Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Review:

Feature selection techniques for microarray datasets: a comprehensive review, taxonomy, and future directions*

Kulanthaivel BALAKRISHNAN, Ramasamy DHANALAKSHMI‡
Department of Computer Science and Engineering, Indian Institute of Information Technology, Tiruchirappalli 620012, India
E-mail: bala.k.btech@gmail.com; r_dhanalakshmi@yahoo.com

Received Dec. 10, 2021; Revision accepted June 7, 2022; Crosschecked July 4, 2022

Abstract: For optimal results, retrieving a relevant feature from a microarray dataset has become a hot topic for researchers involved in the study of feature selection (FS) techniques. The aim of this review is to provide a thorough description of various, recent FS techniques. This review also focuses on the techniques proposed for microarray datasets to work on multiclass classification problems and on different ways to enhance the performance of learning algorithms. We attempt to understand and resolve the imbalance problem of datasets to substantiate the work of researchers working on microarray datasets. An analysis of the literature paves the way for comprehending and highlighting the multitude of challenges and issues in finding the optimal feature subset using various FS techniques. A case study is provided to demonstrate the process of implementation, in which three microarray cancer datasets are used to evaluate the classification accuracy and convergence ability of several wrappers and hybrid algorithms to identify the optimal feature subset.

Key words: Feature selection; High dimensionality; Learning techniques; Microarray dataset
https://doi.org/10.1631/FITEE.2100569 **CLC number:** TP391

1 Introduction

In the last two decades, the development of DNA microarray (MA) datasets has stimulated a new wave of research in bioinformatics and machine learning (ML). The gene expression patterns of malignant and normal cells in MA datasets are used for cancer diagnosis in clinical research. An MA dataset contains the minimum number of samples and the maximum number of features (Hambali et al., 2020). The number of features in the raw data ranges from 6000 to 60 000 because the gene expression is analyzed as a whole, despite the fact that the numbers of training and testing samples are generally quite small (frequently, less than 100) (Alonso-Betanzos et al., 2019; Bolón-Canedo and Remeseiro, 2020).

Several studies have demonstrated that the majority of genes detected in MA research are not important in accurately identifying an ailment. Feature selection (FS) is a pivotal preprocessing step in ML tasks (including classification, clustering, association, and regression) that helps overcome the problem of high dimensionality (Chen RC et al., 2020). Selecting relevant features from the MA dataset provides high accuracy and lowers the computational complexity. The optimality of the chosen function subset is assessed using a predetermined criterion. To illustrate the concept, Fig. 1 gives a non-technical view of FS.

Various researchers have focused on elaborating and evaluating FS on supervised learning tasks due to the prevalence of FS in varied disciplines such as medicine (Remeseiro and Bolon-Canedo, 2019), engineering (Shadravan et al., 2019), and healthcare

‡ Corresponding author
* Project supported by the Department of Science and Technology under the Interdisciplinary Cyber-Physical Systems Scheme (No. T-54)
ORCID: Kulanthaivel BALAKRISHNAN, https://orcid.org/0000-0003-2009-4414; Ramasamy DHANALAKSHMI, https://orcid.org/0000-0003-2928-584X

(Tadist et al., 2019). Fig. 2 illustrates the process of a conventional FS technique. A subset is generated from the original MA dataset with a valid search procedure at the first step. The optimum subset is compared to the antecedent subset in the second phase, which involves evaluating a list of subsets. The update keeps the same if the newly updated subset is more highly recommended than the previous one. The procedure continues until the stopping condition is satisfied. The best subgroup is chosen and is given as the input to the classification procedure for validation.

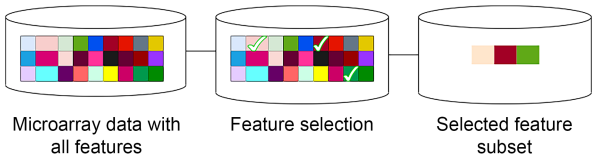

Fig. 1 Non-technical perspective of feature selection

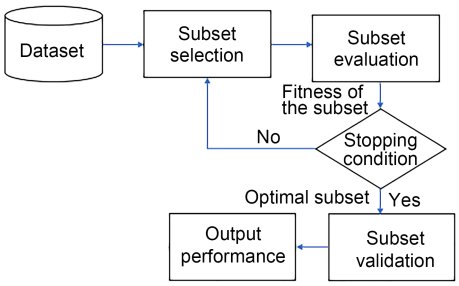

Fig. 2 A general feature selection process

The label status thechnique and search strategy based technique are two types of FS (Saw and Myint, 2019). In label status, methods are classified based on whether the samples are labeled or not. They are classified into supervised and unsupervised methods. The majority of well-known, supervised FS algorithms rely on the creation of a similarity matrix to choose features based on the graph structure. The prominent FS algorithms are fisher score and linear discriminative analysis (LDA). Unsupervised FS is considered a challenging issue as there is no label information to assist in searching for relevant features in the data.

Search strategy based FS approaches are classified into five different categories:

1. Filter method (FM). Filters are employed, rather than a learning algorithm, to choose the optimal function subset based on the general features of the input. In most situations, filters use a set of assessment metrics to calculate a feature's score (Saeys et al., 2007). Pearson correlation (Mangal and Holm, 2018), Chi-squared test (Ranjani and Ramyachitran, 2018), entropy, Fisher score (Prasad et al., 2018), ANOVA (Sahu et al., 2017), relief (Ebrahimpour et al., 2018), information gain (IG) (Arunkumar and Ramakrishnan, 2018), and minimum redundancy maximum relevance (mRMR) (Dong et al., 2018) are well-known examples of conventional FMs. Fig. 3 illustrates the taxonomy of FS strategies.

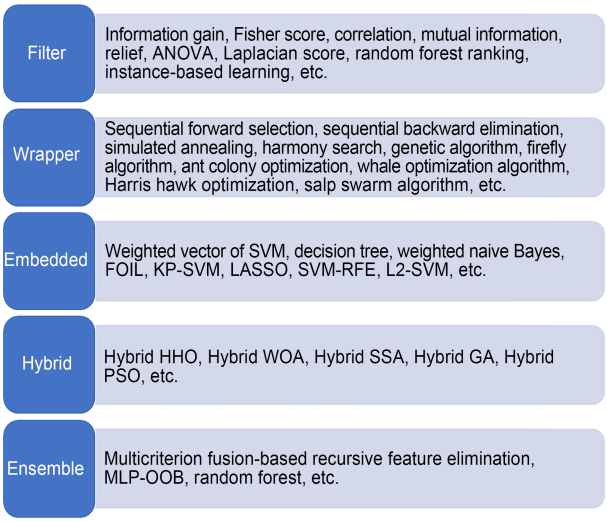

Fig. 3 Taxonomy of feature selection

2. Wrapper method. The wrapper technique chooses the optimal feature subset using learning algorithms. In addition, subsets are chosen using regular search techniques, while the quality of the chosen features is assessed through learning algorithms. The drawback of this strategy is that it requires more computation time than the filter technique. Sequential forward selection (SFS), sequential backward selection (SBS) (Maldonado et al., 2014), beam search (Urbanowicz et al., 2017), genetic algorithm (GA) (McCall, 2005), particle swarm optimization (PSO) (Shao et al., 2012), advanced binary ant colony optimization (ACO) (Kashef and Nezamabadi-Pour, 2013), harmony search (HS) (Diao and Shen, 2012), differential evolution (DE) (Storn and Price, 1997), whale optimization algorithm (WOA) (Mirjalili S and Lewis, 2016), and artificial bee colony (ABC) (Gao et al., 2012) are some notable wrapper methods.

3. Embedded method. This method facilitates the integration of the filter and wrapper methods. This

approach addresses the concerns of the filter and wrapper techniques. Embedded methods keep track of relevant features while learning at the initial stage. Support vector machine (SVM) (Bouazza et al., 2018), decision tree (Shukla et al., 2019), weighted naive Bayes (Rani and Devaraj, 2019), kernel-penalized SVM (KP-SVM) (Maldonado and López, 2018), and LASSO (Rahimipour and Usefi, 2019) are prominent examples of embedded methods.

4. Hybrid method. The hybrid method employs two recombination approaches to hybridize wrapper and filter methods (Aziz et al., 2017). As a preprocessing strategy, the filter approach is employed first, followed by the wrapper approach—a two-stage procedure. Second, either a filter or a wrapper method is used to integrate local search algorithms. Recent examples of the hybrid method are the hybrid salp swarm algorithm (SSA) (Balakrishnan et al., 2021), hybrid WOA (Liu et al., 2020), and hybrid Harris hawk optimization (HHO) (Houssein et al., 2020).

5. Ensemble method. This method uses a set of feature subsets from various base classifiers (Seijo-Pardo et al., 2017). Functionally, two categories of ensemble technique are defined by being heterogeneous or homogeneous. The first method employs several selection algorithms on the same dataset. In contrast, the second method employs the same selection strategy across multiple, dispersed sets of the same dataset, such as the multicriterion fusion-based recursive feature elimination (MCF-RFE) (Yang and Mao, 2011), multilayer perceptron (MLP) (Bramer, 2007), and random forest (RF).

Researchers have used the above-mentioned five approaches to address problems in high-dimensional datasets. This review strives to identify research on this subject and highlight some of the unresolved problems. The primary objectives of this research are as follows:

1. to present an intensive overview of FS methodologies based on search strategies and an insight into contemporary FS techniques used on existing MA datasets;

2. to give a comprehensive literature review on these five types of FS techniques—filter, wrapper, embedded, hybrid, and ensemble;

3. to trace the difficulties and research concerns associated with developing an FS algorithm;

4. to thoroughly show the procedure for computing performance evaluation metrics;

5. to present a case study to understand the performances of different FS approaches; and

6. to evaluate how three MA cancer datasets are used to evaluate the efficiency of a few, well-known wrapper and hybrid FS algorithms.

2 Microarray datasets

This section comprises a meticulous overview of MA datasets. Biologists employ DNA MA technologies to track gene expression in afflicted cells and identify responsible biomarkers (van Hal et al., 2000). It is also understood that MA gene expression data has a significant impact in cancer research. Thousands of genes and a limited number of samples make it a high-dimensional dataset. MA experimentation data is organized and stored as a matrix of dimensions $m\times n$:

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}. \tag{1}$$

MA data's high dimensionality combines superfluous information alongside vital characteristics, making it difficult to identify the best outcomes (Cooper, 2001). MA datasets are often found with binary categories and multiple categories in the target variable. The challenges associated with MA datasets are as follows:

1. The computation time is high due to the problems of high dimensionality and irrelevant and insignificant information.

2. The imbalanced MA dataset negatively impacts the learning process, which leads to lower accuracy.

Table 1 lists several datasets that have been used by academics in recent years and are publicly available for experimentation.

3 Steps involved in feature selection

In general, the FS approach involves a five-step process: search direction, search strategy, FS techniques,

stopping criteria, and evaluation measures. Fig. 4 shows the steps involved in FS. The following subsections discuss these steps in detail.

3.1 Search direction

FS selects a search direction at the initial stage. There are four typical search directions:

1. Forward search: The forward search process begins with an empty set of characteristics. For each succeeding phase, it adds a random feature or a unique feature.

2. Backward search: The entire list of features is used to begin a backward search. It removes a random feature or a feature that optimizes some objectives in each successive phase.

3. Bi-directional: The benefits of forward search and backward search are combined in bi-directional search. For each step of the search, a feature is either added or removed.

4. Random search: Showing excellent results, random search may be used for FS. It selects features at random, tests the model's performance, and iterates as required. Random search creates a random integer N between 1 and the number of attributes. It produces a random series of N integer values ranging from 0 to N–1 with no repetitions. Then, the model is run on such characteristics and is validated, and the average value of a certain performance metric is preserved. Furthermore, the feature array that provides the best efficiency based on the performance metric of choice is obtained.

3.2 Search strategy

An efficient search strategy must achieve rapid convergence and deliver an optimum solution with low computing costs and strong global search capabilities. The following are the three most common search strategies:

1. Sequential search: Sequential forward search, as an example, follows a certain sequence in selecting the optimal feature subset. This technique is vulnerable to feature interaction and risks, achieving local minima.

2. Exponential search: This is a comprehensive search that ensures an optimal solution, but it is costly. This method identifies all potential feature subsets before selecting the best one, which is computation demanding, especially in large datasets.

3. Heuristic search: Heuristic search is carried out using a cost measure or a heuristic function that optimizes the result iteratively. It often does not provide the best option, but it does provide an acceptable amount of time and memory space.

3.3 Feature selection techniques

This subsection includes a comprehensive summary of five main FS approaches and a detailed analysis.

Table 1 Microarray datasets

No.	Name	URL
1	ArrayExpress	https://www.ebi.ac.uk/arrayexpress/search.html?query=MICROARRAY
2	Gene Expression Omnibus	https://www.ncbi.nlm.nih.gov/geo/
3	Get data from NCBI Gene Expression Omnibus (GEO)	http://www.bioconductor.org/packages/release/bioc/html/GEOquery.html
4	Microarray Gene Expression Cancer Data	https://data.mendeley.com/datasets/ynp2tst2hh/4
5	DRIVE: Digital Retinal Images for Vessel Extraction	https://www.isi.uu.nl/Research/Databases/DRIVE/
6	Inspire datasets	https://medicine.uiowa.edu/eye/inspire-datasets
7	VOPTICAL Databases	http://www.varpa.es/research/optics.html#databases
8	BCI Competitions	http://www.bbci.de/competition/

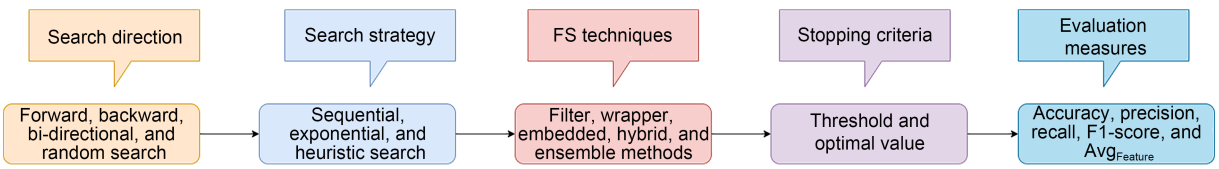

Fig. 4 Steps involved in feature selection

3.3.1 Filter methods

FM checks whether a particular feature is relevant in knowledge discovery regardless of the error rate. The feature/relevance score is initially calculated and compared with those of the remaining features in the dataset. Later, features with a low-score subset from the MA dataset are ignored. The recommended optimal subset of features is then transferred to the classification approach for validation. The two categories of FM are univariate and multivariate. The univariate category examines each characteristic independently, whereas the multivariate category evaluates features concerning the relationship between the features. Fig. 5 shows the basic steps involved in the filter approach.

1. Traditional filter methods

Some of the traditional FMs that are documented in the literature include the following types:

(1) Correlation is a mathematical tool that uses the correlation coefficient to select features or estimate the linear relationship between two variables or features. It is a relationship that aids in determining the correctness of the features in the dataset pertinent to hypothesis testing. The correlation value lies in the range of [–1, 1]. It includes the Pearson correlation, Kendall correlation, Spearman correlation (SC), concordance correlation, intraclass correlation, Moran's I, Phi coefficient, point biserial correlation, polychoric correlation, and zero-order correlation (Sakae et al., 2019). These are examples of some widely used correlations. IG selects the features in an ordered ranking based on the threshold value (Hira and Gillies, 2015). The threshold value helps select the features with positive IG for the next process.

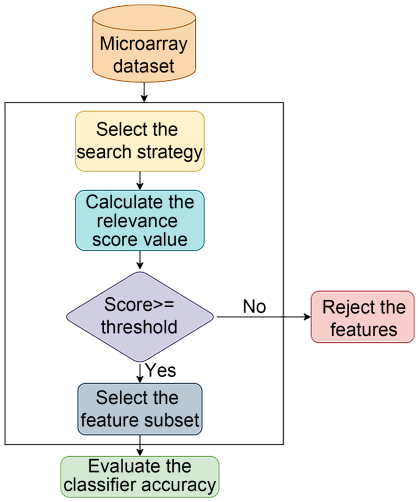

Fig. 5 Filter approach

(2) Entropy of a subset plays a vital role in calculating IG. Mutual information (MI) is a measure for calculating the dependency of two random characteristics on one another. This approach determines how much information one variable has over another (Vergara and Estévez, 2014).

(3) mRMR focuses on MA between two classes in a high-dimensional dataset. This approach identifies the minimum redundancy and the maximum relevance of both discrete and continuous variables. It is simple to implement, and it produces more accurate results than other approaches (Peng et al., 2005). Symmetrical uncertainty (SU) is a suitable metric for assessing the quality of features. SU is a condensed version of MI that ranges from 0 to 1. As a result, the total number of contrasts is limited (Hall, 1999).

(4) Relief evaluates the relevance score of a certain attribute by comparing neighboring distances in the same class and between classes. The modified relevance score, which ranges from –1 (irrelevant) to 1 (relevant), chooses relevant features. The main flaw of the algorithm is its rigid binary categorization. As a result, a new relief algorithm has been created to effectively handle missing data in the dataset (Jain et al., 2018).

(5) ANOVA aims to find the difference between the means of two or more groups. It also assesses the differences between groups as well as within groups (Arowolo et al., 2016). The result of ANOVA produces a p-value and an F-statistic value. The p-value is used to rank relevant features, which decreases the computational complexity. The Laplacian score uses an unsupervised feature-filter system with similar features potentially connected to the same class (He et al., 2016). Independent component analysis (ICA) is an FS approach for extracting independent components from non-Gaussian data by finding the linear representation. ICA decomposes a feature when it is statistically independent of others (Zheng et al., 2006). The instance-based learning (IBL) filtering technique ranks the features by monitoring the instances. In IBL, various instances assign different ratings to various characteristics (Aha et al., 1991). The most important genes are chosen for the Bhattacharya

distance by lowering the overall probability of an upper limit. The co-variance and the vector mean of each class are used to measure the Bhattacharyya difference (Xuan et al., 2006).

2. Recent filter methods

Mazumder and Veilumuthu (2019) proposed a novel FM using Joe's normalized MI to improve the FS FM and its application to select optimal features. Multi-class MA cancer datasets were assessed using five different classifiers. In all the five classifier situations on all MA datasets studied, the suggested technique exhibited strong improvements in terms of classification accuracy (improvement of 5.1%) and area under the curve (AUC) values while reducing classification time (a median of 2.86 s was conserved during training). By integrating SC and distributed filter FS approaches, Shukla and Tripathi (2019) developed a unique, two-stage FS methodology that is able to choose effective feature genes for discriminating samples from MA datasets. The suggested model's aim was to enhance classification accuracy on MA datasets. It was also employed to estimate the gene–gene and gene–class relationships, as well as to find groups of important genes at the same time. On six datasets, the experimental findings showed that the technique suggested offers additional assistance for a large reduction of features employing the SVM classifier and achieves high prediction performance in terms of the accuracy, sensitivity, precision, and *F*-measure. Using filter-based FS, Kavitha et al. (2020) proposed to reduce computation time and improve classification and prediction accuracy. An SVM classifier was used to evaluate the Leukemia dataset with different, conventional, filter-based FS methods. The use of a score-based strategy for FS can yield better results than employing separate methods. Ke WJ et al. (2018) suggested a score-based criteria fusion (SCF) for FS, which uses two composite score filters to choose factors based on an estimate of feature–class significance. The authors tested the technique suggested using SVM and *K*-nearest neighbor (KNN) based on five MA datasets. When compared to previous approaches, the method suggested showed its high efficiency in picking highly relevant features.

Yuan et al. (2019) introduced a novel FS for MA data categorization called partial maximum correlation information (PMCI). This study introduced a novel, general, class-encoding approach for multi-class situations. To optimize classification accuracy, SVM and KNN classifiers were used with 10 benchmark MA datasets and compared to traditional techniques. Zare et al. (2019) suggested a novel FS for MA datasets. They suggested using matrix factorization and singular value decomposition. The approach suggested was used to evaluate the criteria for relevance and redundancy. The KNN classifier was used to evaluate the proposed model's accuracy on nine benchmark datasets.

3. Inferences from filter-based feature selection

Much research has successfully employed state-of-the-art filters as a pre-processing step in the hybrid FS technique approach to minimize the number of features that must be approved in the wrapper stage. It is a promising approach for researchers because it yields more accurate findings than the FM FS technique while taking less time than the wrapper FS methods.

Following this, we look at the classification accuracy of features chosen by the FM as well as the runtime required for FS, and develope a suitable classification model based on the chosen feature. We observe that no subgroup of FM outperforms the majority of FMs across all datasets.

4. Specific ideas to handle filter-based feature selection problems

One of the obvious issues in FM is the discrepancy of results that occurs when various methods are applied to the same dataset and provide conflicting outcomes. This discrepancy in FM shows that this issue might contribute to the improper feature being selected, affecting the quality of classification models. In this situation, the process of normalizing feature scores and then combining them into a single unified score is effective in lowering the volatility of FS results.

All FM FS approaches employ a "ranker" to assess the obtained attribute scores generated using statistics, information theory, or some functions of the classifier's output. Domain experts employ feature ranking as a basic approach in determining the best feature subset. However, ranker search methods do not offer the variety of features to be chosen; instead, they leave it up to the domain experts. We conclude that no ranker approach is sophisticated enough to

distinguish significant traits from redundant ones without the help of domain experts. Furthermore, no study discovered an intelligent solution for ranking inside filter techniques. Additional research and analysis are required to produce more complex rankers that can be employed successfully with any FS approach.

Only a few studies have been undertaken to emphasize the significance of detecting feature-to-feature correlation to achieve an improved performance of the FS process. Using relevance and redundancy analysis, fast correlation based FS (FCBF), mRMR, and *F*-statistic are used to choose relevant characteristics and find the optimal features from the specified set to improve the selection process.

To provide fair and reliable findings, FS necessitates properly balanced data. However, having completely balanced data is not always realistic. So, we emphasize the necessity for a viable technique to balance unbalanced data prior to the FS to achieve better outcomes. Automated sampling strategies should be included into FMs to find the imbalanced data without affecting the original data.

3.3.2 Wrapper methods

In wrapper methods, adding a feature at each stage exponentially increases the size of the subset. The model hypothesis is coupled with the classifier in the search space to provide a more reliable classification result. The wrapper method often employs evolutionary or bio-inspired algorithms to aid the search process. The fitness function based learning technique is used to test the feature subset. The wrapper technique usually has higher computing costs and is more likely to overfit, but it outperforms the filter approach in terms of efficiency. Fig. 6 shows the basic steps involved in the wrapper approach.

1. Traditional wrapper methods

GA is inspired by the biological evolution process (Siedlecki and Sklansky, 1989). GA is undoubtedly the most commonly used technique for FS problems. The goal of GA is to spontaneously generate a population with the same hereditary traits. The algorithm consists of three operations: selection, crossover, and mutation. By the selection mechanism, the fittest chromosomes are selected as the algorithm progresses to the following generation.

PSO is inspired by the foraging technique of a flock of birds (Eberhart and Kennedy, 1995). The goal of PSO is to locate the most optimal positions of all particles. Each search agent in the design matrix updates its position concerning the personal and global best values gathered to find the optimal solution.

ACO was employed in the process of FS for the first time in Ke LJ et al. (2008). ACO mimics the behavior of ants by determining the shortest route between the ants' nest and the food supply. As a result, ACO seeks to find the best way inside the weighted graph. Ants can make their way back to the nest using the pheromone they have left on the trail.

Heidari et al. (2019) developed the HHO, a swarm-based optimization method that mimics the Harris hawk's attacking methodology. The main objective of HHO is to accomplish single- and multi-objective assignment by imitating the hawk's hunting methodology in nature. HHO is a theoretically efficient optimizer that solves complicated nonlinear problems quickly and effectively.

Mirjalili S et al. (2017) developed the SSA, a population-based optimization method inspired by the foraging technique of sea salps. SSA mimics the formation of the salp chain in the search for food sources. Individuals are classified as leaders or followers

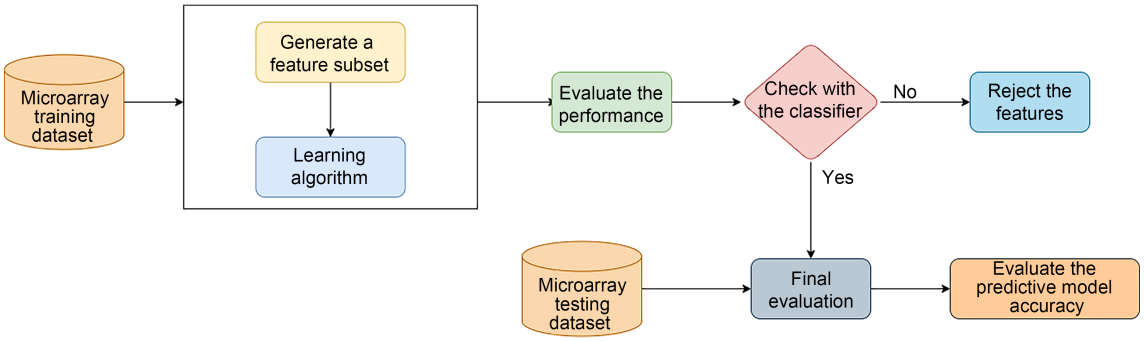

Fig. 6 Flow diagram of the wrapper approach

in SSA based on their place in the salp chain. The position of the leader determines the movement of the salp chain.

The grasshopper optimization algorithm (GOA) is inspired by swarms of grasshoppers (Mirjalili SZ et al., 2018). Grasshopper positions in this algorithm represent candidate solutions. The location of the grasshopper is determined by its social interaction, gravity, and wind advection. WOA is a swarm-based metaheuristic (MH) inspired by the hunting methodology of humpback whales (Tubishat et al., 2019). Humpback whales use a unique hunting technique known as bubble-net feeding to hunt schools of tiny krill fish. The hunting technique of the humpback whales is classified into three typical phases: encircling the prey, bubble-net attacking, and searching for prey.

2. Recent wrapper methods

Almugren and Alshamlan (2019) suggested the FF-SVM, a wrapper FS approach, for categorizing cancer MA gene expression that employs the firefly algorithm (FFA) and an SVM classifier. Five classical microarray datasets (Leukemia1, SRBCT, Lung, Colon, and Leukemia2) were used to assess the suggested model's accuracy. Over the Leukemia1 dataset with three genes and the Lung dataset with two genes, FF-SVM achieved a classification accuracy of 100%.

Ragunthar and Selvakumar (2019) suggested a wrapper-based FS approach called ABC-SDS, which combines hybridized ABC and stochastic diffusion search (SDS) algorithms. The superiority of the suggested approach is assessed with two MA datasets (GDS 531 and GDS 2643), evaluated using the SVM classifier, and compared with the classical FS methods in terms of accuracy, sensitivity, specificity, and the F1-score (Ragunthar and Selvakumar, 2019).

Tawhid and Ibrahim (2020) proposed a wrapper method based FS model using a binary variant of WOA. The suggested model uses three different classifiers (LR, C4.5, and NB) on a breast cancer dataset. The suggested algorithm was assessed with 32 benchmark University of California Irvine (UCI) datasets. The suggested approach achieved the highest mean accuracy for the MA dataset using LR (97%), C4.5 (98%), and NB (97%).

Recently, Balakrishnan et al. (2021) suggested a wrapper-based FS approach based on enhanced SSA using the Lévy flight approach. Six different high-dimensional MA datasets were used (Breast Cancer, CNS, Ovarian, OSCC, Colon, and Leukemia) in this study. The authors used an SVM classifier to assess the selected features in terms of precision, recall, F1-score, and accuracy. The suggested model outperformed SSA by offering a 0.1033% higher confidence in the specified characteristics.

Ghosh et al. (2019) employed a wrapper-based FS method using the recursive memetic algorithm (RMA). To assess the superiority of the suggested approach, seven different MA datasets were employed in the three well-known classifiers. RMA surpassed both the GA and standard memetic algorithm in terms of performance. It achieved 100% accuracy in all circumstances while using a relatively minimum number of genes.

Based on the unique fitness function and binary bat algorithm, Nakamura et al. (2012) suggested a novel, wrapper-based FS approach. This study aimed to reduce intra-class distances while increasing inter-class distances. To test the performance of the suggested strategy, they employed an extreme learning machine as a classifier as well as eight MA datasets. The new fitness function surpassed conventional fitness functions in terms of classification, accuracy, precision, recall, specificity, and F1-score metrics, according to the findings.

Table 2 highlights several other studies that used this wrapper-based method.

3. Inferences from wrapper-based feature selection

Researchers are paying some attention to this area. Some of this research combines two different swarm intelligence (SI) algorithms in a wrapper to combine their benefits, whereas other research simply uses a single SI algorithm. The issues with this technique are that feature space is large and examining every potential mixture takes a large amount of processing time. Some of the algorithms (such as simulated annealing, local beam search, and hill-climbing search) employ the local search algorithms to enhance the proposed approach. Few algorithms (such as binary bat algorithm, WOA, and SSA) use the fitness function to enhance the suggested approach. In the last decade, wrapper-based algorithms such as PSO and GA have been widely used. The use of high-dimensional MA datasets in the majority of studies has been observed. SVM and KNN are the often used classifiers in this research area to examine classification tasks.

Table 2 Recent wrapper approaches based on MA datasets

Reference	FS algorithm	Classifier	Datasets	Performance measure(s)	Findings
Khan et al. (2022)	GA based on community detection	AdaBoost (AB), SVM, KNN	Spam base, Sonar, Arrhythmia, Madelyn, Isolet, Colon	Accuracy, computation time, converging ability	The accuracy of the proposed approach was shown to be 0.52% higher than that of PSO, 1.20% higher than that of ACO, and 1.57% higher than that of the ABC algorithm for three classifiers
Sönmez et al. (2021)	GA-KNN and GA-SVM hybrid methods	KNN, SVM	Colon, DLBCL, Breast, Prostate, CNS, BRCA, COAD	Number of selected features, accuracy, precision, specificity, F1-score, and sensitivity	The recommended approaches were found to enhance all performance indicators. Additionally, the suggested approaches yielded the greatest values for all datasets
Rathee et al. (2022)	Multi-objective genetic algorithm	KNN	WDBC, Lung cancer, Leukemia, Prostate, DLBCL, GSE 412, GSE 2535, GSE 2443	Sensitivity, specificity, G-mean, precision, F1-score	The findings supported the superiority of the MOGA-based parallel method over alternative approaches based on several performance metrics
Deng et al. (2022)	XGBoost-MOGA	SVM, NB	DLBCL, Breast, CNS, Colon, Leukemia, Lung cancer, Lymphoma, Prostate, Myeloma, Ovarian	Accuracy, precision, F1-score, recall	The experimental findings showed that XGBoost-MOGA outperforms state-of-the-art algorithms
Ram and Kuila (2019)	GA-based model	SVM	DLBCL	Sensitivity, specificity, accuracy	Controlling the goal functions was carried out using the weight sum method
Shukla et al. (2019)	Filter-wrapper feature selection	SVM, KNN, DT, NB	Colon, DLBCL, Prostate tumor, SRBCT, Lung cancer, Hepatitis, Ionosphere, Sonar, Lymphography	Accuracy, fitness, number of selected features	The results revealed that the SRBCT dataset had the lowest classification accuracy of 61.24%, and that the Lymphoma dataset had the best classification accuracy of 99.32 % (DLBCL)
Almazini and Ku-Mahamud (2021)	Enhanced graph clustering ACO	SVM, KNN, DT, RF	Wine, Hepatitis, Ionosphere, Spam base, Colon, Leukemia	Accuracy	Research demonstrated that the suggested EGCACO outperformed five other standard UFS algorithms on four classifiers in terms of classification accuracy
Dabba et al. (2021b)	Multi-objective binary Harris hawks optimization	SVM, KNN	CNS, Colon, Leukemia, Ovarian, Brain_Tumor1 , 9_Tumors, 11_Tumors	Accuracy, number of genes	In six datasets, the suggested methodology achieved above 98 % classification accuracy with a maximum accuracy of 100%
Balakrishnan et al. (2022b)	ROBL-MPA	KNN	Breast cancer, CNN, Ovarian, OSCC, Colon, Leukemia	Accuracy, *F*-measure, recall, precision, convergence curve	Based on multiple benchmark performance analysis tests, the suggested ROBL-MPA outperformed the standard MPA
Balakrishnan et al. (2022a)	OBL-MPA	KNN, NB, RF, NN, SVM	Breast cancer, CNN, Ovarian, OSCC, Colon, Leukemia	Accuracy, *F*-measure, recall, precision, convergence curve	According to the findings, the suggested strategy surpassed other standard FS strategies

4. Specific ideas to handle wrapper-based feature selection problems

Features are chosen using the fundamental learning process in the wrapper approach, although it is computationally expensive owing to the iterative picking of the optimal subset of features. The search overhead associated with sequential search techniques is also a disadvantage. To get around this, heuristic search and optimum search techniques based on a bio-inspired algorithm are used to find the best characteristics with the least amount of overhead. As a result, there is much to improve in search algorithms for improved feature subset selection.

The stability of wrapper-based FS approaches is a critical issue that must be addressed. In this context, stability is described as the capability to pick the same set of features regardless of variance in training sample partitioning. Employing empirical aggregation of several trials to form a stronger approximation of key features, essentially wrapping the wrapper, is one strategy to mitigate the instability of the sequential feature of selection strategies. Parallel search techniques and evolutionary algorithms are two alternative search strategies that may provide a solution to the problem of stability.

Another key challenge is determining which classifiers are acceptable for the dataset. This may lead to a reduction of classification performance in the predictive model. SVM and KNN are often used classifiers in this research area to examine classification tasks.

3.3.3 Embedded methods

Despite the deployment of numerous embedded FS techniques in recent years, a unified conceptual model is yet to be developed. Embedded techniques are computationally less expensive than wrapper approaches. They use the core of the learning algorithm for rating the features. The purpose of the embedded methodology is to minimize the computation time required to reclassify distinct feature groups. Fig. 7 shows the basic steps involved in the embedded approach.

1. Embedded methods from 2002 to 2012

Guyon et al. (2002) suggested an SVM based on recursive feature elimination (SVM-RFE) as one of the most prominent embedded techniques. It was designed with the sole purpose of identifying genes that would be used to discover types of cancers.

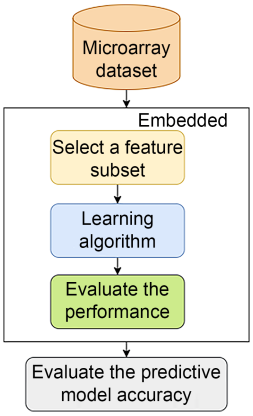

Fig. 7 Flow diagram of the embedded approach

Maldonado and Weber (2011) proposed a unique embedded methodology that simultaneously picks key features throughout classifier modeling by penalizing individual features employed in the dual form of SVM. Kernel penalized SVM (KP-SVM) outperformed competing techniques with increasingly fewer relevant characteristics.

Anaissi et al. (2011) developed a novel embedded technique based on the RF algorithm to address the issue of data imbalance predominantly in MA datasets. The methodology used several tactics and algorithms to tackle the problems of complicated gene expression in the Leukemia dataset. A strategy was used to find the optimal training error cost for a specific class and to deal with data imbalance. Finally, the RF chose the most important characteristics and prevented the learning model from overfitting.

Canul-Reich et al. (2012) proposed an iterative feature perturbation (IFP) approach as an embedded gene selector. They used four distinct MA datasets to test the proposed approach. If adding noise to a feature substantially changes classifier performance, the feature is considered significant. When compared to the SVM-RFE approach, the IFP technique achieved similar or greater average class accuracy on three of the four datasets tested.

2. Recent embedded methods

Zhang G et al. (2020) introduced a rule-based FS technique based on the first-order inductive learner (FOIL). The suggested technique generated a classification rule using a modified propositional version of

the FOIL algorithm. The subset features obtained in previous rules were then combined to create a candidate feature subset that eliminates duplicate features while maintaining interactive and relevant features.

El Kafrawy et al. (2021) proposed a new embedded technique based on SVM-mRMRe. The model was tested using eight of the most widely employed MA datasets for diverse forms of cancer. Four types of classifiers, i.e., RF, MLP, KNN, and SVM, were used to assess the chosen subset feature. The suggested model minimized the time consumption and complexity while improving the distinction of cancerous and benign tissues, according to the findings.

Albashish et al. (2021) suggested an embedded FS model using binary biogeography optimization (BBO) and SVM-RFE. The basic purpose of FS approaches was to maximize the classification performance while reducing the number of features used. SVM-RFE was incorporated in the BBO to increase the quality of the produced results in the mutation operator, thus improving exploitation capabilities and establishing an appropriate balance between exploitation and exploration of the existing BBO. In terms of accuracy and quantity of chosen features, the BBO-SVM-RFE technique surpassed the BBO method as well as other current wrapper and filter methods.

Dabba et al. (2021a) suggested a new embedded technique to deal with gene selection in MA datasets. MI maximization (MIM) was used to determine the genes' relevance and redundancy, whereas mMFA was used to develop gene sets and assessed using the fitness function. The findings in this study, which were executed on 16 binary-class and multi-class benchmark datasets, showed that the MIM-mMFA technique delivers better classification accuracy. Kang et al. (2019) suggested a new embedded approach termed the relaxed LASSO-GenSVM (rLGenSVM) for classification tasks. To test the proposed FS strategy, they employed four sets of MA data of two classes and four sets of multi-class data with GenSVM as a classifier. rLGenSVM achieved 100 % accuracy in six datasets.

By optimizing parameters in kernel functions, Zhu et al. (2018) suggested a unique embedded technique (KPD-SVM) for enhancing accuracy and choosing the most suitable characteristics. An SVM classifier was used to assess the predictive model. For the Wisconsin breast cancer (WBC) dataset, KPD-SVM had an accuracy of 95% using five features, and roughly 88% for the Sonar dataset using around 15 genes. The KPD-SVM approach surpassed the F1-score, filter-based method, RFE-SVM, and wrapper-based methods according to the findings. Kernel-penalized SVDD and KP-CSSVM are two embedded techniques for FS and SVM categorization developed by Zhu et al. (2018). The proposed method achieved this by extending the concept of KP-SVM to two SVM models for class imbalance distribution (SVDD and CS-SVM) to address a skewed class supply problem by combining categorization and variable penalization. Maldonado and López (2018) employed 12 MA datasets using SVM as a classifier to assess the suggested approach.

Zhang L and Huang (2015) suggested a new embedded approach for multiples of SVM-RFE for multi-class FS and classifications. Three different MA datasets, including CNS tumors, Leukemia, and Lung cancer, were evaluated using the SVM classifier in terms of classification accuracy. Because it is effective for CNS tumors and Leukemia datasets, the suggested strategy can increase the performance of each class.

3. Inferences from embedded-based feature selection

According to this review, this is an area to which researchers are paying moderate attention. In many works, the embedded technique was combined with the SVM classifier. RF, MLP, and KNN classifiers have been used only in a limited amount of research and have not outperformed the SVM classifier in terms of accuracy. Little research has been performed on multi-class classification issues. The use of high-dimensional MA datasets has been noted in most of the investigations. Most of the embedded FS study was evaluated based on accuracy, the number of selected features (NSF), and convergence ability.

4. Specific ideas to handle embedded feature selection problems

Identifying the appropriate combination is challenging since it involves the embedding of two different FS approaches (filter and wrapper). We suggest to employ the classical filter approach rather than experimenting with new filter approaches.

Another key challenge in the embedded approach is classifier orientation. The kernel-based SVM classifier is often used in this research area to examine

classification tasks. KPD-SVM provides superior performance compared to conventional embedded approaches. We suggest that this strategy be modified based on FS considerations.

3.3.4 Hybrid methods

Rather than relying exclusively on classic FS approaches such as univariate and multivariate statistical tools, a contemporary approach combines traditional FS methods and new hybrid and ensemble FS techniques. A hybrid methodology combines an independent test with the performance estimation method for the feature subset. Hybrid methods are ideal for selecting important features from the high-dimensional dataset because they reduce the runtime. The goal of the hybrid strategy is to achieve a trade-off between time complexity and feature space size by the filter technique to eliminate unnecessary data from the original dataset. The wrapper approach is then used to extract the best feature subset from the feature pool that has been chosen. This technique accelerates FS since the filtering process quickly removes superfluous characteristics from the dataset. The hybrid method's core workflow is shown in Fig. 8.

1. Traditional hybrid methods

The following are a few contemporary, well-known, and hybrid FS methods:

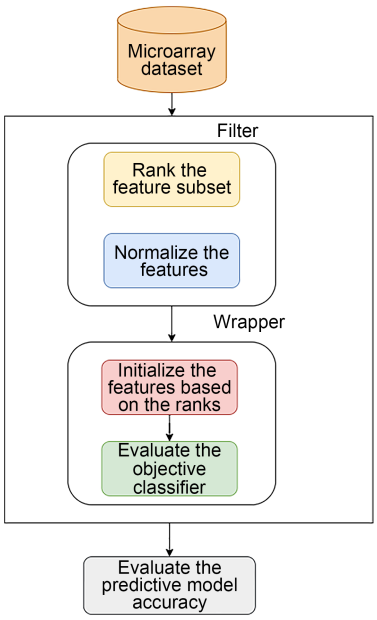

Fig. 8 Flow diagram of the hybrid approach

(1) Hybrid GA. GA variations are divided into five categories: real and binary coded, multi-objective, parallel, chaotic, and hybrid GA (Katoch et al., 2021). Based on chromosome expression, GAs are classified into two types: binary GA (BGA) and real coded GA (RGA) (Payne and Glen, 1993). BGA is created to detect molecule similarities, positions, and conformations. The depiction of chromosomes in an RGA is directly linked to real-life issues. Most RGAs are produced by experimenting with the crossover, mutation, and selection operators (Chuang et al., 2016).

(2) Hybrid FFA. This algorithm refracts optical flashes to simulate the mating and information sharing of fireflies. Emary et al. (2015) presented the first binary variant of the FFA, which uses a threshold value to solve function selection problems. The recommended method is exposed to extensive testing, which results in a straightforward solution to the problem. Kanimozhi and Latha (2015) proposed an image-retrieval strategy based on SVM classifiers and FFA to improve the algorithm's accuracy using optimal functionality.

(3) Hybrid WOA. WOA is a swarm intelligence algorithm designed to solve problems involving continuous optimization. Ling et al. (2017) proposed the Lévy flight trajectory based WOA (LWOA) to accelerate and strengthen WOA movement, thus preventing premature convergence. The Lévy flight trajectory can increase population diversity and potentially hop out of local optima concerning premature convergence. To solve large-scale global optimization (LSGO) problems, an updated WOA (MWOA) has been proposed. A cosine function is used to update the control parameter in a nonlinear dynamic approach to establish equilibrium between exploration and exploitation skills (Sun et al., 2018).

(4) Hybrid HHO. HHO was modified by Gupta et al. (2020) to solve general engineering design problems. The proposed model maintained an equilibrium between the discovery and extraction phases while optimizing HHO's population diversity and convergence efficiency. Thirty-three benchmark problems were used to validate the efficiency of the proposed algorithm. Improved HHO based on SSA—assuming that SSA's powerful explorative capacity will facilitate the exploration of the original HHO—was proposed by Zhang G et al. (2020). The proposed method

contained two stages: initialization and updating. A hybrid version of HHO based on cuckoo search and chaotic maps to boost the efficiency of the original HHO was proposed in Sihwail et al. (2020).

(5) Hybrid SSA. S- and V-shaped transition functions created an efficient binary SSA with a crossover operator (Faris et al., 2018). The proposed method was combined with a KNN classifier and applied to 22 well-known UCI machine learning repository datasets, yielding the best results. The salps' location was modified using Singer's chaotic map and local search algorithm to avoid being trapped into local optima and improve the SSA's discovery and exploitation (Tubishat et al., 2021). Twenty benchmark datasets were used to evaluate the performance of the proposed approach.

2. Recent hybrid methods

ENSVM is a hybrid FS model for cancer classification suggested by Qaraad et al. (2021). The suggested model employs the elastic net technique for FS in high-dimensional MA data, which controls and chooses variables. The authors used the SSD-SVM model and SVM with an RBF kernel without any particular FS techniques in the proposed model. The suggested model was evaluated using seven high-dimensional MA datasets in terms of specificity, sensitivity, and classification accuracy.

Xie et al. (2021) suggested a two-stage hybrid FS approach to enhance the accuracy of classifiers. The suggested model combines the mRMR and improves binary differential evolution (BDE). Four cancer datasets were employed to assess the superiority of the suggested approach in terms of accuracy. The suggested technique may regulate the number of chosen characteristics while balancing global search efficiency and local optimization capability.

Al-Rajab et al. (2021) suggested a framework for exclusive colon-cancer categorization that includes a two-stage, multifilter hybrid approach of FS. The suggested model employs a combination of IG and GA to choose features. To rank the genes, the standard filter approach mRMR was used. The suggested approach was evaluated using DT, KNN, NB, and SVM classifiers. The model outperformed all the classifiers.

To enhance classification accuracy, Wang et al. (2022) introduced the MMPSO technique, which combines the feature-ranking method with the heuristic search method. The suggested approach employed 10 benchmark datasets and the LIHC MA dataset in terms of accuracy (Xie et al., 2021).

Zhang G et al. (2020) suggested a novel hybrid approach using IG and the improved binary krill herd algorithm for MA data classification. To help search for a feature using a hyperbolic tangent function, an adaptive transfer factor and a chaotic memory weight factor were proposed. The KNN classifier was used to assess the predictive model using nine MA datasets.

Albaldawi and Almuttairi (2021) proposed a hybrid FS model using the ANOVA and LASSO methods. Five different cancer datasets were used using a linear support vector, MLP, and RF classifier in terms of accuracy. The findings showed that the models in the spark environment are particularly successful at processing high-dimensional data that cannot be handled using traditional implementations of some techniques.

3. Inferences from hybrid-based feature selection

Based on our review, we find that a hybrid approach is an effective and competitive strategy because it incorporates the benefits of both filter and wrapper or embedded approaches. Furthermore, most of the well-known SI algorithms are used to hybridize novel approaches to solve FS problems. A high-dimensional MA dataset has been used in the majority of research. SVM, NB, and KNN have been commonly used in the literature for classification. The accuracy and NSF are two important evaluation measures that are used frequently.

4. Specific ideas to handle the hybrid-based feature selection problems

The major issue with a hybrid approach is finding the appropriate SI algorithm. Hybridization should be used with two SI algorithms along with any of the suggested algorithm improvements, such as Lévy flight, Brownian motion, opposition-based learning, novel control factor, and parameter tuning.

Another key challenge is determining which classifiers are acceptable for the dataset. This may lead to a reduction of the classification performance in the predictive model. SVM and KNN are used often as classifiers to examine classification tasks.

Stability is also a key challenge in the hybrid approach. The trade-off between bias and variance of the classification error rate is aided by stability. The stability

of the FS algorithm is determined by factors such as the dataset's dimensionality, NSF, sample size, and variability of the data.

3.3.5 Ensemble methods

To solve a specific problem, ensemble learning relies on integrating several models rather than a single model. In recent years, ensemble learning models have demonstrated their usefulness in tackling FS issues. Bagging and boosting are the most common methods for ensemble learning. These methods vary by altering the training set to run the learning algorithm several times through diverse training sets. Fig. 9 depicts the basic flow of the homogenous and heterogeneous approaches of ensemble models. The homogeneous solution employs the same FS approach but with different training data subsets. The heterogeneous strategy employs a variety of FS algorithms, but they are all applied to the same training data. RF is a popular homogenous ensemble technique that combines several DT models, with the added feature that the trees are built from diverse, random subsets of data (Del Río et al., 2014).

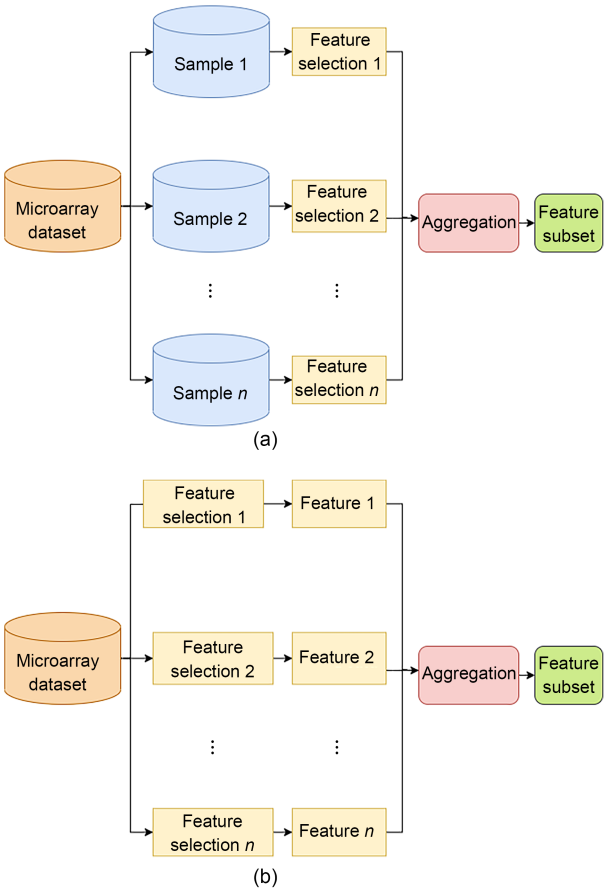

Fig. 9 Homogenous (a) and heterogeneous (b) ensemble methods

1. Traditional ensemble methods

Five separate filters were used in Bolón-Canedo et al. (2012), each of which chose a different subset of features to train and evaluate five classifiers; the results were then merged using basic voting. The MCF-RFE algorithm improved the accuracy and stability of FS (Bonilla-Huerta et al., 2016). Further research suggested a function-rating scheme for MLP ensembles using an out-of-bootstrap (OOB) approximation as the stopping criterion (Windeatt et al., 2011). Three widely used, filter-based attribute-ranking strategies for text classification problems, with the lowest, top, and average rank mixing strategies, were applied in Olsson and Oard (2006). In some other research, the findings of the basic selectors were merged using several combination techniques, also known as aggregators (Olsson and Oard, 2006). Fig. 10 illustrates the benefits and drawbacks of various FS methods.

2. Recent ensemble methods

Hengpraprohm and Jungjit (2020) suggested an ensemble approch, named EnSNR, for breast cancer classification using MA data. EnSNR approach's fundamental concept is to merge relevant features derived from two independent sets of feature assessment. In the suggested approach, three cancer datasets—Ovarian, Lung, and Prostate—were employed with the SVM classifier in terms of accuracy. The EnSNR method considerably decreased the number of irrelevant features (genes) that must be evaluated for cancer classification.

Wang AG et al. (2022) suggested another ensemble model for breast cancer classification. The suggested model was evaluated based on four different cancer datasets based on two classifiers, SVM and KNN. The effectiveness of the suggested approach was assessed with accuracy and stability measures.

Hashemi et al. (2022) suggested a multi-criterion decision-making (MCDM) procedure to examine ensemble FS for the first time. Initially, the authors suggested the EFS-MCDM approach to create a decision matrix using the ranking of each feature according to various rankers. The simple decision matrix and the VIKOR technique were then used to allocate a score to each characteristic. Hashemi et al. (2021) proposed a novel ensemble approach for bi-objective

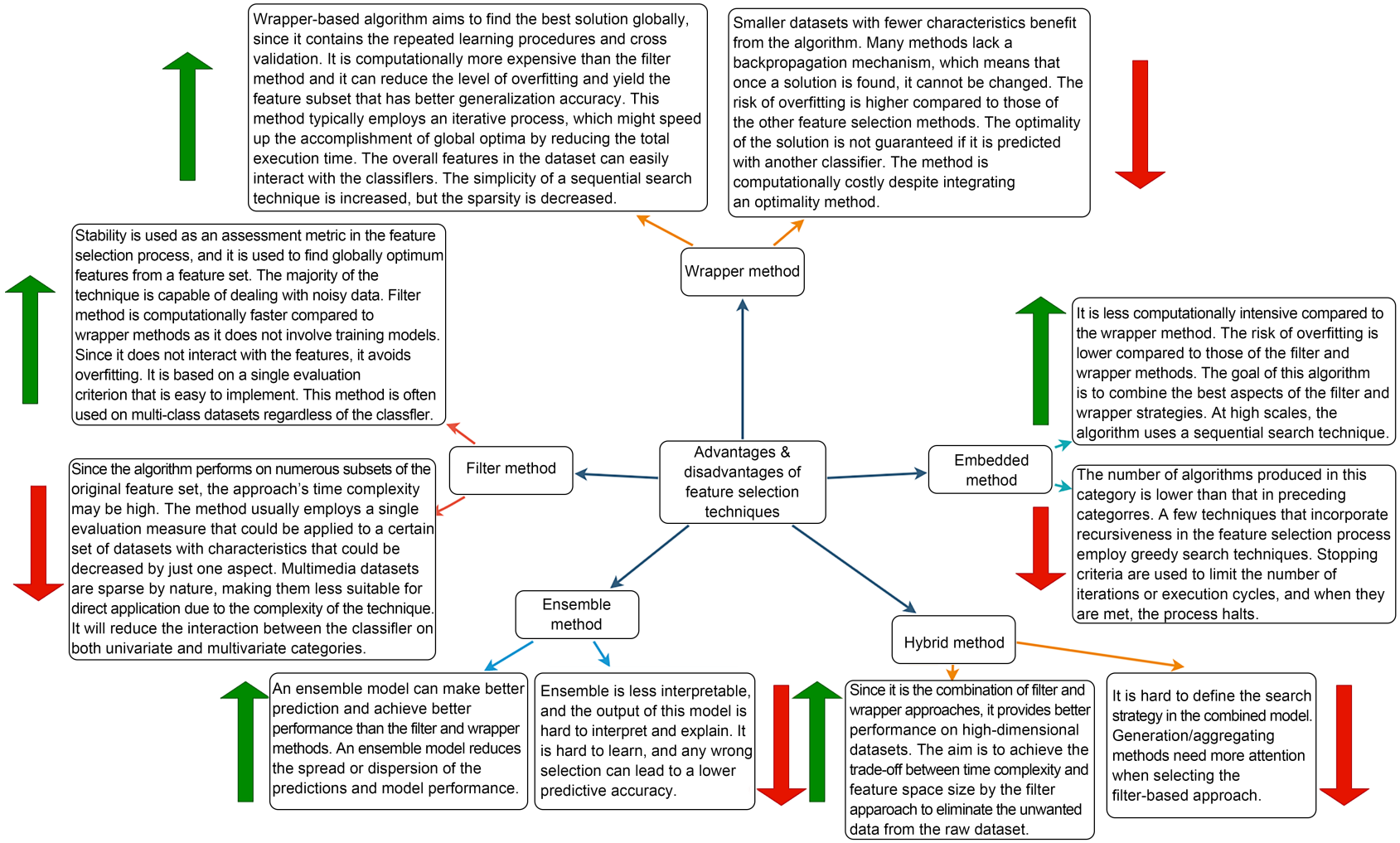

Fig. 10 Advantages and disadvantages of feature selection techniques

optimization problems involving the relevance and degree of redundancy of features. First, a simple decision matrix was generated by various FS techniques, and the modeled bi-objective optimization problem was used to locate non-dominated features. Next, these features were sorted using the crowding distance.

Tsai and Sung (2020) suggested a novel ensemble approach based on a parallel serial combination approach for high- and low-dimensional MA datasets. Ensemble FS performed better than single FS in terms of classification accuracy when compared to nine parallel and nine serial combinations, as well as three single-baseline FS approaches. The serial combination strategy yielded the highest rate of feature reduction. Kalaimani and Umagandhi (2020) proposed an ensemble FS approach for MA data classification. They used SCF, FEHO, and SVM-t. This worked by combining the FS results obtained from the single-feature pickers into a final WMV file. The suggested approach's fitness was determined using KNN, SVM, and RNN.

3. Inferences from ensemble feature selection

This is one of the approaches receiving the least amount of research attention in the literature. It is worth noting that the research in this area has ranged from proposing a novel thresholding technique to developing and comparing alternative approaches to aggregate data, developing cost-sensitive ensemble FS algorithms, and developing various ensemble designs.

4. Specific ideas to handle ensemble feature selection problems

Stability is the major key challenge in the ensemble approach. The trade-off between bias and variance of the classification error rate is aided by stability. Stability of the FS algorithm is determined by factors such as the dataset's dimensionality, NSF, sample size, and variability of the data.

Complex computational search problems are another issue in the ensemble approach. To solve this issue, the term hyper-heuristics has expanded rapidly to denote a learning process or search strategy for creating or choosing heuristics. Hyperparameters for MH algorithms are a type of parameters that may be used to evaluate different control model parameters during the assessment phase of determining the algorithm's feasibility.

3.4 Stopping criteria

Stopping criteria are a type of criteria that manage the classifier when to stop selecting features. Adequate stopping conditions will prevent a model from overfitting, resulting in superior results that are computationally expensive. The following are two of the most common stopping factors:

1. When the search approaches its limit, the limit might be a number of iterations or a large number of characteristics.

2. The findings do not need to be enhanced when some other features are removed.

3.5 Evaluation measures

Various assessment metrics have been used in the literature to evaluate the performance of the FS approach. The following are the metrics that are commonly used to monitor the effectiveness of algorithms:

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}}, \tag{2}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{5}$$

$$\text{F1-score} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{6}$$

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

The average fitness value (a standard deviation of fitness values), convergence ability, and average number of selected features from the source datasets are used to assess the performance. The mean fitness value over the number of complete runs in the algorithm is calculated as

$$\text{Avg}_{\text{Fit}} = \frac{1}{\text{Tot\_runs}} \sum F_I^*. \tag{7}$$

The ratio between the total number of selected features and the available number of features in a high-dimensional dataset is used to compute the average number of selected features:

$$\mathrm{Avg}_{\mathrm{Feature}}=\frac{1}{\mathrm{Tot\_runs}}\sum\frac{\mathrm{Len}\left(F_I^*\right)}{|s|}. \tag{8}$$

The average computation time is calculated by dividing the mean value of the time by the number of total runs:

$$\mathrm{Avg}_{\mathrm{Time}}=\frac{1}{\mathrm{Tot\_runs}}\sum\mathrm{Time}_i. \tag{9}$$

Large volumes of high-dimensional MA data present opportunities as well as obstacles for FS. The importance of appropriate computational paradigms (such as distributed and parallel, multi-label learning, and fusion data mining) for new issues is growing in FS. As a result, we offer these latest ongoing issues and solutions in the following section.

4 Advanced feature selection models

4.1 Distributed feature selection

FS is frequently carried out in a centralized fashion, which means that just one learning model is used to address issues. However, if the data is spread out, FS may be able to analyze numerous subsets sequentially or simultaneously. Fig. 11 shows the vertical partitioning of the dataset. There are two reasons for using distributed feature selection (DFS) (Bolón-Canedo et al., 2015): First, with the advancement of network technology, data is sometimes dispersed over numerous sites, so it may be skewed; Second, most contemporary FS algorithms may not scale well, and their efficiency suffers drastically when handling a massive amount of data. By distributing smaller datasets across multiple cores and learning and integrating the results simultaneously, learning can be parallelized to improve the speed. The two basic strategies for splitting and distributing data are either vertical (i.e., by characteristics) or horizontal (i.e., by samples). Datasets that are too large for batch learning in terms of the number of samples have been scaled up using DFS.

It is worth emphasizing the work of Das et al. (2010), in which a method was provided that executes FS in an asynchronous mode with low communication cost using a horizontal split (by samples) and allows each peer to set its own privacy requirements. Banerjee and Chakravarty (2011) suggested a DFS approach derived from virtual dimension reduction, in which data was partitioned vertically or horizontally. Bolón-Candeo et al. (2013) suggested a distributed filter strategy to increase the accuracy over MA data while reducing the execution time. The model was formed using three steps: (1) datasets were partitioned, (2) filtering was applied to the subsets, and (3) the findings were combined. The findings on eight MA datasets demonstrated that the execution time is significantly reduced while the performance is maintained or even increased when compared to non-partitioned datasets using traditional approaches.

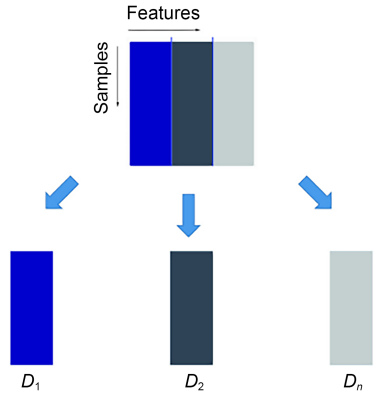

Fig. 11 Vertical partitioning of a dataset

Potharaju and Sreedevi (2018) suggested a DFS approach for complex, high-dimensional datasets. Using the suggested method, the features were spread evenly throughout multiple clusters without duplication. After applying the suggested technique to seven high-dimensional datasets and one low-dimensional dataset, it achieved a 57% success rate and an 18% competitive rate against established methods.

In Morán-Fernández et al. (2017), a distribution strategy was presented that includes splitting data horizontally and vertically and then combining the partial outputs. They used four classifiers to assess the proposed technique on 11 datasets (five of which were MA datasets). They gave users some suggestions for splitting high-dimensional datasets that were appropriate for their aims. The horizontal split was advised by users if reducing storage requirements and processing time was more important than increasing categorization accuracy.

Ye et al. (2019) suggested a DFS model based on intermediate representation. Each party in the suggested methodology found intermediate representations from the original data and distributed them for collaborative FS. The original data from many parties was translated to the same low-dimensional space via shared intermediate representations. The suggested strategy can increase the performance of FS in the local party according to the experimental data.

A novel distributed FS strategy for identifying gene expression data was proposed by Ayyad et al. (2019). The objective was to find the most likely cancer-related genes in a dispersed fashion, which assists in categorizing the data more efficiently. First, a massive quantity of characteristics to be evaluated were split and spread across numerous processors. Next, for each subset of the dataset, a new filter selection approach based on a fuzzy inference system was implemented. Last, all the features that have been generated were ranked, and a wrapper-based selection approach was used.

Jung (2021) proposed a DFS for multi-class classification using the alternating direction method of multipliers (ADMM). Convex optimization and ADMM were used to create a distributed FS algorithm. By employing parallel calculations, the distributed approach scaled effectively with an increasing number of classes. The suggested model was evaluated using two case studies (defect classification and the MNIST dataset), which illustrated a large, multi-class classification challenge.

The least amount of study has been done in the inferences about DFS. In comparison to horizontal partitioning, we note that vertical partitioning is more often employed.

Problem 1: How to perform DFS when the data is stored in a central database?

Solution: MPI and Google's MapReduce protocols are used to store the sophisticated distributed-programming model (Chu et al., 2007). On high-performance computer grids or clusters, these models are helpful for the implementation of DFS. The following four stages can be used to parallelize FS: (1) Across training instances, we break the FS procedure into summation forms; (2) We split data into segments and store them on cluster nodes; (3) On the cluster nodes, we calculate local FS results in parallel; (4) We integrate the local results to acquire the ultimate FS results.

Problem 2: Data is disseminated over a vast number of nodes in a network rather than being stored in a single repository. In these situations, standard FS methods cannot be used directly. As a result, data analysis in such networks will necessitate the creation of a new type of DFS which is capable of operating in such large-scale, distributed contexts.

Solution: These measures are initially examined in a peer-to-peer (P2P) network without data centralization. The program then operates depending on local interactions among participants. In contrast to centralization, the procedure is demonstrably valid because it leads to the appropriate results.

A DFS technique must provide an information sharing and fusion process when data is dispersed over a large number of nodes in a network. This ensures that all distributed sites can operate together to reach a global optimization goal. The data in each node has an identical set of features in the current DFS. Furthermore, since the data on various nodes may have distinct feature representations, more research in DFS is needed.

4.2 Parallel feature selection

Problem: How can vital information for FS be extracted and represented from several sources in parallel FS?

Solution: Venkataramana et al. (2019) introduced a parallel FS framework, termed HFS, which uses a correlation-based feature subset. Furthermore, in parallel, they employed ranking-based FS algorithms to rank features and choose the best ones with a KNN classifier.

Using the Hadoop MapReduce technology, Kečo et al. (2018) built a parallel GA. They employed 11 GEMS datasets in their research for assessing the proposed technique, as well as two classifiers. For fewer than 25 genes, the suggested technique obtained a 100% accuracy rate.

Ray et al. (2016a) implemented an MI FS technique on the Spark framework and applied it to several MA datasets using multiple classifiers, which is an example of parallel-based methodology.

To choose a resilient collection of genes, Boucheham and Batouche (2014) devised a massively parallel

meta-ensemble FS technique. It combines the results of each filter within each ensemble and all ensemble results in parallel. They tested the suggested approach on five MA datasets with three classifiers. The proposed approach was adaptable and produced acceptable results.

Ray et al. (2016b) proposed an sf-ANOVA statistical test using the Spark framework for selecting one of most relevant features. Two different classifiers with three different, high-dimensional datasets were used to assess the suggested model. When compared to logistic regression, the suggested technique using naive Bayes obtained greater accuracy.

Some other research offers a parallel Chi-squared FS approach using Spark for FS. On the Childhood Tumor Gene dataset with a binary class, parallel logistic regression and SVM were used in Lokeswari and Jacob (2017). With 25 genes, the suggested system achieved 63% accuracy using parallel logistic regression and 75% accuracy using a parallel SVM classifier.

4.3 Feature fusion selection

In the process of FS, feature fusion is a procedure that fuses different types of features. Feature fusion's main goal is feature reduction, which can help eliminate noisy features. Feature fusion techniques in particular deal with the selection and combination of characteristics to eliminate duplicate and unnecessary features. They may also be combined with two or more distinct types of characteristics, and thus the dimensionality is reduced. The FS procedure is the most important part of feature fusion. In summary, feature fusion is a step forward in information fusion, and its related strategies may be broadly classified as linear weighted fusion, maximum entropy fusion, neural networks, and Bayesian inference.

A fusion-based FS framework was suggested by Almutiri et al. (2021) who attempted to use different FS approaches and integrate them using ensemble methods. Out of three layers, the first layer runs separately to rank genes and award a score to each gene. A threshold is used in the second layer to filter each gene based on its estimated score. The ultimate choice regarding which genes are significant is determined in the last layer using one of two decision-voting techniques: plurality or agreement. In contrast to other earlier techniques, the suggested framework shows an improvement in accuracy and dimensionality reduction.

Ke WJ et al. (2018) presented a score-based criterion fusion FS model for cancer prediction with the purpose of improving the prediction performance of the classification model. The suggested model was evaluated using different dimensions of datasets using two classifiers. The results showed that the suggested model can uncover extra discriminative features when compared to alternative methods, and that it may be used as a pretreatment algorithm to be successfully integrated with classical models.

Shalabi (2022) developed a new FS technique based on feature stability and correlation to choose the most appropriate minimal subset of features. The suggested technique was compared to various conventional DR algorithms using benchmark datasets to determine its efficiency. The findings showed that the suggested method is the first to reduce a dataset with excellent predicted accuracy.

In several studies, the linear fusion approach has been used to conduct various analytical tasks. In comparison to classical methods, this method is less computationally costly. A fusion system must determine and alter the weights for optimal task completion. Maximum entropy fusion is a statistical model that uses a data theoretic method to determine the likelihood of an occurrence belonging to a specific class based on its information content. Neural network (NN) is another method for combining characteristics extracted from raw data. It is a black box that may be trained to tackle problems that are ill-defined and computationally costly. Although the NN approach appears to be suited for dealing with high-dimensional problems and performing high-order, nonlinear mapping, choosing the right network topology for a specific application may be problematic. Furthermore, the NN approach is prone to sluggish training. Because of these drawbacks, the NN technique has not been used as widely as fusion methods (Zhang R et al., 2019).

4.4 Multi-label feature selection

Problem: The training sample has a distinct label in classic FS. In many real-world applications, however, the same instance can be assigned to many class labels. Thus, there is FS for multi-label data, which attracts much research attention (Chen WZ et al., 2007).

Solution: The most straightforward technique for implementing FS on a multi-label dataset is to turn it into a single-label dataset and then to use a typical FS method. There are several ways for converting a multi-label dataset into a single-label dataset, including: (1) simple transformation; (2) copy transformation; (3) label powerset transformation; (4) binary relevance transformation.

The connection between class labels is typically not taken into consideration when multi-label FS methods based on transformation are used. Sparse training samples and an uneven class distribution present difficulty in every imbalanced single-label dataset. As a result, FS algorithms that deal directly with multi-label data are preferable. Most of today's multi-label FS approaches are simple extensions of single-label FS procedures. A future study will focus on developing multi-label FS software that will be able to deal with multi-label datasets, will not require any transformation, and will take into account the relationships among labels.

5 Open research issues and challenges

Challenges in the FS process and additional issues related to MA datasets are discussed in this section.

5.1 Stability

Stability is a critical consideration when constructing an FS method concerning high-dimensional datasets. An FS algorithm is considered a robust algorithm if identical output is produced regardless of perturbation in input data. Neglecting the FS algorithm's stability problem can lead to incorrect inference and unreliable outcomes. The abandoning of characteristics associated with the chosen features and aligned with the dependent variables is one of the most prominent sources of instability. The trade-off between bias and variance of the classification error rate is aided by stability. The stability of an FS algorithm is determined by factors such as dataset dimensionality, NSF, sample size, and variability of the data.

5.2 Selection of objective functions

Wrapper FS algorithms maximize a given objective function to find the optimum feature subset. FS objective function design differs depending on the classification issues. Initially, an objective function is created during the first phase of the metaheuristic optimization approach that optimizes classification accuracy or decreases the number of characteristics used. Feature counts and classification accuracy are integrated into a single fitness function in a single-objective method.

5.3 Selection of classifiers

Different classifiers, such as KNN, naive Bayes, SVM, RF, and artificial NN (ANN), have been employed to solve FS problems. The choice of a classifier is the most important aspect in achieving the best results from a high-dimensional dataset. According to the literature, KNN is the most often used classifier. SVM is critical in the classification process.

5.4 Small sample size

Many researchers have attempted to deal with the most well-known and targeted challenge, that is, small sample size in MA datasets. The fundamental concern is that small size samples have a significant impact on the performance of the learning technique. To address this issue, it is important to employ an appropriate validation approach to assess the misclassification rate.

5.5 Class imbalance

Class imbalance occurs when a class has more occurrences in a dataset than the other classes, hindering the learning process. Multi-class MA datasets are well-known instances of imbalanced MA datasets. This problem becomes arduous when the test set's imbalance is more evident than that of the training set. Preprocessing methods such as under-sampling and oversampling are commonly employed to overcome this problem. Recently, an ensemble classifier has been presented as a possible solution to the problem of class imbalance.

5.6 Outliers

Outlier detection in MA data is one of the key topics that received little attention in the literature. Outliers are samples in databases that have been polluted owing to measurement errors or the malfunctioning of devices. Outliers are not suitable for the learning

process because they obstruct the selection of informative genes.

6 Case study

This section demonstrates the application of the wrapper and hybrid FS approaches on diverse MA datasets.

6.1 Dataset description

A few real-life MA datasets with a large number of features are discussed in this subsection. The datasets used in this study are obtained from publicly accessible archives. Table 3 lists three high-dimensional MA datasets used to assess the effectiveness of different FS techniques.

Table 3 Overview of datasets

Dataset	Feature count	Sample number
Central nervous system (CNS)	7129	60
Colon cancer	2000	60
Leukemia	7129	72

6.2 Results and discussion

We compare the performances of different wrapper FS methods, such as GA, PSO, WOA, SSA, and HHO, along with a few hybrid approaches, such as ISSA, IWOA, and IHHO. A detailed discussion of all the methodologies has been given in Section 4. These models are evaluated based on their accuracy and convergence capability. The cross-entropy objective function evaluates the model's efficiency by computing the error rate for each iteration. The drop-in error rate, as the model progresses through each iteration, demonstrates the model's capacity to converge to the global minimum. Fig. 12 compares the convergence ability of various wrapper approaches, GA, PSO, WOA, SSA, and HHO, based on three cancer MA datasets. The convergence graph in Fig. 12 demonstrates that WOA experiences premature convergence by revealing that its convergence capability remains unchanged after 5–10 epochs. GA and PSO are effective on both Colon and CNS, but HHO is exclusively effective. In the instance of Leukemia, none of the optimizers except SSA produce an optimum solution.

Fig. 13 compares the classification performances of GA, WOA, SSA, and HHO based on the three MA datasets. The optimizers exhibit the best accuracy based on Leukemia. However, the HHO model produces the greatest accuracy for all datasets compared to the other traditional MH optimization approaches.

Fig. 14 depicts the convergence ability of three hybrid models, ISSA, IWOA, and IHHO. Based on the Colon dataset, IWOA outperforms the two other models in terms of convergence ability. Based on Leukemia, ISSA gradually moves to find the optimal solution, whereas both IWOA and IHHO experience premature convergence. However, in the case of CNS, both IWOA and IHHO perform effectively as compared to ISSA.

The comparative analysis of the classification performances of IWOA, ISSA, and IHHO is shown in Fig. 15. All three hybrid models show significant outcomes based on Leukemia. IWOA has the highest accuracy on CNS, Colon, and Leukemia compared to ISSA and IHHO.

7 Conclusons

MA data analysis offers valuable insight that is helpful in the resolution of problems related to disease discovery. The task of classification is challenging due to the high complexity of gene expression and a limited sample size. As a result, the most practical solution to these problems is to use an FS approach. This study scrupulously consolidates methodologies, methods, datasets, and prospects in recent years of research on the MA dataset. Based on the critical review, this research critique has analyzed numerous research areas such as multi-class classification, improving learning algorithms' performance by different approaches, fixing the dataset imbalance problem, and validating researchers' efforts on MA datasets. To summarize, the following are the contributions of this study:

1. This study gives a thorough overview of FS methodologies based on search strategies and insights into contemporary FS techniques used in existing MA datasets.

2. It provides a comprehensive literature review for five types of FS techniques: filter, wrapper, embedded, hybrid, and ensemble.

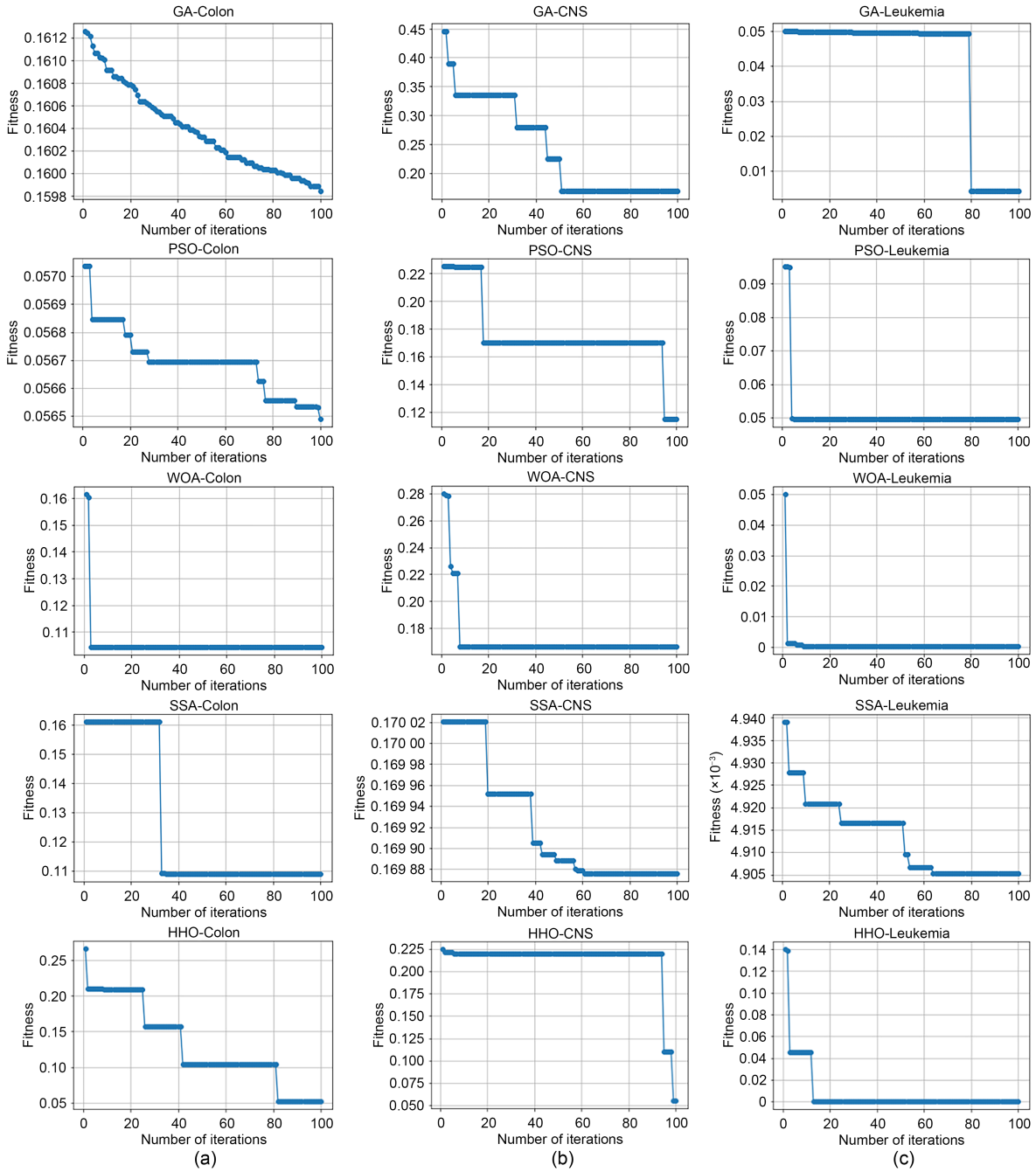

Fig. 12 Converging ability of GA, PSO, WOA, SSA, and HHO based on Colon (a), CNS (b), and Leukemia (c) datasets

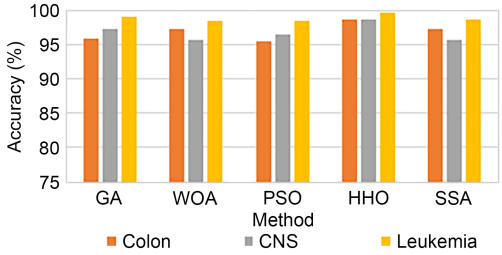

Fig. 13 Classification performances of GA, WOA, PSO, HHO, and SSA based on three MA datasets

3. The difficulties and research concerns associated with developing an FS algorithm are discussed.

4. A procedure for computing performance evaluation metrics is presented.

5. To further understand the performance of alternative FS techniques, a case study is presented.

6. Three MA cancer datasets are used to assess the performances of several well-known wrapper and hybrid FS algorithms.

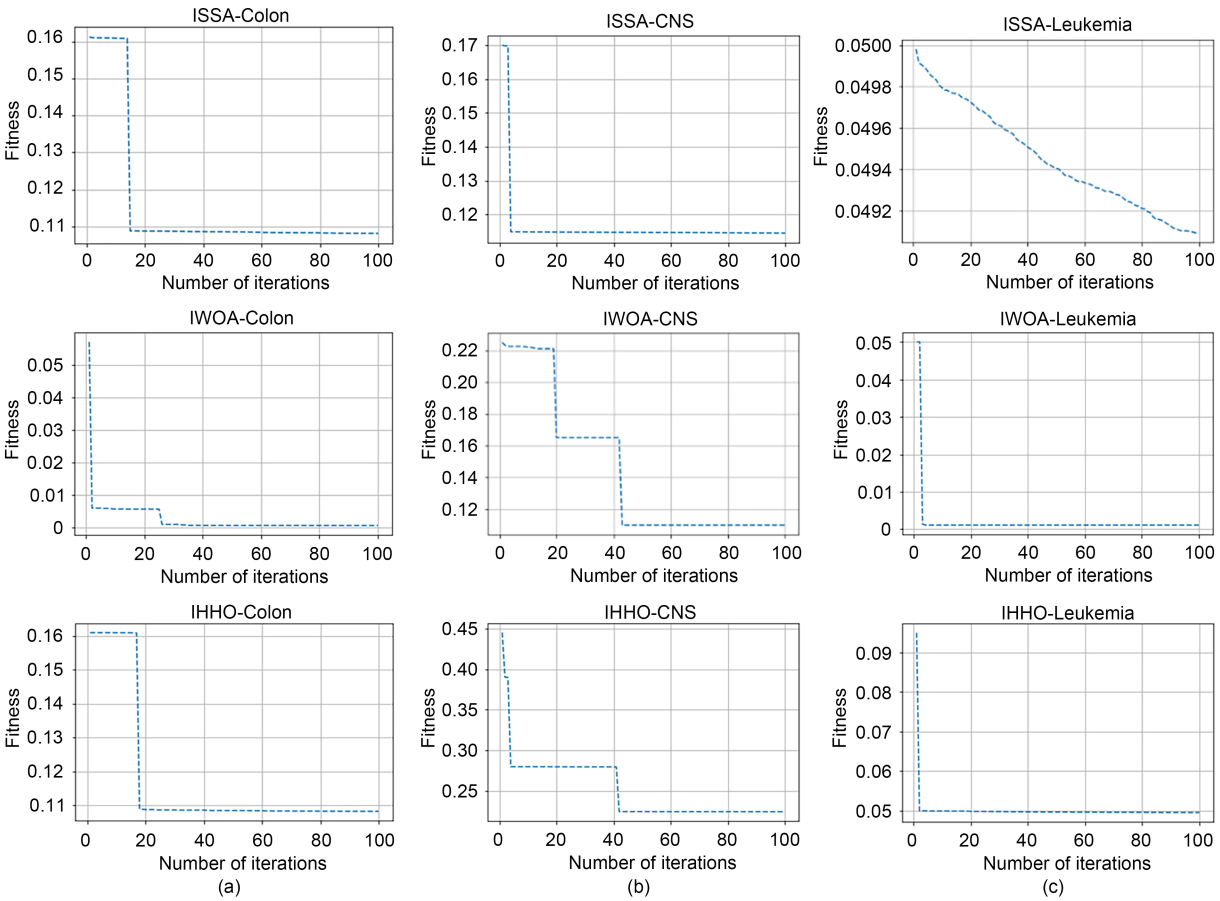

Fig. 14 Converging ability of ISSA, IWOA, and IHHO based on the Colon (a), CNS (b), and Leukemia (c) datasets

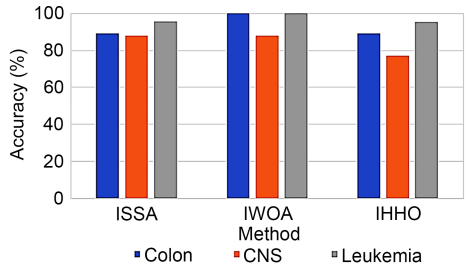

Fig. 15 Classification performance of ISSA, IWOA, and IHHO based on three MA datasets

Contributors

Kulanthaivel BALAKRISHNAN designed the research. Kulanthaivel BALAKRISHNAN and Ramasamy DHANALAKSHMI processed the data. Kulanthaivel BALAKRISHNAN drafted the paper. Ramasamy DHANALAKSHMI helped organize the paper. Kulanthaivel BALAKRISHNAN revised and finalized the paper.

Compliance with ethics guidelines

Kulanthaivel BALAKRISHNAN and Ramasamy DHANALAKSHMI declare that they have no conflict of interest.

References

Aha DW, Kibler D, Albert MK, 1991. Instance-based learning algorithms. *Mach Learn*, 6(1):37-66. https://doi.org/10.1007/BF00153759

Albaldawi WS, Almuttairi RM, 2021. Hybrid ANOVA and LASSO methods for feature selection and linear support vector, multilayer perceptron and random forest classifiers based on spark environment for microarray data classification. *IOP Conf Ser Mater Sci Eng*, 1094(1): 012107. https://doi.org/10.1088/1757-899X/1094/1/012107

Albashish D, Hammouri AI, Braik M, et al., 2021. Binary biogeography-based optimization based SVM-RFE for feature selection. *Appl Soft Comput*, 101:107026. https://doi.org/10.1016/j.asoc.2020.107026

Almazini H, Ku-Mahamud KR, 2021. Adaptive technique for feature selection in modified graph clustering-based ant colony optimization. *Int J Intell Eng Syst*, 14(3):332-345. https://doi.org/10.22266/ijies2021.0630.28

Almugren N, Alshamlan H, 2019. FF-SVM: new firefly-based gene selection algorithm for microarray cancer classification. IEEE Conf on Computational Intelligence in Bioinformatics and Computational Biology, p.1-6. https://doi.org/10.1109/CIBCB.2019.8791236

Almutiri T, Saeed F, Alassaf M, et al., 2021. A fusion-based feature selection framework for microarray data classification.

Int Conf of Reliable Information and Communication Technology, p.565-576. https://doi.org/10.1007/978-3-030-70713-2_52

Alonso-Betanzos A, Bolón-Canedo V, Morán-Fernández L, et al., 2019. A review of microarray datasets: where to find them and specific characteristics. In: Bolón-Canedo V, Alonso-Betanzos A (Eds.), Microarray Bioinformatics. Humana, New York, USA, p.65-85. https://doi.org/10.1007/978-1-4939-9442-7_4

Al-Rajab M, Lu J, Xu Q, 2021. A framework model using multifilter feature selection to enhance colon cancer classification. *PLOS ONE*, 16(4):e0249094. https://doi.org/10.1371/journal.pone.0249094

Anaissi A, Kennedy PJ, Goyal M, 2011. Feature selection of imbalanced gene expression microarray data. Proc 12th ACIS Int Conf on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, p.73-78. https://doi.org/10.1109/SNPD.2011.12

Arowolo MO, Abdulsalam SO, Saheed YK, et al., 2016. A feature selection based on one-way-ANOVA for microarray data classification. *Al-Hikmah J Pure Appl Sci*, 3:30-35.

Arunkumar C, Ramakrishnan S, 2018. Attribute selection using fuzzy roughset based customized similarity measure for lung cancer microarray gene expression data. *Fut Comput Inform J*, 3(1):131-142. https://doi.org/10.1016/j.fcij.2018.02.002

Ayyad SM, Saleh AI, Labib LM, 2019. A new distributed feature selection technique for classifying gene expression data. *Int J Biomath*, 12(4):1950039. https://doi.org/10.1142/S1793524519500396

Aziz R, Verma CK, Srivastava N, 2017. Dimension reduction methods for microarray data: a review. *AIMS Bioeng*, 4(1): 179-197. https://doi.org/10.3934/bioeng.2017.1.179

Balakrishnan K, Dhanalakshmi R, Khaire UM, 2021. Improved salp swarm algorithm based on the levy flight for feature selection. *J Supercomput*, 77(11):12399-12419. https://doi.org/10.1007/s11227-021-03773-w

Balakrishnan K, Dhanalakshmi R, Khaire UM, 2022a. Analysing stable feature selection through an augmented marine predator algorithm based on opposition-based learning. *Expert Syst*, 39(1):e12816. https://doi.org/10.1111/exsy.12816

Balakrishnan K, Dhanalakshmi R, Utkarsh K, 2022b. Excogitating marine predators algorithm based on random opposition-based learning for feature selection. *Concurr Comput Pract Exp*, 34(4):e6630. https://doi.org/10.1002/cpe.6630

Banerjee M, Chakravarty S, 2011. Privacy preserving feature selection for distributed data using virtual dimension. Proc 20th ACM Int Conf on Information and Knowledge Management, p.2281-2284. https://doi.org/10.1145/2063576.2063946

Bolón-Canedo V, Remeseiro B, 2020. Feature selection in image analysis: a survey. *Artif Intell Rev*, 53(4):2905-2931. https://doi.org/10.1007/s10462-019-09750-3

Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A, 2012. An ensemble of filters and classifiers for microarray data classification. *Patt Recogn*, 45(1):531-539. https://doi.org/10.1016/j.patcog.2011.06.006

Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A, 2013. A review of feature selection methods on synthetic data. *Knowl Inform Syst*, 34(3):483-519. https://doi.org/10.1007/s10115-012-0487-8

Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A, 2015. Distributed feature selection: an application to microarray data classification. *Appl Soft Comput*, 30:136-150. https://doi.org/10.1016/j.asoc.2015.01.035

Bonilla-Huerta E, Hernández-Montiel A, Morales-Caporal R, et al., 2016. Hybrid framework using multiple-filters and an embedded approach for an efficient selection and classification of microarray data. *IEEE/ACM Trans Comput Biol Bioinform*, 13(1):12-26. https://doi.org/10.1109/TCBB.2015.2474384

Bouazza SH, Auhmani K, Zeroual A, et al., 2018. Selecting significant marker genes from microarray data by filter approach for cancer diagnosis. *Proc Comput Sci*, 127:300-309. https://doi.org/10.1016/j.procs.2018.01.126

Boucheham A, Batouche M, 2014. Massively parallel feature selection based on ensemble of filters and multiple robust consensus functions for cancer gene identification. Science and Information Conf, p.93-108. https://doi.org/10.1007/978-3-319-14654-6_6

Bramer M, 2007. Principles of Data Mining. Springer, London, UK. https://doi.org/10.1007/978-1-84628-766-4

Canul-Reich J, Hall LO, Goldgof DB, et al., 2012. Iterative feature perturbation as a gene selector for microarray data. *Int J Patt Recogn Artif Intell*, 26(5):1260003. https://doi.org/10.1142/S0218001412600038

Chen RC, Dewi C, Huang SW, et al., 2020. Selecting critical features for data classification based on machine learning methods. *J Big Data*, 7(1):52. https://doi.org/10.1186/s40537-020-00327-4

Chen WZ, Yan J, Zhang BY, et al., 2007. Document transformation for multi-label feature selection in text categorization. Proc 7th IEEE Int Conf on Data Mining, p.451-456. https://doi.org/10.1109/ICDM.2007.18

Chu CT, Kim SK, Lin YA, 2007. Map-Reduce for machine learning on multicore. Proc 19th Int Conf on Neural Information Processing Systems, p.281-288.

Chuang YC, Chen CT, Hwang C, 2016. A simple and efficient real-coded genetic algorithm for constrained optimization. *Appl Soft Comput*, 38:87-105. https://doi.org/10.1016/j.asoc.2015.09.036

Cooper CS, 2001. Applications of microarray technology in breast cancer research. *Breast Cancer Res*, 3(3):158. https://doi.org/10.1186/bcr291

Dabba A, Tari A, Meftali S, et al., 2021a. Gene selection and classification of microarray data method based on mutual information and moth flame algorithm. *Expert Syst Appl*, 166:114012. https://doi.org/10.1016/j.eswa.2020.114012

Dabba A, Tari A, Meftali S, 2021b. A new multi-objective binary Harris Hawks optimization for gene selection in microarray data. *J Amb Intell Human Comput*, early access. https://doi.org/10.1007/s12652-021-03441-0

Das K, Bhaduri K, Kargupta H, 2010. A local asynchronous distributed privacy preserving feature selection algorithm

for large peer-to-peer networks. *Knowl Inform Syst*, 24(3): 341-367. https://doi.org/10.1007/s10115-009-0274-3

Del Río S, López V, Benítez JM, et al., 2014. On the use of MapReduce for imbalanced big data using Random Forest. *Inform Sci*, 285:112-137. https://doi.org/10.1016/j.ins.2014.03.043

Deng XS, Li M, Deng SB, et al., 2022. Hybrid gene selection approach using XGBoost and multi-objective genetic algorithm for cancer classification. *Med Biol Eng Comput*, 60(3):663-681. https://doi.org/10.1007/s11517-021-02476-x

Diao R, Shen Q, 2012. Feature selection with harmony search. *IEEE Trans Syst Man Cybern Part B*, 42(6):1509-1523. https://doi.org/10.1109/TSMCB.2012.2193613

Dong HB, Li T, Ding R, et al., 2018. A novel hybrid genetic algorithm with granular information for feature selection and optimization. *Appl Soft Comput*, 65:33-46. https://doi.org/10.1016/j.asoc.2017.12.048

Eberhart R, Kennedy J, 1995. A new optimizer using particle swarm theory. Proc 6th Int Symp on Micro Machine and Human Science, p.39-43. https://doi.org/10.1109/MHS.1995.494215

Ebrahimpour MK, Nezamabadi-Pour H, Eftekhari M, 2018. CCFS: a cooperating coevolution technique for large scale feature selection on microarray datasets. *Comput Biol Chem*, 73:171-178. https://doi.org/10.1016/j.compbiolchem.2018.02.006

El Kafrawy P, Fathi H, Qaraad M, et al., 2021. An efficient SVM-based feature selection model for cancer classification using high-dimensional microarray data. *IEEE Access*, 9:155353-155369. https://doi.org/10.1109/ACCESS.2021.3123090

Emary E, Zawbaa HM, Ghany KKA, et al., 2015. Firefly optimization algorithm for feature selection. Proc 7th Balkan Conf on Informatics Conf, p.1-7. https://doi.org/10.1145/2801081.2801091

Faris H, Mafarja MM, Heidari AA, et al., 2018. An efficient binary Salp Swarm Algorithm with crossover scheme for feature selection problems. *Knowl-Based Syst*, 154:43-67. https://doi.org/10.1016/j.knosys.2018.05.009

Gao WF, Liu SY, Huang LL, 2012. A global best artificial bee colony algorithm for global optimization. *J Comput Appl Math*, 236(11):2741-2753. https://doi.org/10.1016/j.cam.2012.01.013

Ghosh M, Begum S, Sarkar R, et al., 2019. Recursive memetic algorithm for gene selection in microarray data. *Expert Syst Appl*, 116:172-185. https://doi.org/10.1016/j.eswa.2018.06.057

Gupta S, Deep K, Heidari AA, et al., 2020. Opposition-based learning Harris hawks optimization with advanced transition rules: principles and analysis. *Expert Syst Appl*, 158: 113510. https://doi.org/10.1016/j.eswa.2020.113510

Guyon I, Weston J, Barnhill S, et al., 2002. Gene selection for cancer classification using support vector machines. *Mach Learn*, 46(1):389-422. https://doi.org/10.1023/A:1012487302797

Hall MA, 1999. Correlation-Based Feature Selection for Machine Learning. PhD Thesis, The University of Waikato, Hamilton, New Zealand.

Hambali MA, Oladele TO, Adewole KS, 2020. Microarray cancer feature selection: review, challenges and research directions. *Int J Cogn Comput Eng*, 1:78-97. https://doi.org/10.1016/j.ijcce.2020.11.001

Hashemi A, Dowlatshahi BM, Nezamabadi-Pour H, 2021. A pareto-based ensemble of feature selection algorithms. *Expert Syst Appl*, 180:115130. https://doi.org/10.1016/j.eswa.2021.115130

Hashemi A, Dowlatshahi MB, Nezamabadi-Pour H, 2022. Ensemble of feature selection algorithms: a multi-criteria decision-making approach. *Int J Mach Learn Cybern*, 13(1):49-69. https://doi.org/10.1007/s13042-021-01347-z

He XF, Cai D, Niyogi P, 2016. Laplacian score for feature selection. Proc 18th Int Conf on Neural Information Processing Systems, p.507-514.

Heidari AA, Mirjalili S, Faris H, et al., 2019. Harris hawks optimization: algorithm and applications. *Fut Gener Comput Syst*, 97:849-872. https://doi.org/10.1016/j.future.2019.02.028

Hengpraprohm S, Jungjit S, 2020. Ensemble feature selection for breast cancer classification using microarray data. *Intel Artif*, 23(65):100-114. https://doi.org/10.4114/intartif.vol23iss65pp100-114

Hira ZM, Gillies DF, 2015. A review of feature selection and feature extraction methods applied on microarray data. *Adv Bioinform*, 2015:198363. https://doi.org/10.1155/2015/198363

Houssein EH, Hosney ME, Elhoseny M, et al., 2020. Hybrid Harris hawks optimization with cuckoo search for drug design and discovery in chemoinformatics. *Sci Rep*, 10: 14439. https://doi.org/10.1038/s41598-020-71502-z

Jain I, Jain VK, Jain R, 2018. Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification. *Appl Soft Comput J*, 62:203-215. https://doi.org/10.1016/j.asoc.2017.09.038

Jung D, 2021. Distributed feature selection for multi-class classification using ADMM. *IEEE Contr Syst Lett*, 5(3): 821-826. https://doi.org/10.1109/LCSYS.2020.3006428

Kalaimani V, Umagandhi R, 2020. Hybrid ensemble feature selection (HEFS) model for gene expression microarray data. *Eur J Mol Clin Med*, 7(3):5022-5036.

Kang CZ, Huo YH, Xin LH, et al., 2019. Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine. *J Theor Biol*, 463:77-91. https://doi.org/10.1016/j.jtbi.2018.12.010

Kanimozhi T, Latha K, 2015. An integrated approach to region based image retrieval using firefly algorithm and support vector machine. *Neurocomputing*, 151:1099-1111. https://doi.org/10.1016/j.neucom.2014.07.078

Kashef S, Nezamabadi-Pour H, 2013. A new feature selection algorithm based on binary ant colony optimization. Proc 5th Conf on Information and Knowledge Technology, p.50-54. https://doi.org/10.1109/IKT.2013.6620037

Katoch S, Chauhan SS, Kumar V, 2021. A review on genetic algorithm: past, present, and future. *Multim Tools Appl*, 80(5):8091-8126. https://doi.org/10.1007/s11042-020-1013

Kavitha KR, Prakasan A, Dhrishya PJ, 2020. Score-based feature

selection of gene expression data for cancer classification. Proc 4th Int Conf on Computing Methodologies and Communication, p.261-266. https://doi.org/10.1109/ICCMC48092.2020.ICCMC-00049

Ke LJ, Eng ZR, Ren ZG, 2008. An efficient ant colony optimization approach to attribute reduction in rough set theory. *Patt Recogn Lett*, 29(9):1351-1357. https://doi.org/10.1016/j.patrec.2008.02.006

Ke WJ, Wu CX, Wu Y, et al., 2018. A new filter feature selection based on criteria fusion for gene microarray data. *IEEE Access*, 6:61065-61076. https://doi.org/10.1109/ACCESS.2018.2873634

Kečo D, Subasi A, Kevric J, 2018. Cloud computing-based parallel genetic algorithm for gene selection in cancer classification. *Neur Comput Appl*, 30(5):1601-1610. https://doi.org/10.1007/s00521-016-2780-z

Khan AH, Sarkar SS, Mali KK, et al., 2022. A genetic algorithm based feature selection approach for microstructural image classification. *Exp Techn*, 46(2):335-347. https://doi.org/10.1007/s40799-021-00470-4

Ling Y, Zhou YQ, Luo QF, 2017. Lévy flight trajectory-based whale optimization algorithm for global optimization. *IEEE Access*, 5:6168-6186. https://doi.org/10.1109/ACCESS.2017.2695498

Liu M, Yao XF, Li YX, 2020. Hybrid whale optimization algorithm enhanced with Lévy flight and differential evolution for job shop scheduling problems. *Appl Soft Comput J*, 87:105954. https://doi.org/10.1016/j.asoc.2019.105954

Lokeswari YV, Jacob SG, 2017. Prediction of child tumours from microarray gene expression data through parallel gene selection and classification on spark. In: Behera HS, Mohapatra DP (Eds.), Computational Intelligence in Data Mining. Springer, Singapore, p.651-661. https://doi.org/10.1007/978-981-10-3874-7_62

Maldonado S, López J, 2018. Dealing with high-dimensional class-imbalanced datasets: embedded feature selection for SVM classification. *Appl Soft Comput*, 67:94-105. https://doi.org/10.1016/j.asoc.2018.02.051

Maldonado S, Weber R, 2011. Embedded feature selection for support vector machines: state-of-the-art and future challenges. Proc 16th Iberoamerican Congress Conf on Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, p.304-311. https://doi.org/10.1007/978-3-642-25085-9_36

Maldonado S, Weber R, Famili F, 2014. Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines. *Inform Sci*, 286:228-246. https://doi.org/10.1016/j.ins.2014.07.015

Mangal A, Holm EA, 2018. A comparative study of feature selection methods for stress hotspot classification in materials. *Integr Mater Manuf Innov*, 7(3):87-95. https://doi.org/10.1007/s40192-018-0109-8

Mazumder DH, Veilumuthu R, 2019. An enhanced feature selection filter for classification of microarray cancer data. *ETRI J*, 41(3):358-370. https://doi.org/10.4218/etrij.2018-0522

McCall J, 2005. Genetic algorithms for modelling and optimisation. *J Comput Appl Math*, 184(1):205-222. https://doi.org/10.1016/j.cam.2004.07.034

Mirjalili S, Lewis A, 2016. The whale optimization algorithm. *Adv Eng Softw*, 95:51-67. https://doi.org/10.1016/j.advengsoft.2016.01.008

Mirjalili S, Gandomi AH, Mirjalili SZ, et al., 2017. Salp Swarm Algorithm: a bio-inspired optimizer for engineering design problems. *Adv Eng Softw*, 114:163-191. https://doi.org/10.1016/j.advengsoft.2017.07.002

Mirjalili SZ, Mirjalili S, Saremi S, et al., 2018. Grasshopper optimization algorithm for multi-objective optimization problems. *Appl Intell*, 48(4):805-820. https://doi.org/10.1007/s10489-017-1019-8

Morán-Fernández L, Bolón-Canedo V, Alonso-Betanzos A, 2017. Centralized vs. distributed feature selection methods based on data complexity measures. *Knowl-Based Syst*, 117:27-45. https://doi.org/10.1016/j.knosys.2016.09.022

Nakamura RYM, Pereira LAM , Costa KA, et al., 2012. BBA: a binary bat algorithm for feature selection. Proc 25th SIBGRAPI Conf on Graphics, Patterns and Images, p.291-297. https://doi.org/10.1109/SIBGRAPI.2012.47

Olsson JOS, Oard DW, 2006. Combining feature selectors for text classification. Proc 15th ACM Int Conf on Information and Knowledge Management, p.798-799. https://doi.org/10.1145/1183614.1183736

Payne AWR, Glen RC, 1993. Molecular recognition using a binary genetic search algorithm. *J Mol Graph*, 11(2):74-91. https://doi.org/10.1016/0263-7855(93)87001-L

Peng HC, Long FH, Ding C, 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Patt Anal Mach Intell*, 27(8):1226-1238. https://doi.org/10.1109/TPAMI.2005.159

Potharaju SP, Sreedevi M, 2018. Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance. *Clin Epidemiol Glob Heal*, 7(2):171-176. https://doi.org/10.1016/j.cegh.2018.04.001

Prasad Y, Biswas KK, Hanmandlu M, 2018. A recursive PSO scheme for gene selection in microarray data. *Appl Soft Comput*, 71:213-225. https://doi.org/10.1016/j.asoc.2018.06.019

Qaraad M, Amjad S, Manhrawy IIM, et al., 2021. A hybrid feature selection optimization model for high dimension data classification. *IEEE Access*, 9:42884-42895. https://doi.org/10.1109/ACCESS.2021.3065341

Ragunthar T, Selvakumar S, 2019. A wrapper based feature selection in bone marrow plasma cell gene expression data. *Clust Comput*, 22(6):13785-13796. https://doi.org/10.1007/s10586-018-2094-2

Rahimipour J, Usefi A, 2019. A comparative study of feature selection methods on genomic datasets. Proc IEEE 32nd Int Symp on Computer-based Medical Systems, p.471-476. https://doi.org/10.1109/CBMS.2019.00097

Ram PK, Kuila P, 2019. Feature selection from microarray data: genetic algorithm based approach. *J Inform Optim Sci*, 40(8):1599-1610. https://doi.org/10.1080/02522667.2019.1703260

Rani MJ, Devaraj D, 2019. Two-stage hybrid gene selection

using mutual information and genetic algorithm for cancer data classification. *J Med Syst*, 43(8):235. https://doi.org/10.1007/s10916-019-1372-8

Ranjani R, Ramyachitran D, 2018. Microarray cancer gene feature selection using spider monkey optimization algorithm and cancer classification using SVM. *Proc Comput Sci*, 143:108-116. https://doi.org/10.1016/j.procs.2018.10.358

Rathee S, Ratnoo S, Ahuja J, 2022. Feature selection using PMOGA for microarray datasets. *J Sci Res*, 66(1):375-385. https://doi.org/10.37398/JSR.2022.660140

Ray RB, Kumar M, Rath SK, 2016a. Fast computing of microarray data using resilient distributed dataset of Apache Spark. In: Meesad P, Boonkrong S, Unger H (Eds.), Recent Advances in Information and Communication Technology. Springer, Cham, p.171-182. https://doi.org/10.1007/978-3-319-40415-8_17

Ray RB, Kumar M, Rath SK, 2016b. Fast in-memory cluster computing of sizeable microarray using spark. Int Conf on Recent Trends in Information Technology, p.1-6. https://doi.org/10.1109/ICRTIT.2016.7569599

Remeseiro B, Bolon-Canedo V, 2019. A review of feature selection methods in medical applications. *Comput Biol Med*, 112:103375. https://doi.org/10.1016/j.compbiomed.2019.103375

Saeys Y, Inza I, Larrañaga P, 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507-2517. https://doi.org/10.1093/bioinformatics/btm344

Sahu B, Dehuri S, Jagadev AK, 2017. Feature selection model based on clustering and ranking in pipeline for microarray data. *Inform Med Unlocked*, 9:107-122. https://doi.org/10.1016/j.imu.2017.07.004

Sakae Y, Straub JE, Okamoto Y, 2019. Enhanced sampling method in molecular simulations using genetic algorithm for biomolecular systems. *J Comput Chem*, 40(2):475-481. https://doi.org/10.1002/jcc.25735

Saw T, Myint P, 2019. Swarm intelligence based feature selection for high dimensional classification: a literature survey. *Int J Comput*, 33(1):69-83.

Seijo-Pardo B, Porto-Díaz I, Bolón-Canedo V, et al., 2017. Ensemble feature selection: homogeneous and heterogeneous approaches. *Knowl-Based Syst*, 118:124-139. https://doi.org/10.1016/j.knosys.2016.11.017

Shadravan S, Naji HR, Bardsiri VK, 2019. The sailfish optimizer: a novel nature-inspired metaheuristic algorithm for solving constrained engineering optimization problems. *Eng Appl Artif Intell*, 80:20-34. https://doi.org/10.1016/j.engappai.2019.01.001

Shalabi L, 2022. New feature selection algorithm based on feature stability and correlation. *IEEE Access*, 10:4699-4713. https://doi.org/10.1109/ACCESS.2022.3140209

Shao LS, Bai Y, Qiu YF, et al., 2012. Particle swarm optimization algorithm based on semantic relations and its engineering applications. *Syst Eng Proc*, 5:222-227. https://doi.org/10.1016/j.sepro.2012.04.035

Shukla AK, Tripathi D, 2019. Identification of potential biomarkers on microarray data using distributed gene selection approach. *Math Biosci*, 315:108230. https://doi.org/10.1016/j.mbs.2019.108230

Shukla AK, Singh P, Vardhan M, 2019. A new hybrid feature subset selection framework based on binary genetic algorithm and information theory. *Int J Comput Intell Appl*, 18(3):1950020. https://doi.org/10.1142/s1469026819500202

Siedlecki W, Sklansky J, 1989. A note on genetic algorithms for large-scale feature selection. *Patt Recogn Lett*, 10(5):335-347. https://doi.org/10.1016/0167-8655(89)90037-8

Sihwail R, Omar K, Ariffin KAZ, et al., 2020. Improved Harris hawks optimization using elite opposition-based learning and novel search mechanism for feature selection. *IEEE Access*, 8:121127-121145. https://doi.org/10.1109/ACCESS.2020.3006473

Sönmez ÖS, Dağtekin M, Ensari T, 2021. Gene expression data classification using genetic algorithm-basedfeature selection. *Turk J Electr Eng Comput Sci*, 29(7):3165-3179. https://doi.org/10.3906/elk-2102-110

Storn R, Price K, 1997. Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *J Glob Optim*, 11(4):341-359. https://doi.org/10.1023/A:1008202821328

Sun YJ, Wang XL, Chen YH, et al., 2018. A modified whale optimization algorithm for large-scale global optimization problems. *Expert Syst Appl*, 114:563-577. https://doi.org/10.1016/j.eswa.2018.08.027

Tadist K, Najah S, Nikolov NS, 2019. Feature selection methods and genomic big data: a systematic review. *J Big Data*, 6(1):79. https://doi.org/10.1186/s40537-019-0241-0

Tawhid MA, Ibrahim AM, 2020. Feature selection based on rough set approach, wrapper approach, and binary whale optimization algorithm. *Int J Mach Learn Cybern*, 11(3):573-602. https://doi.org/10.1007/s13042-019-00996-5

Tsai CF, Sung YT, 2020. Ensemble feature selection in high dimension, low sample size datasets: parallel and serial combination approaches. *Knowl-Based Syst*, 203:106097. https://doi.org/10.1016/j.knosys.2020.106097

Tubishat M, Abushariah MAM, Idris N, et al., 2019. Improved whale optimization algorithm for feature selection in Arabic sentiment analysis. *Appl Intell*, 49(5):1688-1707. https://doi.org/10.1007/s10489-018-1334-8

Tubishat M, Ja'afar S, Alswaitti M, et al., 2021. Dynamic Salp swarm algorithm for feature selection. *Expert Syst Appl*, 164:113873. https://doi.org/10.1016/j.eswa.2020.113873

Urbanowicz RJ, Meeker M, La Cava W, et al., 2017. Relief-based feature selection: introduction and review. *J Biomed Inform*, 85:189-203. https://doi.org/10.1016/j.jbi.2018.07.014

van Hal NLW, Vorst O, van Houwelingen AMML, et al., 2000. The application of DNA microarrays in gene expression analysis. *J Biotechnol*, 78(3):271-280. https://doi.org/10.1016/S0168-1656(00)00204-2

Venkataramana L, Jacob SG, Ramadoss R, et al., 2019. Improving classification accuracy of cancer types using parallel hybrid feature selection on microarray gene expression data. *Genes Genom*, 41(11):1301-1313. https://doi.org/10.1007/s13258-019-00859-x

Vergara JR, Estévez PA, 2014. A review of feature selection methods based on mutual information. *Neur Comput Appl*, 24(1):175-186. https://doi.org/10.1007/s00521-013-1368-0

Wang AG, Liu HC, Yang J, et al., 2022. Ensemble feature selection for stable biomarker identification and cancer classification from microarray expression data. *Comput Biol Med*, 142:105208. https://doi.org/10.1016/J.COMPBIOMED.2021.105208

Windeatt T, Duangsoithong R, Smith R, 2011. Embedded feature ranking for ensemble MLP classifiers. *IEEE Trans Neur Netw*, 22(6):988-994. https://doi.org/10.1109/TNN.2011.2138158

Xie WD, Chi YH, Wang LJ, et al., 2021. MMBDE: a two-stage hybrid feature selection method from microarray data. IEEE Int Conf on Bioinformatics and Biomedicine, p.2346-2351. https://doi.org/10.1109/BIBM52615.2021.9669496

Xuan GR, Zhu XM, Chai PQ, et al., 2006. Feature selection based on the Bhattacharyya distance. Proc 18th Int Conf on Pattern Recognition, p.957-957. https://doi.org/10.1109/ICPR.2006.557

Yang F, Mao KZ, 2011. Robust feature selection for microarray data based on multicriterion fusion. *IEEE/ACM Trans Comput Biol Bioinform*, 8(4):1080-1092. https://doi.org/10.1109/TCBB.2010.103

Ye XC, Li HM, Imakura A, et al., 2019. Distributed collaborative feature selection based on intermediate representation. Proc 28th Int Joint Conf on Artificial Intelligence, p.4142-4149. https://doi.org/10.24963/ijcai.2019/575

Yuan MS, Yang ZJ, Ji GL, 2019. Partial maximum correlation information: a new feature selection method for microarray data classification. *Neurocomputing*, 323:231-243. https://doi.org/10.1016/j.neucom.2018.09.084

Zare M, Eftekhari M, Aghamollaei G, 2019. Supervised feature selection via matrix factorization based on singular value decomposition. *Chemom Intell Lab Syst*, 185:105-113. https://doi.org/10.1016/j.chemolab.2019.01.003

Zhang G, Hou JC, Wang JL, et al., 2020. Feature selection for microarray data classification using hybrid information gain and a modified binary krill herd algorithm. *Interdisc Sci Comput Life Sci*, 12(3):288-301. https://doi.org/10.1007/s12539-020-00372-w

Zhang L, Huang XJ, 2015. Multiple SVM-RFE for multi-class gene selection on DNA microarray data. Int Joint Conf on Neural Networks, p.1-6. https://doi.org/10.1109/IJCNN.2015.7280417

Zhang R, Nie FP, Li XL, et al., 2019. Feature selection with multi-view data: a survey. *Inform Fus*, 50:158-167. https://doi.org/10.1016/j.inffus.2018.11.019

Zheng CH, Huang DS, Shang L, 2006. Feature selection in independent component subspace for microarray data classification. *Neurocomputing*, 69(16-18):2407-2410. https://doi.org/10.1016/j.neucom.2006.02.006

Zhu HQ, Bi N, Tan J, et al., 2018. An embedded method for feature selection using kernel parameter descent support vector machine. Proc 1st Chinese Conf on Pattern Recognition and Computer Vision, p.351-362. https://doi.org/10.1007/978-3-030-03338-5_30