**100.00 (2)**	98.57 (10)
CFS+EBFA	**100.00 (14)**	98.33 (19)	98.33 (5)	**100.00** (13)	**100.00 (3)**	**100.00 (2)**	**100.00** (3)	**99.29 (9)**
U3F+EBFA	**100.00** (15)	**100.00 (12)**	**100.00 (7)**	**100.00** (13)	**100.00 (3)**	**100.00 (2)**	**100.00 (2)**	98.57 (8)

TABLE 7. Comparison between the proposed methods and other approaches for multi-class datasets (in terms of best classification accuracy % and number of selected genes).

Method	Lung_5c	Lymphoma	MLL	SRBCT
[1]	–	–	–	**100.00** (18)
[23]	97.52 (–)	**100.00** (10)	**100.00** (–)	**100.00** (10)
[27]	–	**100.00 (2)**	97.80 (27)	**100.00** (30)
[51]	94.56 (4)	99.87 (3)	–	97.77 (5)
[19]	–	**100.00 (2)**	**100.00 (3.6)**	**100.00** (5.8)
[26]	98.52 (327)	–	–	–
[55]	96.00 (25)	–	–	96.00 (4)
[52]	–	**100.00** (99)	–	**100.00** (52)
IGR+EBFA	96.54 (10)	**100.00** (4)	**100.00** (5)	**100.00** (6)
ReliefF+EBFA	98.02 (12)	**100.00** (4)	**100.00** (5)	**100.00** (9)
CFS+EBFA	**99.51 (63)**	**100.00** (12)	**100.00** (6)	**100.00** (6)
U3F+EBFA	**99.51** (74)	**100.00** (13)	**100.00** (5)	**100.00** (16)

in terms of accuracy and the CFS+EBFA method proves to be the best. For lymphoma, MLL and SRBCT datasets, the reference [34] achieves the ideal performance (100%) with a slightly lower number of genes than the proposed methods.

These findings underscore the effectiveness of our proposed methods, particularly U3F+EBFA, in selecting discriminating genes across various datasets and constructing efficient classification models. Furthermore, the results show that the hybridization of filter and wrapper approaches greatly enhances the robustness of the final selected gene subset.

TABLE 8. List of most frequently repeated genes selected by U3F+EBFA method for each tested dataset.

Dataset	Name/index of gene
Breast	NM_004953, Contig412_RC, Contig47544_RC, AL080059, NM_003079
CNS	844, 5844, 2917
Colon	1346, 1924
DLBCL	M37815_cds1_at, D78134_at, M14328_s_at
Leukemia	U77604_at, HG1612-HT1612_at, M23197_at, X95735_at, Y07604_at
Lung	34320_at, 33328_at
Ovarian	MZ436.63379, MZ435.46452
Prostate	38634_at, 37639_at, 37599_at
Lung_5c	613_at, 319_g_at, 40567_at, 318_at
Lymphoma	162, 3710, 3753, 3761
MLL	34306_at, 1065_at
SRBCT	1386, 1389, 255

4.4.4. Biological results validation

Gene selection from cancer microarray datasets allows the identification of characteristic molecular signatures associated with tumor phenotypes. These signatures can be valuable tools for medical experts for more accurate and personalized diagnosis. The stability of the proposed gene selection method, U3F+EBFA, is evaluated through the analysis of the overlap of selected genes in multiple runs (30 in our study). The indices or names of the most frequently selected genes are presented in Table 8. Heatmaps are widely used in biology to visualize the expression levels of many genes across different samples. They can be used to identify characteristic expression profiles, identify frequently regulated genes, or highlight molecular signatures associated with specific pathologies. Figure 10 presents the heatmaps of the discriminating genes identified in each dataset, as reported in Table 8, with the aim of analyzing differentially expressed genes. In these visualizations, each column corresponds to a sample, while each row represents a gene. Expression variations are coded by a color gradient: red indicates overexpression, blue underexpression, and white stable expression. In several datasets, such as colon, lung, ovarian, MLL and lymphoma, the expression profiles are discriminating and show a clear differentiation between patient subgroups. For the Leukemia dataset, two genes among the 3 genes listed in Table 8 are also found among the top selected genes in the study [45]. For a more in-depth study on the biological interpretation (or biological relevance) of the results found, we plan to integrate a pathway analysis in the future.

5. CONCLUSION

This work tackles the challenge of gene selection in cancer microarray datasets, which are distinguished by a very large number of genes and a limited number of instances. A hybrid filtering-wrapping approach is proposed to enhance the classification performance. The methodology starts with a multi-filtering preprocessing step that combines three feature selection techniques, followed by a wrapper method using an improved binary firefly algorithm to find the best collection of genes for classification. The methodology was evaluated on twelve cancer microarray datasets and compared to several recent methods that have been published in the literature. According to the test findings, the proposed methodology is competitive with alternative gene selection approaches based on filter and wrapper techniques. It achieves high accuracies (above 98.57% on all tested datasets) with a reduced number of selected genes, obtaining a dimensionality reduction rate above 99%. This method offers substantial utility to biologists by facilitating the identification of key biomarker genes associated with specific cancer types.

Future work could focus on reducing the computational time by implementing a parallelized version of the binary firefly algorithm employed in the wrapper method. Additionally, further improvements might be achieved

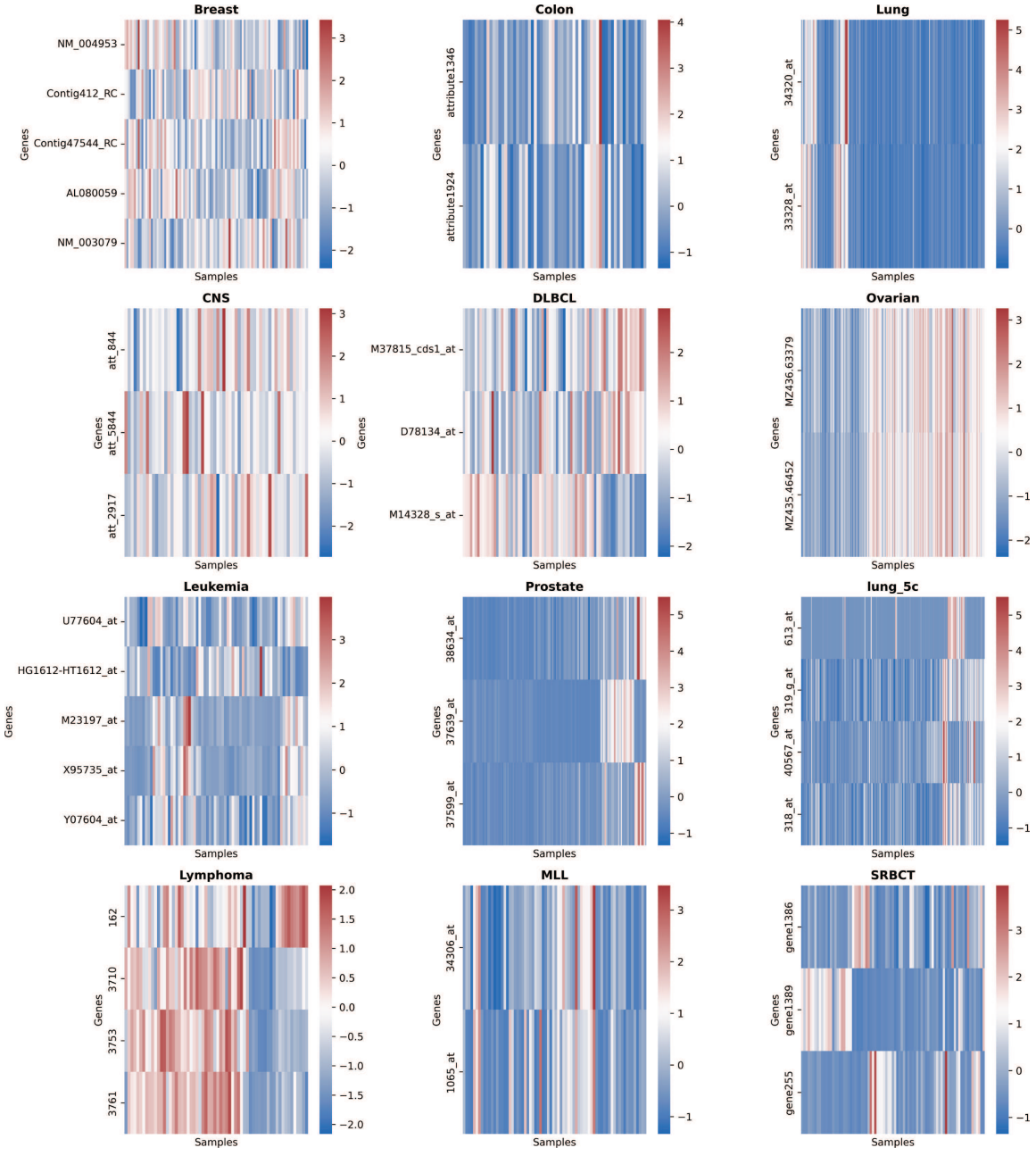

FIGURE 10. Heatmaps representing frequently selected genes by U3F+EBFA method for each dataset.

by integrating alternative filtering techniques and exploring the use of other classification algorithms, such as Random Forests or Artificial Neural Networks, to enhance the overall efficacy of the proposed approach.

Data availability statement

No new data/codes were created or analyzed in this study.

References

[1] M. Dashtban and Mohammadali Balafar, An integer-coded genetic algorithmgene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts. *Genomics* **109** (2017) 91–107.

[2] S. Almuhaideb and M. El Bachir Menai, Impact of preprocessing on medical data classification. *Front. Comput. Sci.* **10** (2016) 1082–1102.

[3] B. Duval and J.-K. Hao, Advances in metaheuristics for gene selection and classification of microarray data. *Briefings Bioinf.* **11** (2010) 127–141.

[4] V. Bolón-Canedo, N. Sánchez-Maroño and A. Alonso-Betanzos, Feature selection for high-dimensional data. *Prog. Artif. Intell.* **5** (2016) 65–75.

[5] K.K. Ghosh, S. Begum, A. Sardar, S, Adhikary, M. Ghosh, M. Kumar and R. Sarkar, Theoretical and empirical analysis of filter ranking methods: experimental study on benchmark dna microarray data. *Expert Syst. Appl.* **169** (2021) 114485.

[6] A.K. Shukla and D. Tripathi, Identification of potential biomarkers on microarray data using distributed gene selection approach. *Math. Biosci.* **315** (2019) 108230.

[7] L. Sun, J. Wang and J. Wei, Avc: Selecting discriminative features on basis of auc by maximizing variable complementarity. *BMC Bioinf.* **18** (2017) 3.

[8] G. Xu, M. Zhang, H. Zhu and J. Xu, A 15-gene signature for prediction of colon cancer recurrence and prognosis based on svm. *Gene* **604** (2017) 33–40.

[9] X. Song, Y. Zhang, D. Gong, H. Liu and W. Zhang, Surrogate sample-assisted particle swarm optimization for feature selection on high-dimensional data. *IEEE Trans. Evol. Comput.* **27** (2023) 595–609.

[10] M. Alzaqebah, K. Briki, N. Alrefai, S. Brini, S. Jawarneh, M.K. Alsmadi, R.M.A. Mohammad, I. ALmarashdeh, F.A. Alghamdi, N. Aldhafferi and A. Alqahtani, Memory based cuckoo search algorithm for feature selection of gene expression dataset. *Inf. Med. Unlocked* **24** (2021) 100572.

[11] X. Song, H. Ma, Y. Zhang, D. Gong, Y. Guo and Y. Hu, A streaming feature selection method based on dynamic feature clustering and particle swarm optimization. *IEEE Trans. Evol. Comput.* 2024.

[12] K. Chatra, V. Kuppili, D.R. Edla and A.K. Verma, Cancer data classification using binary bat optimization and extreme learning machine with a novel fitness function. *Med. Biol. Eng. Comput.* **57** (2019) 2673–2682.

[13] Y. Zhang, D. wei Gong, X. zhi Gao, T. Tian and X. yan Sun, Binary differential evolution with self-learning for multi-objective feature selection. *Inf. Sci.* **507** (2020) 67–85.

[14] Y. Hu, Y. Zhang and D. Gong, Multiobjective particle swarm optimization for feature selection with fuzzy cost. *IEEE Trans. Cybern.* **51** (2021) 874–888.

[15] S. Guo, D. Guo, L. Chen and Q. Jiang, A l1-regularized feature selection method for local dimension reduction on microarray data. *Comput. Biol. Chem.* **67** (2017) 92–101.

[16] C. Kang, Y. Huo, L. Xin, B. Tian and B. Yu, Feature selection and tumor classification for microarray data using relaxed lasso and generalized multi-class support vector machine. *J. Theor. Biol.* **463** (2019) 77–91.

[17] A.M. Alharthi, M.H. Lee and Z.Y. Algamal, Gene selection and classification of microarray gene expression data based on a new adaptive l1-norm elastic net penalty. *Inf. Med. Unlocked* **24** (2021) 100622.

[18] E.H. Houssein, H.N. Hassan, M.M. Al-Sayed and E. Nabil, Gene selection for microarray cancer classification based on manta rays foraging optimization and support vector machines. *Arabian J. Sci. Eng.* **47** (2022) 2555–2572.

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

[20] K. Yu, W. Li, W. Xie and L. Wang, A hybrid feature-selection method based on mrmr and binary differential evolution for gene selection. *Processes* **12** (2024) 2.

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

[22] G. Manikandan and S. Abirami, An efficient feature selection framework based on information theory for high dimensional data. *Appl. Soft Comput.* **111** (2021) 11.
[23] T. Almutiri and F. Saeed, A hybrid feature selection method combining gini index and support vector machine with recursive feature elimination for gene expression classification. *Int. J. Data Mining Modell. Manage.* **14** (2022) 41–62.
[24] W. Li, Y. Chi, K. Yu and W. Xie, A two-stage hybrid biomarker selection method based on ensemble filter and binary differential evolution incorporating binary african vultures optimization. *BMC bioinf.* **24** (2023) 130.
[25] W. Xie, S. Zhang, L. Wang, K. Yu and W. Li, Feature selection of microarray data using multidimensional graph neural network and supernode hierarchical clustering. *Artif. Intell. Rev.* **57** (2024) 3.
[26] W. Ali and F. Saeed. Hybrid filter and genetic algorithm-based feature selection for improving cancer classification in high-dimensional microarray data. *Processes* **11** (2023) 2.
[27] H. Chamlal, T. Ouaderhman and F.E. Rebbah, A hybrid feature selection approach for microarray datasets using graph theoretic-based method. *Inf. Sci.* **615** (2022) 449–474.
[28] A. Bir-Jmel, S.M. Douiri and S. Elbernoussi, Gene selection *via* bpso and backward generation for cancer classification. *RAIRO – Oper. Res.* **53** (2019) 269–288.
[29] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J.M. Benítez and F. Herrera, A review of microarray datasets and applied feature selection methods. *Inf. Sci.* **282** (2014) 111–135.
[30] Z.M. Hira and D.F. Gillies, A review of feature selection and feature extraction methods applied on microarray data. *Adv. Bioinf.* **2015** (2015) 13.
[31] S. Osama, H. Shaban and A.A. Ali, Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: a comprehensive review. *Expert Syst. Appl.* **213** (2023) 3.
[32] X.S. Yang, Nature-inspired Metaheuristic Algorithms. Luniver Press (2010).
[33] X.-S. Yang, Cuckoo Search and Firefly Algorithm: Theory and Applications. Springer Publishing Company, Incorporated (2013).
[34] G. Piatetsky-Shapiro and P. Tamayo, Microarray data mining: facing the challenges. *ACM SIGKDD Explor. Newsl.* **5** (2003) 1–5.
[35] V.N. Vapnik, The Nature of Statistical Learning Theory. Springer-Verlag (1995).
[36] C.J.C. Burges, A tutorial on support vector machines for pattern recognition. *Data Mining Knowl. Discovery* **2** (1998) 121–167.
[37] S. Sathiya Keerthi and C.-J. Lin, Asymptotic behaviors of support vector machines with gaussian kernel. *Neural Comput.* **15** (2003) 1667–1689.
[38] S. Mahdavi, S. Rahnamayan and K. Deb, Opposition based learning: A literature review. *Swarm Evol. Comput.* **39** (2018) 1–23.
[39] Z. Seif and M.B. Ahmadi, Opposition versus randomness in binary spaces. *Appl. Soft Comput. J.* **27** (2015) 28–37.
[40] M. Parmar, A. Sonker and V. Sejwar, A comparative analysis for filter-based feature selection techniques with tree-based classification. *Int. J. Recent Innovation Trends Comput. Commun.* **11** (2023) 360–369.
[41] I. Kononenko, Estimating attributes: analysis and extensions of relief, in European Conference on Machine Learning. Springer (1994) 171–182.
[42] K. Kira and L.A. Rendell, A Practical Approach to Feature Selection. Elsevier (1992) 249–256.
[43] N. Sánchez-Maroño, A. Alonso-Betanzos and M. Tombilla-Sanromán, Filter methods for feature selection – a comparative study. *Lect. Notes Comput. Sci. (including subseries Lect. Notes Artif. Intell. and Lect. Notes Bioinf.)* **4881** (2007) 178–187.
[44] E. Pashaei and E. Pashaei, Hybrid binary arithmetic optimization algorithm with simulated annealing for feature selection in high-dimensional biomedical data. *J. Supercomput.* **78** (2022) 15598–15637.
[45] I. Guyon, J. Weston and S. Barnhill, Gene Selection for Cancer Classification Using Support Vector Machines. Technical report (2002).
[46] R.A. Musheer, C.K. Verma and N. Srivastava, Novel machine learning approach for classification of high-dimensional microarray data. *Soft Comput.* **23** (2019) 13409–13421.
[47] J. Zhang, B. Gao, H. Chai, Z. Ma and G. Yang, Identification of dna-binding proteins using multi-features fusion and binary firefly optimization algorithm. *BMC Bioinf.* **17** (2016) 323.
[48] Microarray datasets. http://csse.szu.edu.cn/staff/zhuzx/Datasets.html. Accessed May 10. (2024).
[49] Elvira biomedical data set repository. http://leo.ugr.es/elvira/DBCRepository. Accessed May 10. (2024).
[50] D. Chen, 10 gene microarray datasets. mendeley data, v1, 2023. https://data.mendeley.com/datasets/cdsz2ddv3t/1. Accessed Julay 06. (2025).

[51] S.K. Baliarsingh, S. Vipsita and B. Dash, A new optimal gene selection approach for cancer classification using enhanced jaya-based forest optimization algorithm. *Neural Comput. Appl.*, **32** (2020) 8599–8616.

[52] J. Pirgazi, M. Alimoradi, T.E. Abharian and M.H. Olyaee, An efficient hybrid filter-wrapper metaheuristic-based gene selection method for high dimensional datasets. *Sci. Rep.* **9** (2019) 12.

[53] J. Apolloni, G. Leguizamón and E. Alba, Two hybrid wrapper-filter feature selection algorithms applied tohigh-dimensional microarray experiments. *Appl. Soft Comput. J.* **38** (2016) 922–932.

[54] J. Lv, Q. Peng, X. Chen and Z. Sun, A multi-objective heuristic algorithm for gene expression microarray data classification. *Expert Syst. Appl.* **59** (2016) 13–19.

[55] N. Dif, M.W. Attaoui and Z. Elberrichi. Gene Selection for Microarray Data Classification Using Hybrid Meta-Heuristics. Vol. 64. Springer (2019) 119–132.