Computers in Biology and Medicine 140 (2022) 105051

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Feature selection methods on gene expression microarray data for cancer classification: A systematic review

Esra'a Alhenawi [a], Rizik Al-Sayyed [a,*], Amjad Hudaib [a], Seyedali Mirjalili [b,c]

[a] King Abdullah II School for Information Technology, The University of Jordan, Amman, Jordan
[b] Center for Artificial Intelligence Research and Optimization, Torrens University Australia, Fortitude Valley, Brisbane, 4006, QLD, Australia
[c] Yonsei Frontier Lab, Yonsei University, Seoul, South Korea

ARTICLE INFO

Keywords:
Feature selection
Filters
Wrappers
Embedded techniques
Hybrid
Ensemble

ABSTRACT

This systematic review provides researchers interested in feature selection (FS) for processing microarray data with comprehensive information about the main research directions for gene expression classification conducted during the recent seven years. A set of 132 researches published by three different publishers is reviewed. The studied papers are categorized into nine directions based on their objectives. The FS directions that received various levels of attention were then summarized. The review revealed that 'propose hybrid FS methods' represented the most interesting research direction with a percentage of 34.9%, while the other directions have lower percentages that ranged from 13.6% down to 3%. This guides researchers to select the most competitive research direction. Papers in each category are thoroughly reviewed based on six perspectives, mainly: method (s), classifier(s), dataset(s), dataset dimension(s) range, performance metric(s), and result(s) achieved.

1. Introduction

Nowadays, classifying microarray datasets, which are called (large-scale biological data analysis), is a popular and attractive area of study for many researchers, as applying microarray technology is one of the most important applications in molecular biology for cancer detection [1]. It depends on developing more effective classification models that can be used for classifying any unseen microarray data after training the model over a specific training dataset. Detecting and classifying cancer, using microarray gene expressed data, have posed a huge challenge for researchers in the field of computer science, as this kind of datasets contains a small number of examples versus a huge number of genes. However, many of these genes are considered irrelevant or redundant, and they must be removed by using an efficient FS method for improving the performance of classification. Therefore, researchers have employed much effort in coming up with more effective FS techniques that can increase classification's accuracy and decrease the computation time using a smaller number of genes in diagnostic and prognostic prediction of tumor cancer [2].

FS is the process of selecting the most relevant and efficient features for improving classification's performance in high dimensional datasets [3]. Filters, wrappers, embedded, and hybrid methods are the main types of FS methods, but there is a new kind of FS methods that has been recently developed, which is called ensemble [4].

Filters use statistical measures for evaluating features against a class label. There are two categories of filters, mainly: ranking-based (univariate) and search-space-based (multivariate). The first category selects features that have higher ranks based on a specific threshold value, where ranks are provided according to the relationships between each feature and the specified class label for removing the most irrelevant features. In contrast, the second category takes care of the relationships within features. Therefore, it can remove irrelevant features in addition to redundant ones [5].

Wrappers depend on a classifier evaluation to select features as it selects feature sets that satisfy best results based on a fitness value for a classifier. This kind of methods consists of three parts: search algorithm, classifier, and fitness function [6], while embedded methods select features that automatically improve classification performance as a part of the learning stage [2]. See Fig. 1.

Ensemble methods have recently appeared for obtaining stability which does not exist in many FS methods. This will be achieved by aggregating the results of different feature subsets which were generated either by using the same FS method on various training data (homogeneous ensemble FS approach) or by applying various FS methods over

* Corresponding author.
E-mail addresses: esra_a_2008@live.com (E. Alhenawi), r.alsayyed@ju.edu.jo (R. Al-Sayyed), ahudaib@ju.edu.jo (A. Hudaib), ali.mirjalili@gmail.com (S. Mirjalili).

https://doi.org/10.1016/j.compbiomed.2021.105051
Received 9 July 2021; Received in revised form 1 November 2021; Accepted 15 November 2021
Available online 23 November 2021

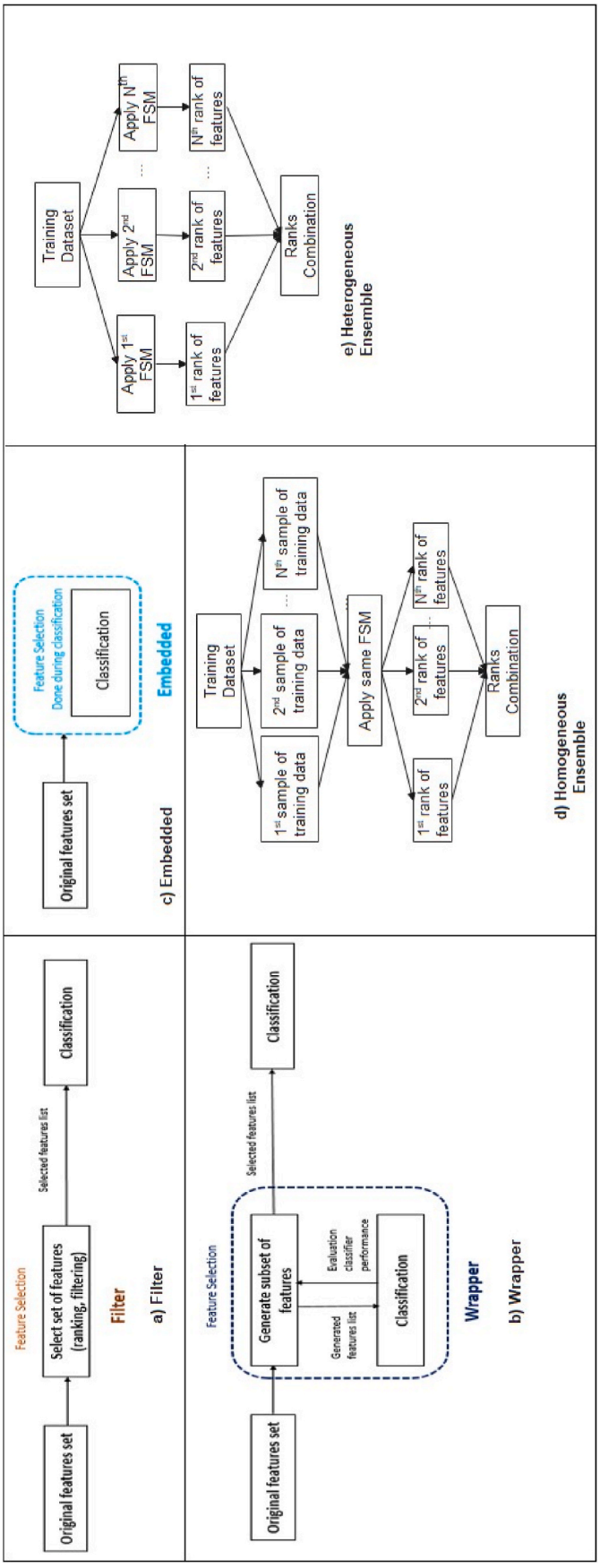

Fig. 1. FS Research main directions for microarray data.

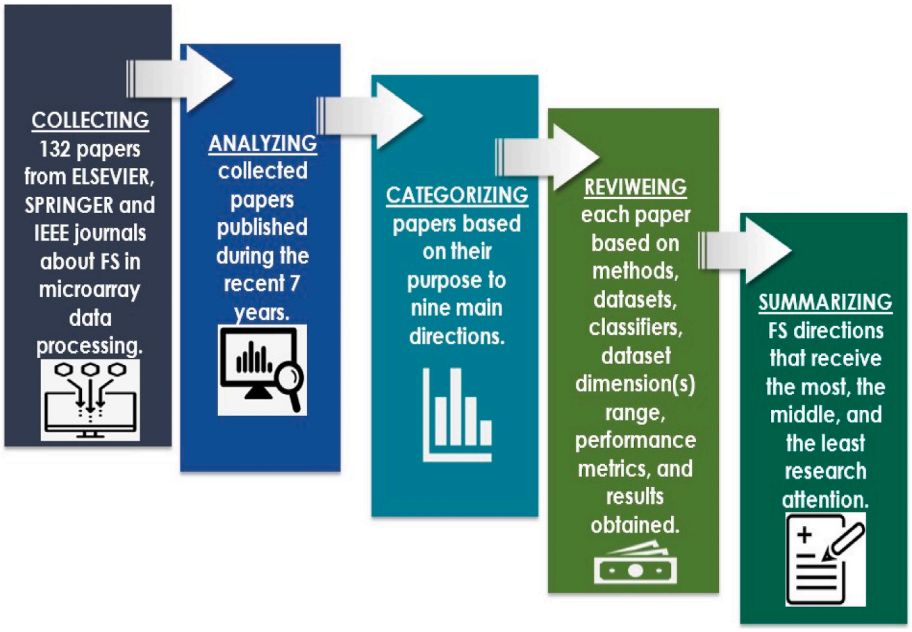

Fig. 2. Methodology.

the same training data (heterogeneous ensemble FS approach) [7]. See Fig. 1 (d and e).

Each approach has its negative and positive properties. For example, filter-based FS methods need less computational complexity and avoid over-fitting problems more than wrapper and embedded-based FS methods. In contrast, wrappers and embedded-based FS methods provide better accuracy than filters. Hybrid methods have advantages of both wrappers and filters as they result in achieving better accuracy than filters, and need less computational cost than wrappers. Ensemble methods are the most flexible of FS methods in high dimensional data, and are the least prone to over-fitting problems [2].

In literature, there are many reviews that concern FS in microarray data subjects, such as the works of [8–23], but to the best of our knowledge, there are no reviews done on categorizing these studies based on the main purpose of the reviewed papers.

In this systematic review, we categorize studies done recently on FS for microarray data processing into nine directions based on the main objectives of each research. Where some of these studies developed new FS methods which are (filter, wrapper, embedded, ensemble, distributed, parallel), other studies selected some FSM from the literature and compared their performance over the same environment. On the other hand, there are some researchers who conducted surveys on the subjects of FS.

Each research in each direction was reviewed based on methods, datasets, classifiers, performance metrics, datasets dimension's range used, and results obtained. Finally, we summarized some of the general observations that we noticed along with our conclusions.

The main contributions of this systematic review can be summarized as follows:

- It analyses 132 research papers from Elsevier, Springer, and IEEE publishers in the literature published in the last seven years about FS in microarray data processing according to their main objectives and categorizes these researches into nine main directions.
- It summarizes FS directions that received the highest, the middle, and the lowest research attention in the recent seven years for guiding researchers who plan to pursue research on FS for microarray data processing to choose their research direction.
- In each direction, each research paper is reviewed based on methods, datasets, classifiers, performance metrics, datasets dimensions range used, and results obtained. This aims to provide researchers with valuable information about the related work of their selected directions if they have already specified their research direction in FS for microarray data classification field as displayed in Fig. 2

The rest of this paper is organized as follows: Section 2 presents a background. Section 3 displays the methodology. Section 4 displays and discusses the main nine directions of FS in microarray data researches in the recent seven years that we proposed in this survey. The development of FS publications over the recent seven years is displayed in Section 5. In Section 6, the main observations are identified, and in Section 7, the conclusion and future work are presented.

2. Background

2.1. Gene expression microarray data

It is a structured medical data, where features in gene expression microarray datasets represent gene expression coefficients in samples for each instance that represents a patient. Usually, microarray datasets are highly dimensional, as they contain a huge number of features versus a small number of samples [24].

2.2. Importance of FS to microarray datasets analysis

Detecting cancer-infected genes and normal healthy genes from the microarray dataset is challenging in high dimensional microarray datasets which contain many redundant and irrelevant genes that negatively affect the classification's performance. Therefore, researchers competed for developing more efficient FS methods for selecting the most informative genes from a huge number of genes [8].

2.3. FS methods

It is used for detecting and removing irrelevant and redundant features from high dimensional datasets, such as microarray data, for the purpose of improving classification's performance and decreasing the computation time. These metrics represent a challenge for all researchers in the field of FS methods regarding high dimensional datasets. FS methods can be grouped based on different perspectives, and

they can be classified based on the availability of supervision, selection strategy, or data perspective [25].

2.3.1. FS based on the availability of supervision

FS methods are grouped into three groups: Unsupervised, Semi-supervised, Supervised, based on the availability of class labels in the samples of datasets.

- Unsupervised FS:

It is unassisted by labeled classes for FS and classification processes. This approach almost cannot produce optimal subsets, as it neglects the possible correlation between features. It also uses some mathematical principles without guaranteeing that the principles are universally valid for all data [26]. Conversely, it does not need any prior knowledge to classify new samples. Therefore, it is considered unbiased [27].

- **Semi-supervised FS** Semi-supervised is an extension of the supervised group that works on both labeled and unlabeled data. Usually, the margins between data points of different classes are maximized using labeled data, and the geometrical structure of the space is discovered using unlabeled data [28].
- **Supervised FS** In the past thirty years, the importance of the supervised feature selection exploded in machine learning for analyzing datasets with high dimensions such as microarray datasets [29]. It depends on the labeled data for FS process and classification process.

2.3.2. FS based on selection strategy

FS methods are grouped into five groups: Filter, Wrapper, Embedded, Hybrid, and Ensemble, based on the selection strategy.

- **Filter** selects features based on the statistical properties of the data without using any learning models. There are two categories of filters: Univariate and Multiivariate, and they are used to evaluate the relevance of any feature.
  * **Univariate Filter** It evaluates and independently provides ranks for each feature using a specific criterion for selecting the k-features with the highest ranks. This type of filters cannot detect the redundant features as it ignores dependencies between features which affect the classification results compared to multiivariate strategy. On the other hand, univariate filters are faster than multiivariate ones. Fisher Score, Mutual Information, and Laplacian Score (LS) are the most famous univariate filters that were used in some papers in this survey.
    1 Fisher Score

It selects features that have a higher fisher score, which compute the discriminative power of each feature [30].

2 Mutual information

It measures the amount of information that can be obtained about one random variable using another random variable. In FS, it works by calculating the statistical dependence between any two random features [31].

3 Laplacian Score (LS) This method can be used with any dataset from any supervision fusion. It evaluates features based on their locality in preserving power as it depends on the assumption that data belongs to the same class that can be found near to each other in many classification problems [32].

* **Multiivariate Filter** It evaluates features in the context of other features. Therefore, this kind of filters can assess both irrelevant and redundant features and provides better classification performance than other kinds of filters. ReliefF and Minimal redundancy maximal relevance are the most famous univariate filters that were used in some papers in this survey.

1 ReliefF

It selects random instances then searches for a specific number of the nearest neighbors that have the same classes (hits) and for K nearest neighbors that have different classes (misses). After that, the average of all hits and misses are computed by repeating this process for a specific time for each feature [33].

2 Minimal redundancy maximal relevance

It selects features that are more relevant (have a large score of correlation with the target class), and not redundant (have a small correlation score with other features) [34].

- **Wrapper** It is one of the classifier-based FS methods. It works by selecting a subset of features that provides the best results of a specific learning model. Wrappers can be classified into greedy and stochastic search strategy [35,36]. The greedy search strategy (such as forward selection and sequential backward selection) can trap in a local optimum as it is a single-track search, while the stochastic search strategy deploys the randomness nature in FS such as using some of meta-heuristic algorithms.
  * **Meta-heuristic algorithms** Here we briefly describe some of the most commonly used meta-heuristic algorithms, such as Simulated Annealing (SA), Genetic Algorithm (GA), particle swarm optimization (PSO), and Ant Colony Optimization (ACO). There are other bio-inspired meta-heuristic algorithms that were used in some papers in this survey such as (Artificial Bee Colony (ABC), Firefly Algorithm (FF), Krill Herd Algorithm (KHA), and Whale optimization algorithm (BWOA)) that were developed originally in 2005, 2008, 2012, and 2016, respectively.
    1 **SA**

Simulated annealing was originally developed by Ref. [37] as a trajectory-based meta-heuristic. It simulates the annealing process that is used to harden metals, starting with a high temperature that is gradually cooled down based on the specific cooling rate value until reaching the final temperature. The algorithm begins with an initial single solution that is randomly generated, then it updates it with a new neighboring solution if it is better in each iteration until reaching the final temperature [38].

2 **GA**

It is an evolutionary-based meta-heuristic algorithm that was developed in 1992. It simulates the process of natural selection. It searches for the best solutions for any research problem using mutation, crossover, and selection operators [39].

3 **PSO**

It is a swarm-based meta-heuristic algorithm that was originally developed by Kennedy et al. [40]. It simulates the social behavior of organisms in a bird flock or a fish school.

4 **ACO**

It is a swarm-based meta-heuristic algorithm that was proposed in 1992 by an author in Ref. [41]. It depends on the behavior of ants in ant colony through their journey of searching for food.

- **Embedded** It works by training specific learning models using an initial subset of features in order to establish a criterion for computing the rank values of features [42].
- **Hybrid** It is a classifier-based FS method that was developed to combine advantages of both wrapper and filter FS methods. This type of FS methods is usually applied through two stages. In the first stage, filter methods or an ensemble of filters are used to decrease the number of features that will be passed to the wrapper stage, where classification performance will be increased with an acceptable computation time [42].
- Ensemble

It is developed for exploiting advantages of multiple FS methods which may obtain better results than the output of any individual method. An ensemble FS method has two categories: Homogeneous and Heterogeneous. The first category works by splitting the dataset into different training datasets, then by applying the same FS method. In Heterogeneous strategy, different FS methods are applied over the same training data. The final set of selected features in both categories is generated using specific combining and thresholding processes. Combination can be done using union, intersection, or voting processes. Thresholding can be done using static thresholds based on a fixed percentage, or by using complexity measures automatically [7].

2.4. Classifiers

This subsection provides an overlook at some of most popular classifiers which were mentioned in this survey.

2.4.1. Support vector machine (SVM)

It is a computer algorithm that learns by example to assign labels to objects [43]. It is a sophisticated and computationally expensive classifier. The original domain can be divided linearly (straight hyper-plane) using simple mathematical models SVM called linear. However, if this domain cannot be divided linearly, then the data domain will be mapped into a feature space using a kernel function [44]. Then, the feature space will be mapped into a response set to divide the data domain [45].

2.5. Naive Bayes (NB)

NB is one of the most popular supervised classification algorithms, especially in the medical science field, as it is simple and easy to build without any complicated iterative parameter estimation. It applies a Bayes' theorem with strong independence among features [46].

2.5.1. K-nearest neighbor (KNN)

It is a supervised classification algorithm. It finds a label for a new point by looking at the nearest labeled (K) points [47]. KNN ranks depend on some similar scores such as Euclidean distance measure [48].

2.5.2. Decision tree (DT)

DT represents datasets as trees. This is done by organizing instances as a decision or leaf nodes. Decision nodes have two branches, and leaf nodes contain one decision [49].

2.5.3. C4.5

It is developed by Ref. [50]. It is an extension of the Iterative Dicotomiser 3, where both algorithms are based on decision trees.

3. Methodology

This survey was conducted by applying five main steps as summarized in Fig. 2:

- At the beginning, we collected 132 papers done on FS in microarray data processing from three popular publishers (Elsevier, Springer, and IEEE) during the last seven years.
- These papers were analyzed and distributed into nine main directions based on their main objectives as new FS methods do not exist in the literature. Surveys were conducted and performance was compared based on some of the existing FS methods.
- Each paper was reviewed based on six perspectives: method(s), classifier(s), dataset(s), dataset dimension(s) range, performance metric(s), and result(s) achieved.
- Directions with the least, middle, and most attention were summarized in the literature.
- Some recommendations were given for researchers who intent to pursue research on FS for microarray data processing, based on our observations.

4. FS in microarray data researches main directions

In the literature, there are many studies done on FS for microarray data processing which we can categorize into nine main research directions as illustrated in Fig. 3. In each direction, papers were grouped and connected based on two levels. At the first level, these papers were arranged in the provided order presented in the presentation according to specific criteria that varies from direction to another (for example, D3 and D5 connected papers that used the same meta-heuristic algorithm together). At the next level, they were connected based on the year of publication, in order to help researchers making sense of how feature selection of publications was developed during the recent seven years.

4.1. First direction (D1)

It is a survey conducted on FS in microarray data processing. One of the recent surveys was done by Manikandan and Abirami [8], where they reviewed some of the art filters, wrappers, embedded, and hybrid FS methods. They presented the importance of FS over various applications such as microarray data analysis.

In 2019, Authors in Ref. [9] reviewed FS methods (FSM) and then compared them based on datasets, feature selection methods, classifiers, and accuracy results. They also applied three filter methods, mainly: *t*-Test, Pearson's Correlation Coefficient (PCC), and Bhattacharyya

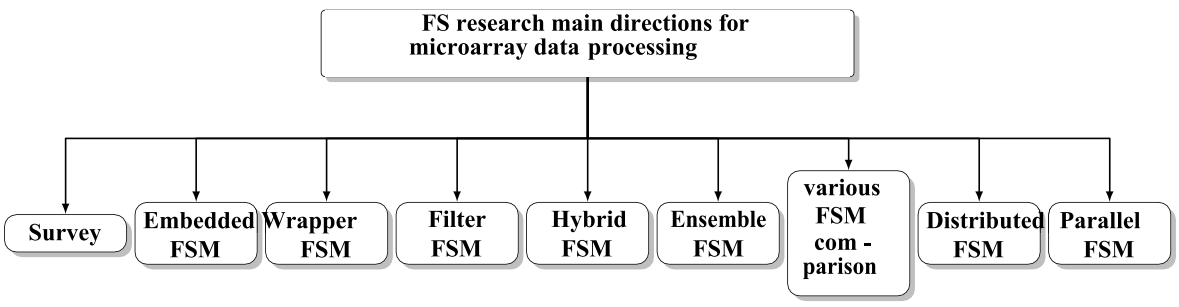

Fig. 3. FS research main directions for microarray data.

distance, in order to eliminate irrelevant features and compare the effects of combining filters with wrapper and embedded methods. They concluded that SVM is the most common classifier used in most of the studies they collected. Additionally, there are no other ways to evaluate what method is the best for each cancer dataset. Experimental results show that using filters improve classification's performance by increasing the accuracy when running time has been decreased.

Shukla et al. [10] provided a survey about the meta-heuristic approaches used for gene selection in microarray data, including: GA, PSO, and ACO. They applied four meta-heuristic FS methods, which are called ACO, PSO, Differential Evolution (DE), and GA. These methods were applied on five microarray datasets with SVM and KNN as classifiers and one regression Lasso approach to demonstrate its applicability in specific situations. Some recommendations were then provided based on the results that were got, in order to guide researchers to choose a suitable FS method. Open challenges for microarray data processing were discussed, such as scalability, computational cost, search mechanisms, measures, and dataset structure.

Almugren, and Alshamlan [15] conducted a comparative survey of hybrid methods that used bio-inspired evolutionary methods for the wrapper part. They concluded that hybrid feature selection methods give a superior performance as they avoid over-fitting problems in microarray datasets due to the high dimensionality, because this type of datasets has a small number of samples versus a large number of features. On the other hand, they discussed the parameters of fitting processes which affect the classification accuracy. They noticed that the fitting parameter depends on gene expression datasets and the applied feature selection methods. Therefore, each dataset has specific parameters, which are manually adjusted in most of the studies in order to get accepted classification results.

After 2018, Zamri et al. [14] provided an overview of some work that has been carried out on FS based on swarm intelligence for microarray data expressions in cancer detection in the literature. They conducted their review by examining eleven studies published between 2010 and 2016, where they mentioned that PSO, ACO, and Artificial Bee Colony (ABC) are the most popular swarm intelligence algorithms used for microarray data processing.

Alonso-Betanzos et al. [11] discussed how feature selection improves the classification accuracy in different microarray data scenarios through displaying four different study cases: using a previous discretization step, carrying out data complexity analysis, employing distributed approaches, and finally, using ensemble approaches for the study of DNA microarray datasets.

Table 1
Summary of surveys from 2019 to 2021.

Reference	Summary
[8]	Reviewed some filters, wrappers, embedded, and hybrid FS methods in the literature and presented the importance of FS over various applications like microarray data analysis.
[9]	Provided a comparative review of FS methods, datasets, classifiers, and the best-obtained results for a microarray data processing studies done in the literature.
[10]	Presented a comprehensive survey about studies that adopted a meta-heuristic approach for feature selection in microarray data classification.
[11]	Discussed four different study cases: using a previous discretization step, carrying out a data complexity analysis, employing distributed approaches, and finally, using an ensemble FS approaches for microarray datasets.
[12]	Presented an exhaustive review of the most successful techniques and datasets applied in FS for microarray data.
[17]	Discussed the main challenges and the future trends for microarray data analysis.
[15]	Presented a comparative review of the most recent hybrid-based FS methods that used a bio-inspired evolutionary algorithms as a wrapper in the literature.
[51]	Analyzed some of the recent studies that were done based on neural network models for cancer prediction. They displayed the practical issues that can be considered to increase models predictability in the future.

Table 2
Summary of surveys from 2014 to 2018.

Reference	Summary
[18]	Reviewed some of the studies concerning FS techniques, datasets, and performance metrics employed in microarray data processing for cancer detection in the literature.
[19]	Provided a detailed review of various gene selection methods based on different aspects.
[23]	Collected some state-of-the-art reviews about microarray data processing and categorized FS methods in these reviews based on an extended taxonomy.
[14]	Reviewed some of the most recent swarm intelligence usage in FS for microarray data expressions in cancer detection and classification. They also provided an overview of microarray data sources that were used in the literature.
[13]	Reviewed some of the FS techniques that were used in microarray datasets for cancer detection in the literature.
[22]	Reviewed some of the FS methods that were developed in DNA microarray classification field.
[16]	Reviewed some of the feature selection methods in the literature as supervised, unsupervised, and semi-supervised FS methods.
[20]	Provided an overview of the FS methods for microarray datasets and categorized these methods into three categories: filters, wrappers, and embedded FS methods.
[21]	Conducted a comprehensive survey about the most recent FS techniques, challenges, and critical problems of FS in the machine learning field. They concluded that FS methods improve the dataset performance, especially the hybrid (filter – wrapper) FS approach.

Researchers in Ref. [12] provided an overview of the most used techniques for classifying microarray datasets from 2010 based on the types of datasets used as binary or multiclass datasets, classifiers, preprocessing, and validation techniques used in each study. In addition, they summarized the most used microarray datasets, their sources, and the most common classifiers used in each of them.

Authors in Ref. [51] discussed recent studies that were done based on neural network models for cancer prediction. They analyzed all studies according to Neural networks used, gene expression datasets, data preprocessing, model configuration, learning parameters, and evaluation metrics. They finally displayed the practical issues that can be considered to increase models predictability in the future.

Singh et al. [13] discussed some advances of feature selection techniques that have been used in microarray data analysis. They noticed that there are no feature selection algorithms suitable for all datasets as each feature selection technique has its own advantages and disadvantages.

In [20], authors provided an overview on some filters as well as wrapper and embedded FS methods used for microarray datasets in the literature. Authors in Ref. [23] generated an extended taxonomy of FS methods and used it to categorize FS methods that were mentioned in some of the-state-of-the-art reviews that they have collected.

Bolon et al. [22] reviewed some of FSM that were developed in DNA Microarray classification field with an experimental study of the most significant algorithms and evaluation techniques.

Some surveys discussed feature selection methods that were used in the supervised, semi-supervised, and unsupervised learning approaches. An example is the study done by the authors in Ref. [16], where they documented the advantages and disadvantages of each gene selection approach and illustrated the challenges and problems in gene selection processes for microarray data processing. They concluded that semi-supervised feature selection is the best approach in microarray data processing because it can handle both labeled and unlabeled data. The hybrid method is better than other methods as it integrated the strength points of more than one method.

There are other works done regarding this direction, where some studies conducted a review about microarray data processing tackling the challenges, limitations, and future trends in the microarray data

processing [17,21]. Other studies conducted their review based on feature selection techniques, datasets, and performance metrics employed in microarray data processing [18,19].

Table 1, and Table 2 present a brief summary of all studies discussed in this direction.

4.1.1. Inferences from first direction (D1)

As can be seen in the literature, most surveys conducted on microarray data analysis present an overview of FS methods. The researchers categorize these methods into four categories: filter, wrapper, embedded, or ensemble. Conversely, some review-papers discuss general challenges, limitations, and future trends of microarray data processing, but there are no reviews that were conducted to categorize previous studies based on their main purpose as we aim to do in this manuscript.

4.2. Second direction (D2)

It consists of embedded FSM. One of the researches in this direction was conducted in 2019 by Kang et al. [52], who developed a relaxed Lasso with Generalized multi-class support vector machine (GenSVM) as a new classification model for tumor classification called relaxed Lasso-GenSVM (rLGenSVM). They used four microarray datasets of the two classes, mainly: Diffuse large B-cell lymphoma (DLBCL), Central Nervous System (CNS), Lung, and Ovarian; and four datasets of multi-class, mainly: Brain, Lymphoma, Mixed-lineage leukemia (MLL), and TOX_171 with GenSVM as a classifier to evaluate the proposed feature selection method. They combined the proposed method with some classifiers such as KNN, L1-regularized regression, and L2-regularized regression. They then compared the obtained results versus the results of using GenSVM classifier based on classification accuracy, where the proposed method provides competitive results based on the number of selected genes and a classification accuracy value.

Some studies have developed new embedded FS methods based on applying some modifications or extensions on the kernel function. An example is the work of Zhu et al. [53] that was done in 2018, where authors have developed a novel embedded method using kernel functions called Kernel Parameter Descent Support Vector Machine (KPD-SVM) for improving the classification performance and selecting the most efficient features by optimizing parameters in kernel function. They used a gradient to find out the best subset of features. For evaluation, they used Sonar and Wisconsin Breast Cancer from the UCI repository. Results indicate that the KPD-SVM method outperforms the F-Score filter-based method, Recursive Feature Elimination Support Vector Machine (RFE-SVM), and Wrapper-based method. KPD-SVM needs less consuming time than other embedded methods as it has novelties in stop criterion and step-size settings in executions.

Moreover, in the same year, authors in Ref. [54] came up with an embedded technique for FS and SVM classification called Kernel Penalized Support Vector Data Description (KP-SVDD) and Kernel Penalized Cost-Sensitive SVM (KP-CSSVM). This was done by extending the idea of Kernel Penalized Support Vector Machine (KP-SVM) to two SVM models for skewed class distribution, which are Support Vector Data Description (SVDD) and Cost-Sensitive SVM (CS-SVM) that apply classification and variable penalization simultaneously to solve a skewed class distribution problem. They used Quasi-Newton strategy for updating the scaling factors. For evaluating the proposed strategy, they deployed twelve microarray datasets with SVM as a classifier.

In 2015, Mishra et al. [55] proposed a method called Support Vector Machine Bayesian T-Test Recursive Feature Elimination algorithm (SVM-BT-RFE). This method combines support vector machine *t*-test recursive feature elimination (SVM-T-RFE) and support vector machine recursive feature elimination (SVM-RFE). They used Colon, Leukemia, Medulloblastoma, Lymphoma, and Prostate datasets along with Support Vector Machine (SVM) to evaluate performance of the proposed method performance versus two methods from the literature. Authors in Ref. [56] developed an embedded FS method called multiple SVM-RFE for multi-class classification (MPMCSVM-RFE), which is a multiple of SVM-RFE for multi-class FS and classifications. The proposed method can improve the performance of each class as it is efficient for CNS tumors and Leukemia datasets. Table 3 summarizes the studies discussed in this section.

4.2.1. Inferences from second direction (D2)

As can be noticed, little attention was paid to research done on this direction in the literature. All researches in this direction use the SVM classifier. In order to evaluate the performance, accuracy, and a number of features, (NF) metrics were commonly employed.

Table 3
Embedded feature selection methods.

Ref	FS Algorithm	Classifier	Datasets	Datasets dimensions range	performance metrics	Results
[52]	rLGenSVM	GenSVM	8 Microarray (Ma) datasets	5748–15154	Accuracy and NF	rLGenSVM provides accuracy that reaches 100% in 6 datasets and an accuracy that equals 96% with 26 features for Brain datasets and 81.38% for TOX-171.
[53]	KPD-SVM	SVM	2 Ma datasets	208–569	Accuracy and NF	KPD-SVM provides an accuracy that equals 95% using 5 features for Wisconsin Breast Cancer (WBC) dataset and around 88% of accuracy for Sonar dataset with around 15 genes.
[54]	KP-SVDD and KP-CSSVM	SVM	12 Ma datasets	2308–17404	Accuracy and NF	KP-CSSVM produces better results than KP-SVDD, where KP-CSSVM provides the best accuracy at 98.4% with 1000 features versus 87.4% using KP-SVDD for Gordon's lung cancer dataset. In addition, KP-CSSVM provides accuracy that reaches 100% with 50 genes versus 93.1% with 500 genes using KP-SVDD for BHAT1 dataset.
[55]	SVM-BT-RFE	SVM	Lymphoma, Colon, Leukemia, Medulloblastoma, Prostate	5893–43237	Accuracy and error rate	SVM-BT-RFE method outperformed SVM-T-RFE with 25% of improvement and SVM-RFE with more than 40%. SVM-BT-RFE provides the best accuracy at 97% with 700 genes versus 9% error rate with 20 genes for a Colon dataset that has 43 237 genes. It also provides an accuracy that equals 100% with nine genes for the dataset with 5893 genes.
[56]	MPMCSVM-RFE	SVM	CNS tumors, Leukemia, Lung cancer	989–1000	Accuracy and NF	The proposed method provides an average accuracy that equals 95.42% with 235 genes, 98.33% with 187 genes, and 98.41% with 192 genes for CNS, Leukemia, and Lung cancer datasets, respectively.

4.3. Third direction (D3)

This direction is wrapper FSM. PSO is one of the popular meta-heuristic algorithms used as a wrapper FSM for microarray data processing. One of the recent studies that used PSO is a study done by Jain et al. [57], where authors developed an improved version of a Binary Particle Swarm Optimization namely (iBPSO) to enhance classification accuracy. They used six microarray datasets called ALL-AML, MLL, CNS, small-round-blue-cell tumor (SRBCT), Breast, and Lymphoma. To apply classification, they used three classifiers which are SVM, Naive Bayes (NB), and KNN with 5-fold cross-validation for evaluating the proposed method. Results show that the proposed system provides better classification accuracy by escaping from a local minima stagnation.

Another PSO wrapper-based FSM was developed by Moradi et al. [58] through hybridizing PSO and local search strategy, which is named HPSO-LS. They used LS algorithm for guiding search in PSO algorithm in gene selection process based on their correlations. They used 12 datasets, including 4 microarray datasets which are called Wisconsin Breast Cancer (WBC), Colon Cancer, Lymphoma, and Leukemia datasets. They employed KNN as a classifier to evaluate the proposed method. They compared HPSO-LS with information gain, term variance, fisher score, mRMR filter-based FS methods, genetic algorithm, particle swarm optimization, simulated annealing, and ant colony optimization wrapper-based FS methods. Results show that the proposed method outperformed all methods.

Garibay et al. [59] proposed an Inertial Geometric Particle Swarm Optimization algorithm with some modifications as a wrapper-based FS method for gene selection in DNA microarray data. They used SVM, Self-Organizing Map, back propagation neural network, NB, Decision Tree (DT), Artificial Immune Recognition System, and PSO along with C4.5 DT as classifiers and two datasets which are Colon and Prostate, for evaluating the proposed method performance. It can be noticed from the results that the proposed method enhanced accuracy by a 4% and the number of the selected genes is also competitive, compared to other state-of-the-art methods.

Modified Particle Swarm Optimization (MPSO) is based on wrapper FS method that was developed by Mohapatra and Chakravarty [48]. They used SVM, KNN, and NB as classifiers, along with three biomedical microarray datasets which are called Prostate Cancer, Leukemia, and Colon Tumor for evaluating the proposed method. Results show that the proposed method provides competitive results based on accuracy, precision, recall, F-score, and Area Under Curve (AUC).

Authors in Ref. [60] used PSO combined with DT for gene selection in ten cancer datasets (11_Tumors, 14_Tumors, 9_Tumors, Brain Tumor_1, Brain Tumor_2, Leukemia2, Lung Cancer, SRBCT, Prostate Tumor, and DLBCL). They deployed PSO for generating a subset of genes, then applied C4.5 classifier for evaluating the generated subset accuracy. This process was repeated five times. In each time one subset represented a test set and the remaining four subsets were used as a train set. Average accuracy of all five runs was used as a fitness function of PSO. They compared the proposed method performance with other four popular classification methods (SVM, back propagation neural network (BPNN), DT (C4.5), and self-organizing map (SOM)). The proposed method outperformed all other methods and provided above 90% accuracy in six datasets.

Other studies used Firefly algorithm, such as the study done by Almugren, and Alshamlan [61]. In, they proposed a wrapper FS method using a Firefly algorithm called FF-SVM. They used Five benchmark microarray datasets and SVM as a classifier to evaluate the proposed method by comparing it to other evolutionary wrapper-based and hybrid-based state-of-the-art algorithms for FS. Results showed that FF-SVM outperforms all other wrapper-based FS methods in four out of five datasets, as it achieves more than 90% of accuracy with a small number of genes in all five datasets.

Furthermore, Jinthanasatian et al. [62] deployed a Neuro-fuzzy FS with Firefly algorithm for tuning the algorithm parameters. In this work, classification and FS were conducted concurrently. They used seven microarray datasets called Lung Cancer, Ovarian cancer, Prostate cancer, Leukemia (ALL/AML), Breast cancer, Colon cancer, and DLBCL from the Kent Ridge Bio-medical repository for evaluating the proposed method. Neuro-fuzzy generated rule sets as classifiers. Results showed that the proposed FS method algorithm provided comparable results as the accuracy is not as high in some cases.

Ragunthar and Selvakumar [63] came up with a wrapper-based FS method that uses hybridize artificial bee colony (ABC) with stochastic diffusion search (SDS) algorithms, which are called ABC-SDS. They used Gene Expression Omnibus (GEO) Data sets (GDS) and their database with SVM and two datasets called GDS531 with 173 samples and about 12 625 features, and GDS2643 with 56 samples and 22 283 features. Results revealed that using ABC-SDS together provide better results compared to using SDS alone or ABC alone.

In [64], authors used Genetic Bee Colony (GBC) algorithm for FS process. They used a classification accuracy as an evaluation criterion for the proposed method for classifying performance. They used three microarray datasets, which are Colon Tumor, Leukemia ALL-AML, and Lung Cancer, for analyzing the proposed method of performance. For classification, they used a Conjugate Gradient Back-propagation with Modified Polak Ribiere (MBP-CGP) as a classifier. Results show that the proposed system provides accuracy results ranged between 88.75 and 100% using up to 47–51% of genes for all datasets.

Recently, Tawhid, and Ibrahim [65] developed a wrapper FSM based on a binary whale optimization algorithm (BWOA) with three classifiers. They evaluated the proposed method using 32 datasets from the UCI repository. One of the microarray datasets is called Breast cancer, where results show that the proposed method gives an efficient classification performance.

Zakeri et al. [66] deployed a wrapper-based FS method which is named GOFS, using a modified grasshopper optimization algorithm (GOA). GOA was modified in order to balance its exploration and exploitation capability by using a mathematical model of repulsion and attraction forces between grasshoppers. For evaluating GOFS performance and twelve of well-known FS methods (binary genetic algorithms (BGA), ACO, simulated annealing (SA), PSO, and differential evolution feature selection (DEFS)), they used ten Ma datasets including three high dimensional datasets (Colon, Leukemia, and SRBCT). Results show that GOFS provides an accuracy of 100% with 4 genes out of 2308 genes for Lymphography dataset.

Chatra et al. [67] developed a wrapper FSM based on a binary bat algorithm using a novel fitness function that minimizes the intra-class distances and maximizes the inter-class distances. They used an extreme learning machine (ELM) as a classifier along with eight microarray datasets to evaluate the effectiveness of the proposed method. They compared their methodology with the proposed fitness function with its fitness function that is usually used in the literature. It was noted from the results that the proposed fitness function outperformed the original one based on classification, accuracy, precision, recall, specificity, and F score metrics.

Recursive Memetic Algorithm (RMA) used a wrapper-based FSM by Ghosh et al. [68]. RMA applied over seven Ma datasets (SRBCT, Colon, Prostate, MLL, DLBCL, Leukemia, and AMLGSE2191) using three well known classifiers (SVM, Multi-Layer-perceptron (MLP), and KNN). Results show that RMA outperformed both Genetic Algorithm (GA) and basic Memetic Algorithm (MA). It provides accuracy of 100% in all cases while requiring a very small number of genes (ranges from 2 to 12).

A new wrapper of FS method that is based on binary learning-based optimization algorithm was presented by Allam and Nandhini [69] and called binary teaching learning based optimization (FS-BTLBO). They used different classifiers for computing the fitness of the proposed system's obtained results, such as NB, DT, and SVM. They used the WDBC dataset for evaluating the proposed method. Results show that FS-BTLBO method reduces the number of genes to 16 with the error rate of 0.018 using NB classifier and 16 genes using the Discriminant

Table 4
Wrapper-based feature selection methods from 2019 to 2021.

Ref	FS Algorithm	Classifier	Datasets	Datasets dimensions range	performance metrics	Results
[65]	BWOA	LR, C4.5, NB	One Ma dataset (Breast cancer) from the UCI repository	9	Accuracy and NF	The proposed algorithm provides the best mean accuracy with a percentage of 97%, 98%, and 97% for Breast cancer dataset using Logistic Regression (LR), and C4.5 and NB in 50–50 training-validation test. However, in the ten-fold cross-validation test, BWOA provides mean accuracy that equals 99%, 98%, and 90% for LR, C4.5, and NB with 4 features, respectively.
[71]	Monarch butterfly optimization algorithm (MBO)	KNN	8 Ma datasets including Breast cancer, Breast EW	15–10368	accuracy, recall, precision, specificity, F-score, and NF	MBO outperforms four of meta-heuristic algorithms in the literature called WOASAT, ALO, GA, and PSO as they select less than 40% of features from all datasets with accuracy, recall, precision, specificity, and F-score that reach 100% in 5 out of 8 datasets.
[67]	Binary bat algorithm	ELM with a new fitness function	8 Ma datasets	50–10368	Accuracy, precision, recall, specificity, F score, and NF.	The proposed FS method provides accuracy at a percentage of 100% in 6 out of 8 datasets with an average of feature selection percentage of 25.80%.
[66]	GOFS	SVM and KNN	Ten Ma datasets including three high dimensional datasets (Colon, Leukemia, and SRBCT)	9–7129	Accuracy and NF.	GOFS provides an accuracy of 100% with 4 genes out of 2308 genes for Lymphography dataset.
[68]	RMA	Three well known classifiers (SVM, Multi-Layer-perceptron (MLP), and KNN).	Seven Ma datasets (SRBCT, Colon, Prostate, MLL, DLBCL, Leukemia, and AMLGSE2191)	2308–12616	Accuracy and NF.	Results show that RMA provides an accuracy of 100% in all cases while requiring a very small number of genes (ranges from 2 to 12).
[61]	Firefly algorithm	SVM	SRBCT, Lung, Colon, Leukemia1, and Leukemia2	2000–7129	Accuracy and NF	FF-SVM provide high classification accuracy that reaches the percentage of 100% over Leukemia1 with 3 genes and Lung dataset with two genes. It provides an accuracy that equals 95.2% and 83.3% over SRBCT and Leukemia2, using five genes, respectively. Ten genes were used in a Colon dataset to provide the best accuracy of 95.2% using the proposed method.
[63]	ABC-SDS	SVM	GDS531, and GDS2643	12, 625–22, 283	Accuracy, sensitivity, specificity, and F score	For GDS531 dataset, ABC–SDS satisfies better results than ABC alone by 9.53%, 9.01%, 9.01%, and 9.48% for accuracy, sensitivity, specificity, and F score, respectively. For GDS2643, ABC–SDS achieves better results than ABC alone by y 10.75%, 10.04, 10.04, and 11.88% for accuracy, sensitivity, specificity, and F score, respectively.

Analysis as a cost function with an error rate of 0.016 4. The proposed method selected the fewest number of genes (equal to 6 genes) with decision trees, compared to other learning models.

Sharma et al. [70] came up with a wrapper-based FS method based on an improved regularized linear discriminant analysis for human cancer classification problems. For evaluating the proposed method, authors used J4.8, NB, KNN with K = 1, and SVM for classification, and 3 microarray datasets which are called MLL Leukemia, SRBCT, and Acute Leukemia. It can be noted from the results that the proposed method achieves encouraging classification accuracy using a small number of genes. There are many other works conducted using this approach, such as [71–73], as summarized in Tables 4–7.

4.3.1. Inferences from third direction (D3)

This direction has a medium attention from researchers. Some of these studies employed two Swarm Intelligence algorithms as a search algorithm in wrapper for combining their advantages together [63], while other studies had only used one intelligent swarms algorithm [71, 72]. Conversely, some studies enhanced the fitness function [67], while others used specific local search algorithms [58] for guiding search in the algorithm employed as wrappers. PSO is one of the popular swarm intelligence algorithms used as a wrapper FSM for microarray data processing. For the process of classification, SVM, NB, and KNN are considered the most popular classifiers used by researchers in this direction with accuracy and NF as performance metrics. Most of the studies use datasets with dimensions above 2000 genes.

4.4. Fourth direction (D4)

It concerns Filter FSM. Some studies use a score-based criteria fusion (SCF) as in the study done by Kavitha et al. [74], where authors developed a Score-based criteria fusion (SCF) FS method that represents a combination of SU (Symmetric uncertainty) and Relief using a score normalization for selecting the relevant, non-redundant features, since these two methods use different strategies in FS process. In SU, they observed the effect in the running time while changing a parameter named Delta. However, in Relief, they observed the effect in the running time based on a various number of neighbors and features to be selected. They used SVM with a Leukemia microarray dataset for evaluation. Based on the results, they concluded that using a single method for feature selection is less effective than using two different feature selection methods with entirely different basic criteria. Therefore, their combination contains the features of both methods, and this leads to achieve better results.

Ke et al. [75] came up with a Score-based criteria fusion (SCF) for FS that applied two ranking-based filters for choosing features based on an

Table 5
Wrapper-based feature selection methods in 2018.

Ref	FS Algorithm	Classifier	Datasets	Datasets dimensions range	performance metrics	Results
[64]	GBC	MBP-CGP architecture is used with one and two hidden layers	Colon Tumor, Leukemia ALL-AML, Lung Cancer	2000–12533	Accuracy and NF	The system is able to select features of up to 47–51% for all datasets and accuracy ranged between 88.75 and 100% for all datasets.
[57]	iBPSO	SVM, NB, KNN (K = 1)	ALL-AML, MLL, CNS, SRBCT, Breast and Lymphoma	2308–12582	Accuracy and NF	The proposed method with NB provides the least average number of genes that equals 120.83. In contrast with KNN (K = 1) and SVM, it provides an average number of genes that equals 133.66 and 138.83, respectively. The average highest classification accuracy using the proposed method with NB, KNN (K = 1), and SVM are 93.98, 90.58, and 93.02, respectively.
[72]	Cuttlefish Algorithm	KNN, DT, Hidden Markov models (HMM), and SVM	Prostate Cancer, Diffuse Large B-Cell Lymphoma (DLBCL), Leukemia, Lung cancer-Michigan, Lung cancer, Ontario, CNS, Breast cancer, Colon tumor	2000–24481	Accuracy, NF and time in seconds.	The proposed method decreases the number of features up to 90%. It achieves an accuracy exceeded 92% in 7 out of 8 datasets with lower computation time compared to the other methods in the literature. For Prostate dataset, the accuracy reaches a percentage of 100%.
[69]	Binary teaching learning based optimization algorithm	NB, KNN, DT, SVM and Discriminant Analysis	Breast Cancer Wisconsin dataset	30	Precision, Recall, Sensitivity, F-Measure, Accuracy, Receiver Operating Characteristic (ROC) curve, and NF	The proposed method provides a percentage above 98% of accuracy and about 97% F-measure using NB, SVM, and Discriminant Analysis classifiers.

Table 6
Wrapper-based feature selection methods from 2016 to 2017.

Ref	FS Algorithm	Classifier	Datasets	Datasets dimensions range	performance metrics	Results
[62]	Nero-fuzzy with firefly algorithm for tuning parameters	rule-set generation as classifiers	Seven microarray datasets which are Lung Cancer, Ovarian cancer, Prostate cancer, Leukemia (ALL/AML), Breast cancer, Colon cancer, and DLBCL.	6500–24481	Accuracy and NF	The proposed method provided an accuracy that equals 93.42% with 4 genes, and 96.13% with 12 genes for Lung cancer and Ovarian cancer datasets.
[73]	Artificial fish swarm (AFSO)	SVM	9 datasets that include 1 Ma dataset called SRBCT	2307	Accuracy and NF.	The proposed approach provides an accuracy that equals 82% using 215 features for SRBCT dataset.
[58]	HPSO-LS, which is a particle swarm optimization with a local search strategy	KNN	4 Ma datasets taken from the UCI repository which are Wisconsin Breast Cancer (WBC), Colon Cancer, Lymphoma, and Leukemia	10, 7129	Accuracy and running time in seconds.	HPSO-LS provides an accuracy that equals 98.08% with four genes for the WBC dataset.

estimation of relevance between features and classes. They used SVM and KNN with five microarray datasets called SRBCT, Leukemia, Colon, Carcinomas, and Prostate for evaluating the proposed method. The proposed method presents its efficiency in selecting more informative features compared with other methods.

Rouhi and Nezamabadi-pour [76] used an Improved Binary Gravitational Search Algorithm (IBSGA). They evaluated it using KNN classifier and five microarray datasets called Leukemia, DLBCL, SRBCT, Prostate, and Lung. Results showed that the proposed method provides better accuracy and precision in 4 out of 5 datasets, compared to other common filter methods like Locality Sensitive Laplacian Score (LSLS) in terms of classification accuracy, precision, and recall. LSLS was developed by Liao et al. [77] in 2014. It is a variant of Laplacian score (LS), and it incorporates label information into the graph Laplacian matrix. LSLS works by minimizing local within-class information and maximizing local between-class information, simultaneously. It has two parameters, mainly: weight factor (set to be a cosine similarity) and k-nearest neighbor (set 5 in all experiments). They used six Ma datasets for evaluating the proposed method (Acute leukemia, Lung Cancer, DLBCL, Prostate, MLL Leukemia, and SRBCT) with SVM for evaluating the proposed method by comparing its performance with three filter-based FSM (Kruskal-Wallis (KW), similarity preserving feature selection (SPFS), and ReliefF). Results show that LSLS outperformed all other methods.

Tang and Zhou [78] developed a new MI filter-based FS method by specifying features based on their relationships with class labeling beyond the selected features. They used three classifiers, which are NB, Iterative Dichotomiser 3 (ID3), and Logistic classifier with five microarray datasets, which are Leukemia, DLBCL, Colon, Prostate, and Lung cancer.

There are some studies that developed filter FSM for unsupervised learning as in the study done by Tabakhi et al. [42], where they developed an unsupervised filter-based FSM, called Microarray Gene Selection based on Ant Colony Optimization (MGSACO). The proposed method uses a novel fitness function without using any learning models for evaluating the performance of the found subset of genes in each

Table 7
Wrapper-based feature selection methods from 2014 to 2015.

Ref	FS Algorithm	Classifier	Datasets	Datasets dimensions range	performance metrics	Results
[59]	A modified version of the Inertial Geometric Particle Swarm Optimization (MIGPSO)	SVM, Self-Organizing Map, back propagation neural network, NB, DT, Artificial Immune Recognition System, and PSO with C4.5 Decision Tree	Colon, Prostate	2000–10509	Accuracy, NF, time.	The proposed method improves the classification accuracy of about 4% compared to other methods. It provides an accuracy that equals 95.8% with 6 genes for Colon dataset and 98.04% with 84 genes for Prostate dataset.
[48]	Modified PSO (MPSO)	SVM, KNN and NB	Prostate Cancer, Leukemia, and Colon Tumor	2000–12600	Precision, Recall, F-Score, AUC, Accuracy, and NF.	MPSO provides an accuracy that reaches a percentage above 99% and F-score that equals 90% using SVM classifier over Prostate cancer dataset with 20 genes. For the rest of datasets, MPSO provides an accuracy of 99.01% with 10 genes using SVM classifier.
[70]	An improved regularized linear discriminant (Improved RLDA)	Four classifiers called J4.8, NB, KNN (K = 1), and SVM	MLL Leukemia, SRBCT, and Acute Leukemia	2308–12582	Accuracy and NF.	Improved RLDA provides the best results using SVM classifier as accuracy that reaches 100% for all of the three datasets using 10% of genes.
[60]	PSO	DT (C4.5)	11_Tumors, 14_Tumors, 9_Tumors, Brain Tumor_1, Brain Tumor_2, Leukemia2, Lung Cancer, SRBCT, Prostate Tumor, and DLBCL	83–12601	Accuracy and NF.	The proposed method outperforms all other methods and provides a percentage above 90% of accuracy in 6 datasets.

Table 8
Filter-based feature selection methods from 2017 to 2021.

Ref	FS Algorithm	Classifier	Datasets	Datasets dimensions range	performance metrics	Results
[74]	Score-based criteria fusion FS method	SVM	Leukemia	7128	Run time	The proposed method improves accuracy and decreases the computational time.
[76]	IBSGA	KNN	Leukemia, DLBCL, SRBCT, Prostate and Lung	2308–12600	Accuracy, precision, and recall.	The proposed method provides better accuracy and precision in 4 from 5 datasets compared to other common filter methods in literature.
[75]	SCF	SVM, KNN	SRBCT, Leukemia, Colon, Carcinomas, Prostate	2000–10509	Error rate and NF.	SCF achieves Zero error rate with only 20 genes using SVM classifier.

Table 9
Filter-based feature selection methods from 2014 to 2016.

Ref	FS Algorithm	Classifier	Datasets	Datasets dimensions range	performance metrics	Results
[79]	Dynamic MBPSO (D-MBPSO)	SVM, NB, DT	Leukemia, Lung, Colon, Prostrate, SRBCT	NA	Accuracy and NF.	D-MBPSO outperformed 6 of well-known unsupervised filter-based gene selection methods with an improvement percentage that equals 5–6%, using each of the three classifiers. It provides an average accuracy that equals 72.71% using SVM classifier and 75.67% using NB classifier.
[78]	New MI filter based FS method called Reinforced Mutual Information based Feature Selection (RMIFS).	NB, ID3 (DT), Logistic classifier	Leukemia, DLBCL, Colon, Prostate, and Lung.	2000–12600	Accuracy and NF.	RMIFS provides a better accuracy in 4 out of 5 datasets compared to 4 of the existing FS methods. It provides an accuracy that reaches 100% with 4 genes for Leukemia dataset.
[42]	MGSACO	SVM, NB, DT	Colon, Leukemia, SRBCT, Prostate Tumor, and Lung Cancer.	2000–12600	classification error rate.	MGSACO provides the least error rate that equals 20.14% with 20 genes using NB classifier. It decreases error rate by 12.73%, 18.24%, 9.74%, and 15.44% for UFSACO, MC, RRFS, and TV respectively, using NB classifier.
[77]	LSLS	SVM	Acute leukemia, Lung Cancer, DLBCL, Prostate, MLL Leukemia, SRBCT.	2308–12600	Accuracy, NF, Precision, Recall, F-score, and AUC.	LSLS gives an accuracy of 100% on Lung, MLL, and SRBCT datasets.

NA means not available.

iteration from ACO, and it returns the best subset of genes at the end of all iterations. They used the relevance and redundancy analyses in the proposed fitness function computation, where relevancy is represented by variance and subsets with maximum relevance, so it should have a greater fitness value. For evaluating MGSACO's performance, they compared its classification error rate to seven of the well-known unsupervised and supervised FS methods. Two of the univariate filter methods which are Term variance (TV) and Laplacian score (LS) were used along with three multiivariate filter methods including Relevance-redundancy feature selection (RRFS), random subspace method (RSM), and mutual correlation (MC). UFSACO minimal-redundancy-maximal-relevance method (mRMR) uses five Ma datasets with three classifiers (SVM, NB, and DT). Results showed that MGSACO is significantly superior to the already existing methods over different classifiers and datasets, as it provides the least error rate that equals 20.14% with 20 genes using NB classifier.

In addition, Umamaheswari et al. [79] proposed an unsupervised filter which is called dynamic modified binary PSO (MBPSO) by integrating MBPSO into filter approach through defining new fitness functions. They conducted their proposed approach over five microarray datasets called Leukemia, Lung, Colon, Prostrate, and SRBCT using three classifiers (SVM, NB, and DT). Results show that the proposed method outperforms six well-known unsupervised filter-based FS methods, such as term variance (TV), unsupervised feature selection based on ACO (UFSACO), random subspace method (RSM), relevance-redundancy feature selection (RRFS), Laplacian score (LS), and mutual correlation (MC).

4.4.1. Inferences from fourth direction (D4)

Recently, most studies used the existing filters as a pre-processing stage for decreasing the number of genes that must be passed to the wrapper stage in the hybrid FS method direction. It is a competitive direction for researchers as it provides more accurate results than the filter-based FS method, and at the same time, minimizes computation time compared to the wrapper-based FS methods. Therefore, there are little research done on D4. It is noticeable that most of the filter-based FSM used in microarray datasets are supervised FS methods.

Table 8 and Table 9 summarize these works.

4.5. Fifth direction (D5)

This direction proposes hybrid FSM. GA was used as a part of many studies done on hybrid FS methods in the literature. An example of the most recent studies in this direction is the study conducted by Al-Obeidat et al. [80]. It used GA as a wrapper combined with three filters, called principal component analysis, correlation, and spectral-based FS method for processing six microarray datasets. They employed KNN, RF, SVM for classification, and results show that the

Table 10
Genatic algorithm-based hybrid feature selection methods from 2019 to 2021.

Ref	First-part	Second-part	Classifier	Datasets	Datasets dimensions range	performance metrics	Results
[80]	Three filters are called principal component analysis, correlation, and spectral-based FS method	GA as a wrapper	KNN, RF, SVM	6 Ma datasets	2000–12600	Accuracy, Recall, False positive rate, Precision, F-measure, and Entropy.	The proposed method provides an average accuracy that equals 92%. It provides an average F-measure and average precision at a percentage above 80%.
[81]	Three filters: Symmetrical uncertainty, Chi-square, and ReliefF	GA as a wrapper	multi-layer Perceptron, KNN, SVM	5 Ma datasets	2309–12601	Accuracy and NF.	The proposed S model (EU) gives 100% of accuracy over 3 out of 5 datasets with less than 5 genes.
[82]	Mutual information maximization and reliefF as an ensemble FS method	GA as a wrapper	CS-D-ELM, SVM, RoF	Breast, Lung, Colon, Leukemia	2000–24482	Accuracy and NF.	The proposed method provides classification accuracy above 98% over 3 datasets (Lung, Colon, and Leukemia).
[83]	T-test and MIC as a filter	GA embedded in the recursive feature elimination (RFE) process as a wrapper	KNN, SVM	19 benchmark Ma datasets	2000–54675	Accuracy, Sensitivity, Specificity, and AUC.	MGRFE gets an accuracy that is near to 100% using 5 genes for 10 datasets, and greater than 90% using 10 genes for all datasets.
[89]	Two filter stages	an adaptive GA as a wrapper	SVM and KNN	Leukemia, Prostate Cancer, DLBCL	NA.	Accuracy, sensitivity, specificity, precision, recall, f_measure, gmean, and NF.	The proposed rMRMR-HBA method gives an accuracy that reaches 100% over DLBCL and Leukemia datasets, using SVM classifier.
[88]	mRMR as a filter	GA as a wrapper	5 classifiers	Risinger Endometrial Cancer, Nutt-Brain Cancer, Pomeroy-Central Nervous System Embryonal Cancer	88–1771	Accuracy	For Risinger Endometrial Cancer dataset. The proposed method gives the highest accuracy of 97.05% using NB classifier, while LDA provides an accuracy that equals 97.9% over Pomeroy-Central Nervous System Embryonal Cancer dataset. SVM provides an accuracy of 95% for Nutt-Brain Cancer dataset.
[90]	ReliefF as a filter	an entropy based genetic search as a wrapper	KNN, SVM, NB.	Wisconsin breast cancer	24 482	Accuracy, Sensitivity, Specificity, Precision, Recall, F-Measure, NF, and running time in seconds.	The proposed approach provides the highest classification accuracy (82.05%) with 75 genes and a running time that equals 0.09 s using the random forest classifier.

NA means not available.

proposed method outperformed four related methods in the literature in terms of accuracy, recall, false positive rate, precision, F-measure, and entropy.

In 2019, Ghosh et al. [81] proposed a hybrid FS method for gene identification from microarray data based on an ensemble filter and GA as a wrapper. In the first stage, an ensemble filter that consists of ReliefF, Chi square, and symmetrical uncertainty was applied, where filters were used to rank the features. Then, the union or intersection was used for combining the previous filters to rank lists. In the next stage, GA was applied as a wrapper to select the optimal set of features from the ensemble. They used five benchmark microarray datasets and three classifiers of multi-layers of Perceptron, KNN, and SVM for evaluating the proposed model. Results illustrate that the proposed method provides better performance compared to some methods in the literature.

In the same year, authors in Ref. [82] developed a hybrid FS method by combining GA as a wrapper with information maximization and reliefF as an ensemble filter. They used four microarray datasets, which are Breast, Lung, Colon, and Leukemia for evaluating the proposed method. For classification, they used three classifiers called CS-D-ELM, SVM, and ROF. The proposed method is considered suitable to handle various cancer diagnostic datasets, and it provides the highest classification accuracy compared to some state-of-the-art FS methods for microarray datasets.

Furthermore, Peng et al. [83] came up with a hybrid FS method, which is multi-layer recursive feature elimination method, based on an embedded integer-coded genetic algorithm (MGRFE). It combines two filter methods, which are *t*-test and maximal information coefficient (MIC) and GA embedded in the RFE process as a wrapper. They used SVM and KNN classifiers over nineteen benchmark microarray datasets, including binary, multi-class, balanced, and imbalanced datasets, that include diffuse large B-cell lymphoma, Prostate, acute lymphoblastic leukemia, embryonal tumor, and CNS. MGRFE was compared to other evolutionary-computation-based feature selection algorithms that showed higher convergence speed with a slightly small number of genes.

Another study was done by Lu et al. [84], where they developed a hybrid FS method called MIMAGA based on a Mutual Information Maximization (MIM) filter and the Adaptive Genetic Algorithm (AGA) as a wrapper. They evaluated the proposed method over six cancer datasets, which are Leukemia, Colon, Prostate, Lung, Breast, and SRBCT, using an extreme learning machine (ELM) classifier. Results show that MIMAGA outperformed three existing FS algorithms.

Dashtban and Balafar [85] proposed FS method, which is called Intelligent Dynamic Genetic Algorithm (IDGA), based on combining a genetic algorithm with an artificial inelegance. They applied it through five cancer datasets, which are SRBCT, Breast, DLBCL, Leukemia, and Prostate, using three classifiers: KNN, NB, and SVM. IDGA with Fisher provides better results compared to IDGA with Laplacian on four datasets.

Salem et al. [86] proposed a new hybrid FS method named IG/SGA, which combines Standard Genetic Algorithm (SGA) as a wrapper with an Information Gain (IG) as a filter. They used seven microarray datasets called Leukemia, Colon tumor, Central nervous system, Lung Cancer-Ontario, Lung Cancer-Michigan, DLBCL, and Prostate cancer. IG/SGA provides 100% of accuracy over two datasets from seven cancer microarray datasets using genetic programming (GP) as a classifier.

Djellali et al. [87] developed two-hybrid methods. The first method is called FCBF-GA, and it combines a Fast Correlation based Filter with a Genetic Algorithm as a wrapper. In contrast, the second method is called FCBF-POS, and it hybridizes the same filter with Particle Swarm Optimization as a wrapper. They used four microarray datasets with Support Vector Machine as a classifier to compare the performance of the proposed methods. Results show that FCBF-POS outperformed the first hybrid method.

Thangavelu et al. [88] proposed a hybrid FS method that combines Minimum Redundancy Maximum Relevancy mRMR as a filter and GA as a wrapper. They used SVM, NB, Linear Discriminant Analysis, decision trees, and RF classifiers with 3 Ma datasets collected from bio-informatics laboratories of Rutgers University, which are Risinger Endometrial Cancer, Brain Cancer, and Pomeroy Central Nervous System Embryonal Cancer, and this was done to evaluate the proposed method.

Begum et al. [89] proposed a hybrid FS method, which is named rMRMR-HBA based on two filter stages. In the first stage, they applied various filters, namely: T-test, Bhattacharyya, and ReliefF. In the second stage, they applied Mutual Information Maximization (MIM) filter and an adaptive GA as a wrapper. They used SVM and KNN classifiers with Leukemia, Prostate Cancer, and DLBCL datasets to evaluate rMRMR-HBA performance, where it provides good results, as shown in the results below.

There are other GA-based FS methods in the literature, such as [90–99]. All these works are summarized in tables from Tables 10–13.

Other studies deployed RFE as the wrapper part of the hybrid FS

Table 11
Genatic algorithm-based hybrid feature selection methods in 2018.

Ref	First-part	Second-part	Classifier	Datasets	Datasets dimension's range	performance metrics	Results
[91]	GA as a wrapper	an embedded hybrid L1/2 + 2 regularization approaches for tuning the population of GA	Logistic regression	AML, Lymphoma, Prostate, Lung, and DLBCL.	77–240	Sensitivity, Specificity, NF, and Accuracy.	HGAWE provides an accuracy of more than 93% in all datasets.
[92]	Conditional Mutual Information Maximization filter	an adaptive genetic algorithm as a wrapper	ELM, SVM, KNN	Breast Cancer, Colon Cancer, Diffuse Large B-cell, Lymphoma, Leukemia, Small-Blue-Round-Cell Tumor, SBRCT, and Lung Cancer.	2000–24481	Accuracy, Sensitivity, Precision, F-measure.	The proposed approach achieves better results using ELM as it provides an accuracy of more than 96% in DLBCL and SPRCT datasets.
[93]	EGS method using multi-layer approach and F-score filter	an adaptive genetic algorithm as a wrapper	SVM, NB	Breast, Colon, DLBCL, Leukemia, SBRCT, and Lung cancer	2000–24481	Accuracy, Sensitivity, Precision, F-measure.	The proposed method provides better classification accuracy using SVM classifier, as it gives an accuracy at a percentage of 90.07% using SVM over SRBCT dataset and more than 95% for DLBCL and Lung datasets.
[94]	An ensemble filter of MI maximization and relief with voting strategy for combining results	an extended GA that extended a self-adjusting crossover probability and mutation probability as a wrapper	ELM	Leukemia, Breast, and Lung cancer	7130–24482	Accuracy.	The proposed method outperforms three of state-of-the-art FS methods. It provides an accuracy greater than 96%.

Table 12
Genatic algorithm-based hybrid feature selection methods in 2017.

Ref	First-part	Second-part	Classifier	Datasets	Datasets dimensions range	performance metrics	Results
[84]	Mutual information maximization	an adaptive genetic algorithm AGA	BP, SVM, ELM, RELM	Leukemia, Prostate, Lung, Breast, SRBCT	2000–24482	Accuracy and NF.	The lowest classification accuracy that is provided by the proposed method is around 80%, which is in the acceptable range.
[85]	Fisher score and l Laplacian-Score filters	an intelligent dynamic GA as a wrapper	KNN, NB, and SVM.	DLBCL, SRBCT, Leukemia, Prostate, and Breast cancer	2308–12600	Average running time, Error rate, and NF.	The proposed Intelligent dynamic genetic algorithm (IDGA) obtained its best results in combination with Fisher-score and KNN. It gives 100% of accuracy in four datasets.
[86]	Information Gain	GA	MLP, J48, NB	Leukemia, Colon tumor, Central nervous system, Lung cancer Ontario, Lung cancer Michigan, DLBCL, and Prostate cancer	2000–12600	Accuracy, Sensitivity, Specificity and NF.	The proposed method provides a classification accuracy of 100% with IG threshold values equal to 0.4 and 0.7 for Lung cancer-Michigan and Prostate Cancer datasets, respectively.
[87]	Fast Correlation based Filter	GA and Particle Swarm as a wrapper	SVM	Wisconsin Breast cancer, hepatitis, arrhythmia, colon, DLBCL from the UCI repository	19–4026	Accuracy and NF.	FCBF-GA provides 100% of accuracy over DLBCL dataset with 58% of reduction in a number of genes. FCBF-GA provides better results based on accuracy and on the number of features than FCBF–PSO in WDBC and DLBCL datasets.
[105]	Three filters (Fisher criterion, *t*-test, and AUC)	GA as a wrapper	SVM	Breast, Colon, lung	NA.	Accuracy, Sensitivity, Specificity and NF.	Multistage feature selection (MSFS) method gives 100% of accuracy with 24 genes and 13 genes for Breast and Colon datasets, respectively.
[95]	a Spearman's Correlation Coefficient as a filter	GA as a wrapper	SVM, KNN, NB, DT	Colon, SRBCT, Lymphoma	2000–4026	Accuracy.	The proposed method provides the highest accuracy that equals 85.24% for Colon dataset using DT classifier. For Lymphoma dataset, it provides the maximum accuracy of 98.36% using KNN classifier.

NA means not available.

Table 13
Genatic algorithm-based hybrid feature selection methods from 2014 to 2016.

Ref	First-part	Second-part	Classifier	Datasets	Datasets dimensions range	performance metrics	Results
[96]	correlation coefficients filter	Multi objective Genetic Algorithm as a wrapper	KNN	Colon, Leukemia, Lymphoma	2000–7129	Accuracy and NF.	The proposed method provides good performance using simple KNN classifier compared to other methods, which used more sophisticated classifiers like SVM and ANN.
[97]	Information gain as a filter	Deep Genetic Algorithm as a wrapper	Genetic Programming	Leukemia, Colon tumor, CNS, Lung cancer-Ontario, Lung cancer-Michigan, DLBCL, Prostate	2000–12600	Accuracy.	IG-DGA provides an accuracy of 100% for Lung Cancer-Ontario and Prostate Cancer datasets.
[98]	Multiple Fusion Filter that consists of five traditional statistical methods	an embedded method that uses a Genetic Algorithm, with a Tabu Search	SVM	CNS, Leukemia, Colon, Lung, DLBCL	2000–12533	Accuracy and NF.	The proposed method provides around 100% of accuracy in all datasets that have been used in this study.
[99]	Filtering approach based on the KEGG database	GA with the addition of biological information as a wrapper	LDA, SVM	Leukemia, Prostate, Lung Cancer	7129–12600	Accuracy and Robustness.	The proposed method gives an accuracy that exceeded 97% in Leukemia dataset and Lung Cancer datasets. It improves the robustness in the three analyzed datasets.

method. An example is the study done by Gao and Liu [100], where RFE was used with SVM as a wrapper and combined with F-statistics as a filter for developing a hybrid FS method. They used SVM, PSOFOA LSSVM, and LSSVM classifiers with Prostate, Lung Cancer, Colon, and Lymphoma microarray datasets for evaluating the proposed method. For optimizing LSSVM classifier parameters, a hybrid method based on fruit fly optimization and particle swarm optimization was used. Results show that the proposed method achieved an accuracy of 100% with only 4 genes, which is better than all state-of-the-art methods that had been used for comparison.

Abinash and Vasudevan [101] used a Fisher's discriminant criterion method and MRMR (Minimum Redundancy Maximum Relevance) filter as a pre-processing method to decrease the dimension of the dataset. They then applied a backward feature elimination, support vector machine, and recursive feature elimination (SVM-RFE) as a wrapper-based FS method. They used the leukemia gene dataset taken from UCI repository to evaluate feature selection that was measured by kappa statistics with balanced accuracy. Results show that many features are

Table 14
Ant Colony Optimization-based hybrid feature selection methods.

Ref	First-part	Second-part	Classifier	Datasets	Datasets dimensions range	performance metrics	Results
[109]	Mutual information (MI) filter	Hybrid Stem Cell algorithm as a wrapper that utilizes Ant Colony optimization and Stem Cell algorithm	fuzzy classification system	Colon Cancer, Tumor, Insulin Sensitive – Resistant, Type 2 Diabetes, Leukemia, Prostate	2000–22283	Accuracy and interpretability.	Hybrid Stem Cell Algorithm (HSC) generates a near optimal set of genes and improves accuracy compared to other algorithms.
[110]	Fisher score as a filter	a cellular learning automata optimized with ant colony method as a wrapper	KNN, SVM, NB	Prostate tumor, ALL-AML-leukemia, MLL-leukemia, ALL-AML-4	7129–12600	Accuracy and NF.	Results denote that the proposed approach achieves 100% of accuracy with 1 or 2 highly informative genes.

irrelevant, but the relevant features give the high inaccurate results. SVM-based wrapper provides better results than in correlation.

In addition, Alzubi and Ramzan [102] developed a hybrid method that combines SVM-RFE as a wrapper with Conditional Mutual Information Maximization (CMIM) as a filter. For classification, they used SVM, NB, Linear Discriminant Analysis, and k-Nearest Neighbors. They applied the proposed method over five different single nucleotide polymorphisms (SNP) datasets. Results show that the proposed method outperformed four of FS methods from the literature, which are Minimum Redundancy Maximum Relevancy, Fast Correlation Based Feature Selection, CMIM, and ReliefF as they achieved up to 96% accuracy for the used dataset.

ABC was used as a part of hybrid FS methods in other studies as in Alshamlan's [103] study in which they used ABC with Correlation-based FS filter (Co), where results show that the proposed method outperformed Co-GA and Co–PSO. Additionally, Musheer et al. [104] developed a hybrid FS method that combines an independent component analysis filter with ABC as a wrapper. They used six microarray datasets.

Authors in Alkuhlani et al. [105] proposed a hybrid FS method based on GA as a wrapper and an ensemble filter that combines three common filters. They used SVM as a classifier and Support Vector Machine Recursive Feature Elimination as a fitness function, along with three different cancer datasets.

Some researchers used gravitational search algorithm (GSA), such as the study done by Shukla et al. [106], where they hybridized GSA with characteristics of teaching-learning-based algorithm (TLBO) as a wrapper, named TLBOGSA. They combined it with a mRMR filter for processing ten microarray datasets using NB as a classifier. The proposed method got above 98% classification accuracy in six datasets and 99.62% accuracy (best accuracy) in DLBCL dataset.

TLBO was hybridized with Simulated Annealing (SA) algorithm as a wrapper and then combined with Correlation-based filter to come up with a new hybrid FS method called TLBOSA and used by Shukla et al. [107]. Han et al. [108] hybridized ReliefF with a recursive binary GSA.

Other researchers used ant colony optimizations-based hybrid FS methods as in Vijay and GaneshKumar's [109] study, where they proposed a two-stage FS method that used MI as a filter then combined Ant Colony optimization and a novel Adaptive Stem Cell Optimization as a wrapper. Five Microarray datasets with a fuzzy classification system

Table 15
Particle swarm optimization-based hybrid feature selection methods.

Ref	First-part	Second-part	Classifier	Datasets	Datasets dimension's range	performance metrics	Results
[111]	Correlation-based filter	an improved Binary PSO as a wrapper	NB	Lung, Ovarian, CNS, ALL-AML, Colon, Breast, ALL-AML-3, ALL-AML-4, Lymphoma, MLL, SRBCT	2000–24481	Accuracy and NF.	The proposed method achieves 100% of accuracy for seven datasets with more than 92% for the remaining datasets.
[112]	Correlation coefficient filter	Particle Swarm Optimization as a wrapper	Random trees, decision stump, J48, random forest	MLL, SRBCT, Lymphoma	2308–12582	Accuracy and NF.	The proposed hybrid method selects 2–8% of the relevant and informative genes. It provides the highest accuracy in all datasets using ELM classifier with an accuracy of 93.7%, 96.8%, and 85.6% for SRBCT, lymphoma, and MLL, respectively.
[113]	ReliefF filter	PSO as a wrapper	SVM	Colon cancer, SRBCT, Leukemia, Lung cancer	2000–12600	Accuracy and NF.	RefFPSO selects the relevant genes to obtain high classification of accuracy that exceeds 80%.
[114]	A combination of three filters, called modified Bayesian logistic regression, T-test, and Fisher-score	PSO-dICA as a wrapper	SVM	DLBCL, SRBCT, Leukemia-ALL, Colon, Lung, Prostate	2000–12600	Accuracy and NF.	The proposed method provides an accuracy of 100% in all datasets.
[115]	F-statistic filter	Maximum Relevance Binary Particle Swarm Optimization with Class Dependent Multi-Category Classification system as a wrapper	SVM	9 Tumors, 11 Tumors, 14 Tumors, Leukemia1, Leukemia2, Brain Tumor1, Lung, SRBCT	2308–12600	Accuracy and NF.	Class dependent approaches improved classification accuracy compared to class independent FS methods.

were used to evaluate the proposed method.

Moreover, in the study done by Sharbaf et al. [110], Cellular Learning Automata and Ant Colony Optimization were combined. They used a fisher criterion method as a filter to reduce a number of genes and computation time. Then, Cellular Learning Automata with Ant Colony Optimization was applied as a wrapper to improve accuracy. The proposed method was evaluated using three classifiers (KNN, SVM, and NB) and four microarray datasets (ALL-AML leukemia, Prostate tumor, MLL-leukemia, and ALL-AML-4). Experimental results demonstrated that the proposed method presented the best accuracy rate using 1 or 2 genes where the selected genes are found meaningful in biology texts in the literature.

Particle Swarm Optimization (PSO) was used in some studies for developing hybrid FS methods. An example is the study done by Jain et al. [111], where they proposed a hybrid particle swarm optimization-based FS method for gene selection in microarray data by combining Correlation-based Feature Selection as a filter with Improved-Binary Particle Swarm Optimization as a wrapper. They evaluated the proposed FS method using Eleven microarray datasets and a Naive–Bayes classifier.

Chinnaswamy and Srinivasan [112] developed a hybrid FS method based on correlation coefficient as a filter and PSO as a wrapper for cancer classification on microarray gene expression data. They evaluated the proposed method using an ELM classifier and three multi-class cancer datasets. In addition, Liu et al. [113], Mollaee, and Moattar [114], Zhou, and Dickerson [115] came up with a PSO based hybrid FS methods for microarray data processing. A summary of all ant-colony-based methods is presented in Table14 while a summary of PSO based methods is presented in Table15.

Pashaei et al. [116] proposed a new hybrid FS method based on the Binary Black Hole Algorithm as a wrapper and Random Forest Ranking as a filter. They used seven classifiers KNN, NB, RF, alternating Decision Tree, AdaboostM1, Basis Function Network, and Bagging with 10-fold cross-validation over four microarray datasets, to evaluate the proposed method. The proposed method provided an accuracy that equals 95.71% for Leukemia_4c microarray dataset using 3 genes and 100% for MLL with 5 genes.

Dabba et al. [117] proposed a hybrid FS method based on Minimum redundancy maximum relevance as a filter and a wrapper. It hybridizes a quantum computing and the Moth fam optimization. For evaluating the proposed method, they applied it over Seven multi-classes microarray datasets (Leukemia2, SRBCT, Brain_Tumors2, Brain_Tumors1, Lung_cancer, 9_Tumors, and 11_Tumors) and six binary classes (CNS, Colon, Leukemia1, Breast, Ovarian, and Prostate) using a SVM classifier. Results show that the proposed method detects the informative genes with high accuracy values, and it is able to deliver a competitive number of genes.

Baliarsingh et al. [118] proposed a hybrid FS method that combines ANOVA as a filter and a wrapper. It utilizes the principle of forest optimization algorithm and an enhanced Jaya. They used some binary class microarray datasets (Leukemia-2, Colon tumor, and Ovarian cancer) and other multi-class (Leukemia-3, Lymphoma-3, SRBCT, and Lung cancer-5) for evaluating the performance of the proposed method. The proposed method achieved high classification accuracy and reduced the original size of features set by more than 99%.

In the literature, there are many other studies that used various algorithms to deploy them in the proposed FS methods. An example of this is the study done by Zhang et al. [119], where they used an improved binary krill herd algorithm. In contrast, the study done by Bonilla-Huerta et al. [120] used a strawberry plant algorithm as a wrapper for gene selection in microarray data. Alomari et al. [121] used a hybrid Bat-inspired algorithm as a wrapper while Mufassirin, and Ragel [122] proposed a Wrapper with the best first forward selection searching strategy.

Alanni et al. [123] combined gain ratio filter with an Improved Gene Expression Programming algorithm as a wrapper. They used SVM with 11_Tumors, 9_Tumors, Brain_Tumor1, Brain_Tumor2, Leukemia 1, Leukemia 2, Lung Cancer, and Prostate Tumor datasets for evaluating the proposed hybrid FS method performance.

Alomari et al. [124] deployed flower pollination algorithm. In contrast, Algamal and Lee [125] used an improved adaptive lasso using a new weight as an embedded method. In Brahim and Limam [126], they developed a Cooperative subset search-based wrapper. Tables from Tables 16–19 summarize all these works sorted according to the year of publication.

4.5.1. *Inferences from fifth direction (D5)*

It is noticed that D5 is an attractive and competitive direction for researchers as it handles the advantages of both filters and wrappers or embedded methods. It is clear that GA, when it is combined with an ensemble filter, represents the most popular hybrid FS methods for microarray data processing in the literature. On the other hand, for classification, SVM, NB, and KNN are considered the most popular classifiers used by researchers in this direction with Accuracy and NF as

Table 16
Hybrid feature selection methods using other algorithms in 2020.

Ref	First-part	Second-part	Classifier	Datasets	Datasets dimensions range	performance metrics	Results
[118]	ANOVA filter	principles of enhanced Jaya and forest optimization algorithms as a wrapper	SVM	Leukemia-3, Lung cancer-5 Ovarian cancer, Lymphoma-3, SRBCT, Colon tumor	2000–12600	Matthews correlation coefficient F-measure, Specificity, Recall, Precision, and Accuracy.	The proposed method gives 100% of accuracy with 4 genes for Ovarian cancer dataset. It provides 98.57% of accuracy with four genes and 96.9% with 3 genes for Leukemia-2 and Colon datasets, respectively.
[117]	mRMR filter	quantum Moth fam optimization algorithm as a wrapper	SVM	7 Ma datasets	2000–24481	Accuracy and NF.	The proposed method provides an average accuracy of 100% with less than 4 genes for half of the datasets.
[106]	mRMR filter	TLBOGSA as a wrapper	NB	10 Ma datasets	2000–12600	Sensitivity, Specificity, Matthews Correlation Coefficient, and F-measure.	The proposed method gives higher than 90% of accuracy for 7 out of 10 datasets.
[119]	Information Gain (IG) filter	an improved binary krill herd algorithm as a wrapper	KNN, SVM, NB	Colon, CNS, ALL-AML, Ovarian Cancer, Lung Cancer, ALL-AML-3, ALL-AML-4, MLL, SRBCT	2000–15154	Accuracy and NF.	IG-MBKH outperforms BKH, MBKH, and several newest algorithms based on the number of selected features and classification accuracy. It provides an accuracy of 100% over 4 out of 9 datasets.

Table 17
Hybrid feature selection methods using other algorithms in 2019.

Ref	First-part	Second-part	Classifier	Datasets	Datasets dimensions range	performance metrics	Results
[125]	Screening approach as a filter	an improved adaptive lasso using a new weight as an embedded method	SVM	Colon, Leukemia, Prostate, Lung	2000–12600	Accuracy, gmean and AUC.	The proposed method outperformed SIS-lasso and SIS-SCAD by 4.26 and 2.02%, respectively.
[104]	Independent component analysis filter	artificial bee colony as a wrapper	NB	High-grade glioma, Lung cancer, Colon cancer, Acute leukemia, Prostate tumor, Leukemia2	2000–12600	Accuracy and NF.	The proposed method (ICA-ABC) achieves an accuracy that exceeds 97% for 4 out of 6 datasets. ICA-ABC provides higher accuracy values with a slightly less number of genes using NB, compared to SVM classifier.
[107]	Correlation-based FS as a filter	TLBOSA as a wrapper	SVM, NB, DT, KNN	10 Ma datasets	2000–11225	Accuracy, Sensitivity, Specificity, and F-measure.	The proposed FS method provides high classification accuracy with a small number of genes, especially for Small-Blue-Round-Cell Tumour (SBRCT) dataset where it provides an accuracy that is near 100% with 5 genes.
[108]	ReliefF filter	Recursive Binary Gravitational Search algorithm as a wrapper	Multinomial NB	6 Ma datasets	2000–24481	Accuracy and NF.	ReliefF-RBGSA provides an accuracy of 100% for 5 out of 6 datasets.

a performance metrics. Most of the studies use datasets with dimensions above 2000 genes.

4.6. Sixth direction (D6)

It concerns ensemble FSM. This direction is one of the directions that had the least research attention in the literature.

Some researches proposed new thresholding process in an ensemble based FS method. For example, in the study done by Seijo-Pardo et al. [127], they proposed a new thresholding process in an ensemble-based FS method using six well-known ranker-based FS algorithms, where each ranker provides a specific subset of features using thresholds that depends on the complexity of data, rather than on a fixed threshold. They concluded that the proposed thresholding technique in an ensemble FS method that obtains competitive results, compared to the baseline using six microarray datasets and SVM-RBF as a classifier.

Others developed and compared different ways for combining results of different FS methods as in the study done by Bol'on-Canedo et al. [128], where they developed an ensemble FS method that used five filters, mainly: Correlation-based Feature Selection (CFS), Consistency-based Filter, INTERACT, Information Gain (IG), and ReliefF. They provide a subset of features which is used to train the same classifier. Their results were then combined based on the voting process or cumulative process in Ensemble1, or on using union and intersection for subsets aggregation in Ensemble2. They used four datasets along with four classifiers for evaluating these ensemble strategies.

On the other hand, some researches developed and compared different ensemble designs done by Seijo-Pardo et al. [129], where authors used a combination of some ranking-based filters as an ensemble FS method with two different ensemble designs. The thresholding step was done before finding a combination of various filters which resulted in a Thresholding Combination (TC) design. In contrast, a combination step was performed before the thresholding step in a Combination Thresholding (CT) design. They used SVM classifier over four microarray datasets. It is concluded that TC ensemble design is better than CT design as it reduces the number of features and enhances the test results by reducing the error rate.

Recently, Abdulla et al. [130] proposed G-forest, which is a cost-sensitive ensemble FS method that uses the feature cost through the feature selection process to enforce the selection of a low cost and more informative features. They used random forest over two microarray datasets. The proposed method increases the accuracy up to 14% and decreases the cost up to 56%.

Sayed et al. [131] developed an ensemble FS technique based on *t*-test and Nested-GA. It consists of two nested genetic algorithms: (Outer) which work on microarray datasets, and (Inner) which runs on DNA Methylation datasets. They evaluated the proposed method using SVM over the Lung cancer dataset. This method provides an accuracy that equals 98.4%.

Brahim and Limam [132] presented an ensemble FS method that depends on the assessment of the selector reliability. They used seven microarray datasets with the KNN for evaluating the proposed method. Table 20 provides a summary of these works.

4.6.1. Inferences from sixth direction (D6)

This direction has the least research attention. It can be noticed that researches done on this direction diverse between proposed new thresholding process, developed and compared different ways to combine results, developed a cost sensitive ensemble FS methods, or developed different ensemble designs. SVM is the most used classifier.

4.7. Seventh direction (D7)

This direction concerns various FSM comparison, such as in Shukla et al. [133], where authors demonstrated the effect of Spearman's Correlation (SC) with three filters (ReliefF, Joint Mutual Information, and MRMR) using four classifiers (NB, KNN, DT, and SVM). Results show that SC with MRMR outperforms other combinations over Diffuse Large B Cell Lymphoma dataset.

Pes [134] Different FS methods on different application domains are compared. They used SVM and RF classifiers over Lymphomas and Ovarian datasets to evaluate the proposed method. Results show that the stability patterns for Ovarian and Lymphoma datasets are shown with subset sizes threshold$\leq$20% of features.

Firdausanti et al. [135] compared Crazy Particle Swarm Optimization (CRAZYPSO) and Advanced Binary Ant Colony Optimization (ABACO), while Liang et al. [136] compared the performance of ten FS methods. They used Two breast cancer gene expression datasets extracted from the TCGA and GEO datasets with three classifiers. They documented some observations and recommendations to users.

Bol'on-Canedo et al. [137] analyzed and compared the scalability of some filters, wrappers, and two embedded FS methods, mainly: Recursive feature elimination for support vector machines (SVM-RFE) and Feature selection-perceptron (FS–P). They were compared based on the

Table 18
Hybrid feature selection methods using other algorithms in 2018.

Ref	First-part	Second-part	Classifier	Datasets	Datasets dimension's range	performance metrics	Results
[120]	MI filter	Strawberry Plant Algorithm as a wrapper	SVM	Leukemia, Colon, CNS, Prostate, Lung.	2000–12600	Accuracy and NF.	The proposed method achieves high classification accuracy in 4 out of 5 datasets. It gives 100% of accuracy with 21 genes and 35 genes for Leukemia and Colon datasets, respectively.
[123]	Gain Ratio filter	an Improved Gene Expression Programming algorithm as a wrapper	SVM	8 Ma datasets	5327–12600	Accuracy, NF, and CPU time.	Results showed the effectiveness of the proposed method, as it gives an accuracy that exceeds 98% in 5 out of 8 datasets. It provides an accuracy of 100% for Leukemia1 and Leukemia2 datasets.
[121]	rMRMR filter	Hybrid Bat-inspired Algorithm as a wrapper	SVM	Breast, Colon, ALL AML 4c, and CNS, MLL, ALL AML, ALL AML 3c, Lymphoma, ovarian, SRBCT	2000–24481	Accuracy, NF, Fitness function, and execution time.	The proposed method obtained an equivalent or higher classification accuracy in comparison with those yielded by other competitors in 9 out of 10 datasets.
[100]	F-statistics as a filter	an embedded method called SVMRFE	3 classifiers	4 Ma datasets	2000–12600	Accuracy and NF.	The proposed method provides an accuracy that reaches 100% in the test set with 4 genes for high dimensional datasets.
[122]	Gain Ratio	a Wrapper with the best first forward selection searching strategy	J48, NB, SMO, DL	Colon, Leukemia, Breast, Lung, and Ovarian.	2000–24481	Accuracy and running time in seconds.	The proposed method obtains an accuracy that equals 89.69% over Breast cancer dataset, 95.16% over lung dataset, and 100% for Leukemia and Ovarian datasets.
[124]	mRMR as a filter	flower pollination algorithm as a wrapper	SVM	Colon, Breast, Ovarian.	NA.	Average Accuracy, NF and fitness function.	The proposed method (MRMR-FPA) resulted in producing the smallest gene subset with a competitive accuracy value, compared to the mRMR and combined with GA. For Ovarian and Breast datasets, MRMR-FPA and MRMR-GA provide the same accuracy at 100% and 85.88%, respectively. MRMR-FPA needs 4 genes and 16.80 genes versus 5.87 genes and 22.23 genes, using MRMR-GA for Ovarian and Breast datasets.
[103]	Correlation-based FS filter	Artificial Bee Colony algorithm	SVM	6 binary and multi-class cancer datasets.	2000–7129	Accuracy and NF.	The proposed method outperformed Co-GA and Co–PSO based on a number of selected genes and classification accuracy. It provides 100% of accuracy with less than 5 genes for all datasets except for Colon.
[101]	Fisher's discriminant criterion method and MRMR (Minimum Redundancy Maximum Relevance)	SVM-based wrapper FS method for cancer classification	SVM	Leukemia dataset from the UCI repository	17 129	Accuracy.	SVM- based wrapper FS method outperformed all other methods with an accuracy of 97.13%.

NA means not available.

dependency of the learning method (classifiers and cluster methods). They also compared the performance of eight filters, mainly: Chi-Squared, ReliefF, Information Gain, Minimum redundancy maximum relevance (mRMR), Correlation-based feature selection (CFS), Fast correlation-based filter (FCBF), INTERACT, and Consistency based. For evaluation, they used two real datasets including one microarray dataset called "Colon" with four classifiers, mainly: SVM, KNN, NB, and C4.5. Results show that filters are more scalable than other methods, while wrapper methods' scalability depends on the classifier, as c4.5 provided the best results. IG and Chi-Square observed more stability than other filters as these methods do not take interaction between genes into account. On the other hand, FCBF requires the lowest running time, and mRMR requires the highest one.

Seijo-Pardo et al. [138] analyzed the suitability of different ensemble configurations of FS methods and threshold values with different classifiers over three types of datasets, which are synthetic, real classical, and DNA microarray datasets, which include CNS, Ovarian, Leukemia, Colon, Prostate, Lung, and DLBCL. They used six classifiers, mainly: Ada-Boost, RF, NB, SVM, KNN, and C4.5. Results demonstrated that aggregating E-Stuart, E-RRA, and E-Min FS methods provides the best classification performance, and Fisher ratio is the optimal threshold value.

Kumar et al. [139] compared some filter, wrapper, and fuzzy rough-set-based FS methods. They used two microarray datasets, which are Leukemia and breast cancer, with five classifiers (SVM, KNN, DT, RF, and NB). Results show that a fuzzy rough-set-based FS approach outperformed a correlation-based filter, based on the computation time and a number of selected genes.

Fahy et al. [140] presented a comparative analysis of a univariate filter (Mutual Information) versus a multiivariate filter (Recursive Feature Elimination) method in a hybrid filter-wrapper model using genetic algorithm as a wrapper for FS in DNA microarray data. In their

Table 19
Hybrid feature selection methods using other algorithms in 2016 and 2017.

Ref	First-part	Second-part	Classifier	Datasets	Datasets dimension's range	performance metrics	Results
[102]	CMIM filter	SVM-RFE wrapper	SVM, NB, Linear Discriminant Analysis, KNN.	5 SNP microarray datasets, which are publicly available from NCBI GEO	250000–1000000	Accuracy and F-measure.	The proposed method outperforms all of compared FS methods and achieves up to 96% of accuracy for the used dataset.
[126]	Instance based candidate feature subsets selection as a filter	Cooperative subset search as a wrapper	SVM, KNN	Lymphoma, CNS, Bladder, DLBCL, Gisette, Breast, Lung, Prostate	3036–24482	Error and Stability.	The proposed method provides the best MCE results for 6 out of 8 datasets. HIB-CSS achieved zero-error rate for Lymphoma and Lung cancer datasets.
[116]	Random Forest Ranking	Binary Black Hole Algorithm	7 classifiers include KNN, NB, and RF	Colon Tumor, Central Nervous System (CNS), Leukemia 4C, MLL	NA.	Accuracy and NF.	The proposed method obtained 100% of accuracy on MLL microarray dataset and more than 90% for Colon Tumor and CNS datasets.

NA means not available.

Table 20
Ensemble-based feature selection methods from 2014 to 2021.

Ref	FS Algorithm	Classifier	Datasets	Datasets dimensions range	performance metrics	Results
[130]	G-forest, which is a cost sensitive ensemble FS method	Random Forest	Leukemia and DLBCL	5147–7070	Accuracy, total Cost, Recall, Precision, F-Measure, and NF.	The proposed FS method improves the accuracy up to 14% and decreases costs up to 56%. For Leukemia dataset, it provides an accuracy of 100% and F-measure provides an accuracy of 96% with 1282 genes using G-Forest model. Additionally, for DLBCL dataset, the proposed method provides a percentage above 90% of accuracy and F-measure with 547 genes using G-Forest model.
[131]	An ensemble FS technique based on *t*-test and Nested-GA	SVM	Lung cancer.	17815–27578	Accuracy and NF.	The proposed method provides an accuracy that equals 98.45% using 16 genes.
[129]	Combination-Thresholding (CT) and Thresholding-Combining (TC) are designed based on thresholding and combination steps order in an ensemble FS methods	SVM	Colon, Ovarian, Leukemia2, and Lung	2000–15154	Test error and NF.	TC ensemble design option is better than CT design, as it reduces a number of features and enhances the test results by reducing error rate.
[132]	An ensemble FS approach based on feature selectors reliability assessment (RAA)	KNN	DLBCL, Bladder, Lymphoma, Prostate, Breast, CNS, Lung	3036–12600	F-measure, and stability.	RAA provides better stability for microarray of 5 out of 7 datasets, compared to 7 FS methods.
[127]	New dynamic threshold in an ensemble-based FS method using six well known ranker-based FS algorithms and different complexity measures.	SVM-RBF.	six DNA microarray datasets: Colon, DLBCL, CNS, Leukemia, Lung, and Ovarian.	2000–15154	Classification error rate and NF.	The proposed thresholding technique in ensemble FS methods obtains competitive results compared to the baseline. For Ovarian dataset, it does not provide classification error using 14 genes only.
[128]	Two ensemble strategies for 5 filters	NB, SVM, C4.5, IBI	CNS, Ovarian, CNS, and Leukemia Ma datasets	2000–15154	Test classification error and NF.	Ensemble1 using cumulative probabilities as a combination method (E1-CP) with C4.5 is better for classifying classical datasets, while (E1-CP) with SVM is better when the number of features is greater than the number of samples. Ensemble1 provides less error rate than Ensemble2 with all used classifiers for all microarray datasets, but it needs more genes.

work, they analyzed the performance of MI with GA, RFE with GA, and MI-RFE with GA. They used SVM classifier with three datasets called DLBCL, follicular, and lymphomas for comparison. Results show that multiivariate methods outperform univariate methods across both datasets. In addition, MI-RFE with GA yielded a better performance than using each filter individually.

Drot'ar et al. [141] compared stability, similarity, and influence on prediction performance of ten-filter methods in the literature. They used six high dimensional microarray datasets and two smaller biomedical datasets for differential diagnosis of Parkinson's disease. For classification, they used SVM, AdaBoost, Random Forests, and Deep Belief Network. Results show that univariate FS methods are more stable than multiivariate methods.

Akila and Christe [142] compared mRMR, Relief, and Fast Correlation-based filter FS methods classification performance using DT, RF, NB, KNN, Artificial Neural Network (ANN), and Sequential Minimal Optimization (SMO) classifiers with ALL-ML, Breast, Colon, Ovarian, Lung, and Lymphoma datasets. Results showed that Relief with SMO as a classifier outperformed all other algorithms such as MRMR and FCBF for all datasets.

Table 21
Summary of studies that compare different feature selection methods from 2019 to 2021.

Ref	FS Algorithm	Classifier	Datasets	Datasets dimensions range	performance metrics	Results
[144]	Comparing 10 of the commonly-used entropy-based, similarity-based and statistics-based filter FS methods	KNN, SVM, MLP	10 Ma datasets	2309–12625	Average Accuracy.	Mutual Information gives the best results among all examined entropy-based filter FS methods while ReliefF and Chi-square performs best in the category of similarity-based and statistics-based filter FS methods, respectively. Chi-square is the best choice for a bi-class dataset, as it provides an average accuracy that equals 95.54%, while MI gives better results for multi-class datasets with an average accuracy of 95.635%.
[142]	Compared 3 filters experimentally	6 classifiers	6 Ma datasets	2000–15154	Accuracy and NF.	Relief with SMO as a classifier outperformed all other algorithms like MRMR and FCBF in all datasets. Relief with SMO provides 100% of accuracy in Ovarian and Lymphoma dataset.
[133]	Compared the effect of Spearman's correlation (SC) with three filters	NB, KNN, DT, SVM	diffuse large B-cell lymphoma (DLBCL)	4026	Accuracy and NF.	SC with MRMR using NB classifier for 20 genes outperformed other combinations over Diffuse Large B Cell Lymphoma.
[134]	Compared different kinds of selection algorithms on different application domains	SVM, RF	Lymphomas, Ovarian	4026–15154	Stability, AUC, and NF.	Stability patterns of different ranking methods are shown with subset sizes th$\leq$20% of features for Lymphomas and Ovarian datasets.
[135]	Compared CRAZYPSO and ABACO FS methods	SVM	Colon and Prostate datasets	2000–2135	Accuracy, time, and NF.	ABACO algorithm outperformed CRAZYPSO-based on classification accuracy and the number of the chosen genes, but it needs longer running time. For Prostate dataset, ABACO needs 8 h using 126 genes for providing an accuracy of 96% versus an accuracy of 95.18% using CRAZYPSO with 457 genes after 2 h.
[145]	Used 5 methods of similarity-based techniques (Fisher, Relief, SPEC, Trace Ration, Laplace) to find 2000 most relevant genes,	KNN	miRNA microarray profiles GSE17861 dataset	NA	Accuracy.	Fisher method gave better results with 3 and 5-nearest neighbors. It achieved an accuracy that equals 94.50% and 82.33.

NA means not available.

Table 22
Summary of studies that compare different feature selection methods in 2018.

Ref	FS Algorithm	Classifier	Datasets	Datasets dimensions range	performance metrics	Results
[146]	Compared F-test, T-test, Signal-to-noise ratio, ReliefF, Pearson product moment correlation coefficient FS methods performance	SVM, KNN, Linear Discriminant Analysis, DT, NB	Leukemia cancer, Lung cancer, Lymphoma, CNS, and Ovarian Cancer	7070–15154	Accuracy and NF.	Results show that Signal-to-noise-ratio (SR) method with KNN provides better classification accuracy with less genes, compared to other applied methods. SR using KNN classifier provides 100% of accuracy in 4 out of 5 datasets.
[136]	Compared the performance of ten FS methods.	SVM, NB, and Logistic Regression.	Two breast cancer gene expression datasets extracted from TCGA and GEO datasets.	NA	Accuracy and running time.	WL2Boost method provides the best performance with low dimensional datasets when users do not care about the running time, while RP-CLR and PCU-CLR methods achieves better performance for a high dimensional dataset with a competitive time.
[137]	Compared the scalability of some filter, wrapper, and embedded FS methods in literature	SVM, KNN, NB, C4.5	Colon	2000	Scalability based on accuracy, stability (distance), and computational time (training time) in seconds.	Filters are the most scalable FS approach, while wrapper scalability depends on the choice of the classifier, as C4.5 provides good scalability.

NA means not available.

In [143], authors compared some supervised and unsupervised feature selection methods using eight microarray datasets. They displayed the advantages and disadvantages of the supervised versus unsupervised. They compared the simulation results of wrapper-based feature selection approaches with filter-based approaches. There are other studies in this direction, such as [144–150]. Tables from Tables 21–25 summarize all studies that were conducted in this direction.

4.7.1. *Inferences from seventh direction (D7)*

This direction has a medium attention from researchers. It is noticeable that all comparison researches diverse between comparing filter-based versus wrapper-based FS methods, comparing different filters performance with specific wrapper, comparing univariate filters versus multiivariate filters, comparing different FS methods in different application domains, and comparing different ensemble configurations. Most studies use Accuracy and NF as evaluation metrics while in a little number of studies, the comparison was done in terms of stability,

Table 23
Summary of studies that compare different feature selection methods in 2017.

Ref	FS Algorithm	Classifier	Datasets	Datasets dimensions range	performance metrics	Results
[147]	Compared the effects of combining Kendall Correlation (KC) with some filter-based FS methods.	KNN, SVM, NB, DT	SRBCT	2308	Accuracy and NF.	The combination of KC and mRMR outperforms other methods, as it gives an accuracy of 92.87 with 100 genes using NB classifier.
[138]	Compared different ensemble configurations of FS methods and threshold values with different classifiers based on suitability	Ada-Boost, RF, NB, SVM, KNN, C4.5	CNS, Ovarian, Leukemia, Colon, Prostate, Lung, and DLBCL	2000–15154	Average test error and NF.	The best results are obtained from aggregating three methods, which are named E-Stuart, E-RRA, and E-Min. Fisher ratio (FR) is the optimal threshold value for DNA microarray datasets, as it selects only 0.09% of the features. FR E-min achieved the best results for four datasets (DLBCL, CNS, Prostate, and Ovarian).
[139]	Compared some filter, wrapper, and fuzzy- rough_set_based FS methods	SVM, KNN, DT, RF, NB	Leukemia, breast cancer	7129–24481	Accuracy, time and NF.	Correlation-based feature subset selection provides the highest accuracy for Breast Cancer dataset using Random Forest classifier that equals 90.72%. Fuzzy Rough Set Attribute Evaluation and Correlation-based attribute evaluation that gives 100% of accuracy for Leukemia dataset, using the KNN classifier.

Table 24
Summary of studies that compare different feature selection methods in 2016.

Ref	FS Algorithm	Classifier	Datasets	Datasets dimensions range	performance metrics	Results
[143]	Compared various supervised and unsupervised FS methods used for gene selection in microarray data in the literature	NB	8 Ma datasets (11 Tumors, Leukemia1, Leukemia2, Lung Cancer, SRBCT, Brain Tumor1, Prostate Tumor, and DLBCL).	5470–12601	Accuracy and NF.	In supervised FS methods, it was observed that the classification accuracy is increased to a great extent using the top 50 and 100 selected genes, compared to the situation in which all genes are considered. In unsupervised FS methods, MCFS outperformed LAPLACIAN SCORE using the top 50 and 100 of the selected genes based on classification accuracy.
[148]	Compared different FS methods for microarray data processing and documented best results achieved over each dataset based on the accuracy and the number of the chosen genes	SVM, KNN	Colon Tumor, Leukemia ALL_AML	2000–7129	Accuracy and NF.	The best classification accuracy (95%) is achieved using 50 genes for colon dataset using ReliefF. For Leukemia dataset, Relief-F and Chi Square test provide an accuracy of 97% with 40 genes using KNN classifier.
[149]	Compared 3 Wrapperbased FS methods called Genetic Algorithm, Particle Swarm Optimization, and Ant Colony Optimization	NB, DT, Rule Induction, KNN	CNS, Colon cancer, Lung cancer, breast cancer, ALL MLL, ALL MLL 3, ALL MLL 4, MLL, SRBCT	2001–24482	F-Measure, ROC and NF.	Best classification performance was provided by NB with 4 datasets based on area, and ACO has higher F-measure of 9 datasets and ROC Area of 7 datasets than GA and PSO.

similarity, performance, and scalability.

4.8. Eighth direction (D8)

It concerns the distribution of FSM, which can be done horizontally H (based on samples) or vertically V (based on features). An example of this is the study done by Potharaju and Sreedevi [151] where they proposed a distributed FS strategy based on a Symmetrical Uncertainty (SU) and Multi-Layer Perceptron (MLP). They used KNN, SVM, SC, and Ridor classifiers with seven high dimensional Ma datasets and one lower dimension dataset for evaluating the proposed strategy. The proposed method satisfied 18% of the competitive accuracy against traditional methods.

Ebrahimpour and Eftekhari [152] proposed a distributed FS framework that starts by partitioning the dataset vertically based on a hesitant fuzzy sets, then applying a second partition, which is called shuffling that randomly partitions the dataset. Finally, obtained results were merged to find the final set of the selected genes that depend on the improvements of the classification accuracy. For evaluating the proposed framework, they used four classifiers and eight microarray datasets. The proposed method provided competitive results compared to other methods in the literature, based on accuracy and the number of the selected genes.

Alhamidi et al. [153] proposed a homogeneous distribution ensemble FS method. They used a two-dimensional partition (vertical and horizontal) for partition data. The proposed method speeds the FS process up by almost two times using two microarray datasets and DT classifier.

Authors in Moran-Fern and Bol'on-Canedo [154] proposed a distribution approach for partitioning data horizontally and vertically, and then by merging the partial outputs. They evaluated the proposed method over 11 datasets (five of them are of microarray datasets) using four classifiers: C4.5, KNN, NB, and SVM. They provided some recommendations for users regarding the suitable partitioning of high dimensional datasets based on their goals. Users recommended using the horizontal partition if reducing storage requirements and running time

Table 25
Summary of studies that compare different feature selection methods from 2014 to 2015.

Ref	FS Algorithm	Classifier	Datasets	Datasets dimensions range	performance metrics	Results
[141]	Compared stability, similarity, and influence on performance prediction of 10 filter FS methods in the literature	SVM	six high dimensional Ma datasets and two smaller biomedical datasets for differential diagnosis of Parkinson's disease	204–22283	Accuracy, AUC, Stability.	Univariate FS methods are more stable than Multiivariate methods.
[140]	Compared a univariate filter versus a multiivariate filter method in a hybrid filter-wrapper model for FS	SVM	DLBCL and follicular lymphomas, Prostate, and tumor	5470–10510	Accuracy and NF.	Results show that multiivariate methods outperform univariate methods across both datasets.
[150]	Compared some filterbased versus wrapper-based gene selection techniques	NB, Discriminant Analysis, NN, AdaBoostMl, GentleBoost, RobustBoost, Bootstrap Aggregation	Ovarian Cancer, Lymphomas, and Leukemia	4000–7129	Accuracy and processing time.	Using filter as ReliefF algorithms with most of the supervised classifiers provides better accuracy while accuracy achieved 99.08% for ovarian cancer dataset using wrapper approach with SVM, with the minimum computation time.

Table 26
Summary of a distributed-based feature selection strategy related works from 2017 to 2019.

Ref	FS Algorithm	Classifier	Datasets	Datasets dimensions range	performance metrics	Results	Type (V or H)
[151]	Distributed FS strategy based on Symmetric Uncertainty (SU), Correlation-based Feature Subset Selection (CFS), and Multi-Layer Perceptron (MLP), called DFS	4 classifiers	8 Ma datasets	2000–12582	Accuracy and RMS error rate, and the number of features in the best cluster (SF).	DFS provides a successful rate that equals 57% and 18% improvement rate, compared to the traditional methods.	V.
[152]	Hybrid partitioning method that combines greedy and shuffling partitioning ways	C4.5, NB, KNN, SVM	Colon, DLBCL, CNS, Leukemia, Prostate, Lung, Ovarian, Breast	2000–24481	Average Accuracy and NF.	The proposed method provides an average accuracy at a percentage above 90%, using SVM, KNN, and C4.5 classifiers with an acceptable number of the selected genes	V.
[153]	Homogeneous distribution ensemble FS method. They used two dimensional partitions (vertical and horizontal) for partition data	DT	seven datasets, where two of them are of Ma datasets	12533–12600	Accuracy, sensitivity, specificity, precision, AUC, and time complexity.	The proposed method can improve the previous method at a percentage of 2% for some datasets, and it can speed up the process by almost 2 times	H and V.
[154]	Distribution approach for partitioning data horizontally and vertically, and then by combining partial outputs using a merging process, based on the theoretical complexity of these feature subsets	C4.5, KNN, NB, SVM	11 datasets, 3 of them are of Ma.	12534-24, 481	Accuracy, NF, and running time in seconds.	Horizontal partition is better if reducing storage requirements and running time was more important to them than the classification accuracy. Otherwise, using vertical partition is better	H one time and V next time.

was more important to them than classification accuracy. All of these works are summarized in Table26.

4.8.1. *Inferences from eighth direction (D8)*

This direction has the least research attention. It is observed that partitioning the dataset vertically is widely used, compared to horizontal partitioning.

4.9. *Ninth direction (D9)*

It concerns parallel FSM. An example is the study done by Venkataramana et al. [155], where they proposed a parallel FS framework called HFS that involves using a Correlation-based feature subset selection. They then used ranking-based feature selection methods to rank features and select the optimal features only along with a specific classifier in parallel. This method works by splitting the dataset into a number of chunks, where each chunk contains some features. Then it applies Correlation-based feature subset selection overall chunks in parallel, and obtains the best features subset from each chunk. Finally, it constructs the learning model, which is called parallelized decision tree and random forest, for evaluating a subset of features to choose the subset that provides the best accuracy after applying four rank-based FS methods in parallel. The proposed methodology was applied over two microarray datasets, which are Childhood leukemia and Gastric cancer. Experimental results show that the proposed method outperformed both RWFS and parallelized ranking-based feature selection methods.

Keco et al. [156] developed a Genetic algorithm (GA) in parallel using the Hadoop Map-Reduce framework. They used 11 GEMS data sets and two classifiers for evaluating the proposed method. Using the proposed method achieved an accuracy that reached 100% for less than 25 selected genes.

The work done by Ray et al. [157] is one of the parallel-based approach examples, where a mutual information FS method was used on Spark framework and applied over some microarray datasets using various classifiers.

Boucheham and Batouche [158] developed a massively parallel meta-ensemble FS method to select a robust subset of genes. It aggregates results for each filter within each ensemble and aggregates the outcomes of all ensembles in parallel, using MATLAB Parallel Computing Toolbox (PCT). They evaluated the proposed method using three classifiers over five microarray datasets. The proposed method is flexible, and it provides good performance.

Table 27
Summary for parallel-based feature selection strategy related works from 2018 to 2021.

Ref	FS Algorithm	Classifier	Datasets	Datasets dimensions range	performance metrics	Results
[160]	Two partition strategy for preparing data for parallelism as Vertical partitioning based on feature space, and horizontal partitioning based on samples. They also proposed a Parallel Multilevel FS algorithm to select optimal features set	Parallel logistic regression, Parallel decision tree, Parallel multinomial linear regression, Parallel SVM, Parallel gradient boosted trees, Parallel linear SVM, Parallel Random Forest	MGE, P53 Mutant	4522-12 625	Accuracy, TPR, FPR, Precision, Recall, F1-Score, F-Measure, NF, and computation time in seconds.	Parallel multilevel feature selection method improves cancer classification accuracy with a range from 85% to 99% with improvements in the computation time, compared to other parallel methods using parallel Random Forest classifier.
[155]	Parallelized hybrid feature selection (HFS) framework, which involves using a Correlation-based feature subset selection and ranking based feature selection methods that rank the features and select the optimal features only, in parallel	Decision tree and random forest in parallel	Childhood leukemia and Gastric cancer	4522–8280	Accuracy, NF, and execution time in seconds.	The proposed method improves classification accuracy compared to RWFS and parallelized RFS method with a percentage that reaches 15%. Parallelized HFS provides an accuracy of 97% and 79% for gastric cancer and childhood leukemia, respectively.
[156]	Cloud computing-based Map Reduce parallel GA	ANN and SVM	11 GEMS data sets (9 tumors, 11 tumors, 14 tumors, brain tumor 1, lung cancer, brain tumor 2, leukemia 1, DLBCL, leukemia 2, SRBCT, and prostate tumor)	5726–15009	Accuracy and NF.	The proposed method can be effectively implemented for real-world microarray data in the cloud environment. Its accuracy reached 100% for less than 25 selected features.

Table 28
Summary for parallel-based feature selection strategy related works from 2016 to 2017.

Ref	FS Algorithm	Classifier	Datasets	Datasets dimensions range	performance metrics	Results
[161]	Chi-Square FS method in parallel using Spark	parallel logistic regression and SVM	Childhood Tumor Gene dataset with Binary class	9945	Accuracy and NF.	The proposed framework provides 63% of accuracy with 25 genes using Parallel logistic regression and 75% using Parallel SVM classifier.
[159]	sf-ANOVA that is a statistical test that is used to select the most informative features using Spark framework	Naive Bayes (sf-NB) and Logistic Regression (sf-LoR)	3 standard large Ma data: GSE13159, GSE13204, and GSE15061 from National center of Biotechnology Information.	1480–54675	Accuracy and time consumed by classifiers on Spark cluster, and conventional system time in seconds.	The proposed approach provides better accuracy with Naive Bayes compared to Logistic Regression. Its performance increases with the increase in data size. For GSE13204, it provides an accuracy that equals 84.76% using NB and 90.03% using Logistic Regression.
[157]	Mutual information FS method using Spark framework	SVM and Logistic Regression based on Spark framework	3 Ma datasets from National center of Biotechnology Information about Leukemia raw dataset: GSE13159, GSE13159, and GSE13204	1480–54675	Accuracy and time consumed by classifiers on Spark cluster, and conventional system time in seconds.	The proposed method performance increases when the data size increases. It provides better accuracy when it is used with SVM than with Logistic Regression. For GSE13204, it provides an accuracy that equals 83.48% using SVM and 78.97% using Logistic Regression.
[162]	MapReduce based statistical test for FS	MapReduce based KNN (mrKNN).	3 Ma benchmark datasets taken from the NCBI GEO repository.	1480–54675	Accuracy, Precision, Recall, and execution time.	The proposed models need less execution time for large data than other models.

There are many other works done on this area such as [159–165]. They are all summarized in Tables 27–29.

4.9.1. Inferences from ninth direction (D9)

This direction has a middle research attention. It is observed that most of the studies evaluated the proposed parallel framework based on Accuracy and NF.

5. Development of feature selection publications over the recent seven years

By tracking the development of research done in the recent seven years over all the proposed directions that we discussed in this survey, it can be noted that there is an increase in the number of research papers published in the last four years, especially the years of 2018 and 2019 compared to years before 2018, and this is noted for most directions.

Table 29
Summary for parallel-based feature selection strategy related works from 2014 to 2015.

Ref	FS Algorithm	Classifier	Datasets	Datasets dimensions range	performance metrics	Results
[164]	ANOVA is a Proposed statistical test using a Hadoop MapReduce	KNN	Leukemia1, Ovarian Cancer, Breast Cancer, MULTMYEL, and Leukemia	7129–54675	Accuracy, running time, and NF.	Results indicate that as the size of data increases, running time of the proposed method decreases. The accuracy reaches 100% in 3 out of 5 datasets using ANOVA.
[165]	MapReduce framework is used to select the relevant genes by proposing various statistical methods	MapReduce based proximal support vector machine classifier	Leukemia, Breast cancer, Ovarian cancer, MULTMYEL, GSE15061, and Leukemia	7129–54675	Accuracy, recall, specificity, precision, and NF.	This study produces good results based on the accuracy rate on the benchmark datasets which were deployed for evaluation process.
[163]	Two stages parallel FS strategy were used. In the first stage, they used a parallel K-means on MapReduce for clustering features, and then they applied an iterative MapReduce that implements a parallel SNR ranking for each cluster.	SVM	2 datasets of cancer RNA-seq gene expression data and 4 gene expression Ma datasets, 2 ovarian cancer datasets, gastric cancer dataset, and Esophageal Squamous Cell carcinoma dataset.	4522–54675	Sensitivity, Specificity, Accuracy, and NF.	The proposed method provides 100% of accuracy in 4 out of 6 datasets.
[158]	Massively parallel meta-ensemble (MPME) FS method to select a robust subset of genes	SVM, KNN, ANN.	Ovarian, Leukemia, DLBCL, Colon, and SRBCT.	2000–15154	Sensitivity, Specificity, and NF.	MPME-FS provides an average sensitivity of 0.963 and specificity of 0.972 with 31 genes using ensemble of SVM, KNN, and ANN classifiers and information gain filter.

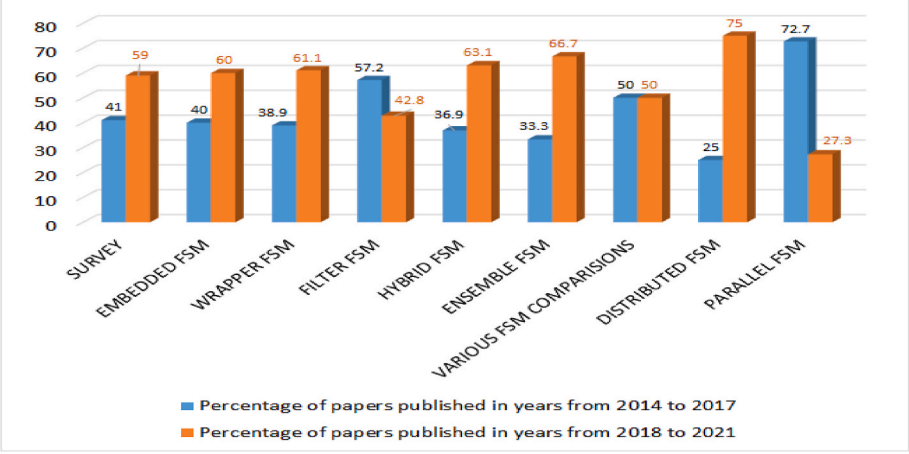

Fig. 4. Comparing the percentages of papers published in [2014–2017] and [2018–2021].

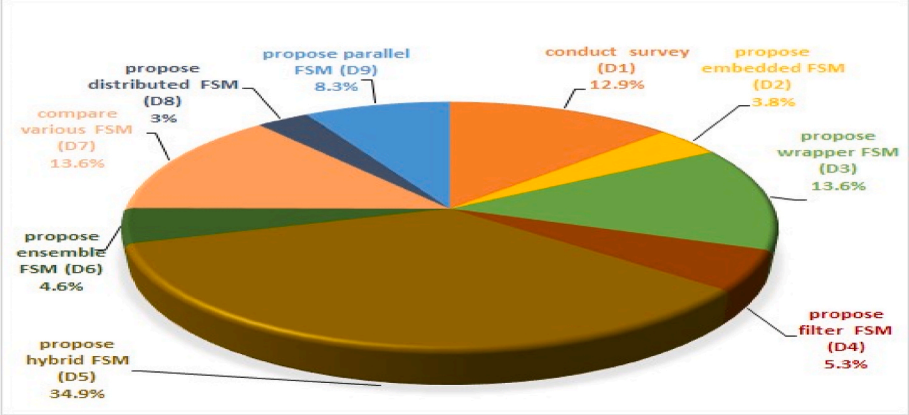

Fig. 5. Comparing the percentages of publications in each direction for Springer, Elsevier and IEEE.

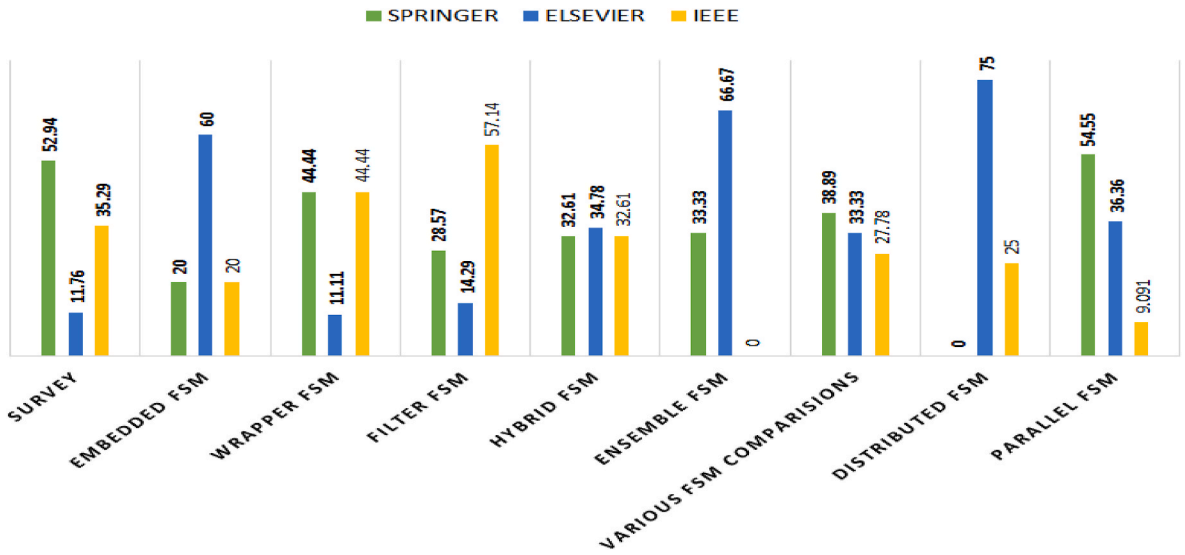

Fig. 6. Comparing the percentages of publications in each direction for Springer, Elsevier and IEEE.

We noticed that there are more works published in relation to D1, D2, D3, D5, D6, and D8 in the recent three years (2018–2021) than works published in the period of 2014–2017, as shown in Fig. 4.

After 2016, work started employing distributed approaches and using an ensemble FS approaches for microarray datasets. In the period of 2014–2015, most research papers used original PSO swarm intelligence algorithm, after approving it as a wrapper FSM for microarray data processing. After 2015, researchers deployed other algorithms.

In 2015 and 2016, researchers developed filter FSM for unsupervised learning. They developed a Score-based criteria fusion (SCF) as a filter FS method after 2018.

GA with an ensemble filter, represents the most popular hybrid FS method for microarray data processing in the literature. In 2016 and 2018, researchers used an Ant colony and PSO as a part of a hybrid FSM for microarray data processing.

6. Observations and analyses

In this systematic review, we investigated 132 research papers available from three famous publishers (Elsevier, Springer, and IEEE) about FS for microarray data processing during the last seven years. We observed that researchers focused on the fifth direction (D5) "Hybrid FSM", as it constituted 34.9% of the researches that were examined. We believe that the reason for this high percentage could be due to the fact that hybrid methods generally improve classification accuracy without causing bad effects on the computation time.

On the other hand, five directions had the least researches attention in the literature, especially the second direction "Embedded FSM (D2)" which constituted a percentage of 3.8%, the fourth direction "Filter FSM (D4)" which constituted a percentage of [5.3%], the sixth direction "Ensemble FSM (D6)" which constituted a percentage of [4.6%], and the ninth direction "Parallel FSM (D9)" with a percentage of [8.3%], and the eighth direction "Distributed FSM (D8)" with a percentage of [3%]. (D9) could reduce computation time, but it has penalty of reducing classification accuracy. The rest of directions had medium attention from researchers. They include the third direction "Wrapper FSM (D3)" and the seventh direction "various FSM comparison (D7)" with equal percentage values that equal 13.6%. The first direction "Survey (D1)" has a quite smaller percentage that equals 12.9%; see Fig. 5 for more details. The percentages are rounded to one decimal digit to simplify readability.

Based on Fig. 6, which presents the percentages of publications for each direction in Elsevier, Springer, and IEEE, we observed that the second direction "propose embedded FSM (D2)", the fifth direction "Hybrid FSM (D5)", the sixth direction "Ensemble FSM (D6)", and the eighth direction "Distributed FSM (D8)" have the highest percentages in Elsevier. Conversely, the first direction "Survey (D1)", the third direction "Wrapper FSM (D3)", the seventh direction "various FSM comparison (D7)", and the ninth direction "Parallel FSM (D9)" represent the most research attention in Springer. On the other hand, papers published in IEEE concentrate more on the third direction "Wrapper FSM (D3)" and the fourth direction "Filter FSM (D4)" directions.

Our findings from this systematic survey can be used as a good source for researchers and PhD students who are engaged in FS research. This study summarizes the directions of FS that received the most, the middle, and the least research attention in literature. We presented a detailed review of the required information for researchers who plan to pursue research on FS in the field of microarray data processing to guide them selecting the most competitive FS directions.

Furthermore, if researchers have already specified their direction, they will be furnished with a good deal of information about the previous studies done in relation to the chosen direction, based on the methods, the classifiers, the datasets, the dimension's range, the performance metrics, and the results obtained. This may be used in order to simplify the comparison process of their work with the most recent state-of-the-art works in that direction.

7. Conclusion and future work

We examined all papers concerned FS field for microarray data processing during the recent seven years, which were published in three famous publishers (Elsevier, Springer, and IEEE). We found that 38% of these papers are published in Springer. The reviewed papers were categorized based on their main purposes into nine directions, then they were summarized according to what studies received the most, the middle, and the least research attention in all 132 papers that were reviewed in this survey. This systematic review compared the publications in each direction for the selected publishers.

We observed from this review that proposing a hybrid FSM takes the most attention of researchers in the literature in all of the 132 papers reviewed, with a percentage of 34.9%. In addition, wrapper-based FSM takes the middle percentage from publications in the literature, while filter and parallel FS methods are one type of the directions that takes the least attention. The hybrid approach tries to improve classification accuracy without negatively affecting the computation time (which is still longer than computation time consumed using filter of parallel FS methods). Therefore, researchers have a challenge to find more efficient

classifier-based FS methods (hybrid and wrapper) with an acceptable computation time, which is approximately near to the time needed by filter or parallel FS methods. They may also face the challenge to search for parallel and filter FS methods that provide better accuracy. The chance is still open for proposing an embedded, an ensemble, and distributed FS methods.

Furthermore, we observed that "Embedded FSM", "Hybrid FSM", "Ensemble FSM", and "Distributed FSM" have the highest percentages in Elsevier, compared to the other two publishers. On the other hand, "Survey", "Wrapper FSM", "compare various FSM", and "propose parallel FSM" directions represent the most research attention in Springer. The IEEE, however, had higher percentages than the other two publishers regarding "Wrapper FSM" and "Filter FSM" directions.

Based on these observations, we recommend that researchers may do searches in order to find more efficient hybrid or wrapper-based FS methods for microarray data processing that can provide better computation time than the already existing ones, or they may propose strategies that can effectively increase the accuracy in "Parallel FSM" or "Filter FSM" directions.

For future work, we will continue exploring these directions, and we will develop new hybrid or parallel feature selection methods, since these directions still need more investigation.

CRediT author statement

Esra'a Alhenawi: Methodology, Writing- Original draft preparation, validation. **Rizik Al-Sayyed:** Supervision, Writing - review & editing, Visualization, Validation. **Amjad Hudaib:** Supervision, Writing - review, Validation. **Seyedali Mirjalili:** Supervision, Review & editing, Visualization, Validation.

Declaration of competing interest

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Acknowledgements

We would like to thank all people who provided a technical help, assisted in reviewing and editing the language of the manuscript, and to those who offered general support and useful comments regarding this manuscript.

References

[1] O. Dagliyan, F. Uney-Yuksektepe, I.H. Kavakli, M. Turkay, Optimization based tumor classification from microarray gene expression data, PLoS One 6 (2) (2011), e14579.
[2] G. Manikandan, S. Abirami, A survey on feature selection and extraction techniques for high-dimensional microarray datasets, in: Knowledge Computing and its Applications, Springer, 2018, pp. 311–333.
[3] B. Remeseiro, V. Bolon-Canedo, A review of feature selection methods in medical applications, Comput. Biol. Med. 112 (2019) 103375.
[4] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Electr. Eng. 40 (1) (2014) 16–28.
[5] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, Distributed feature selection: an application to microarray data classification, Appl. Soft Comput. 30 (2015) 136–150.
[6] T. Saw, P.H. Myint, Swarm intelligence based feature selection for high dimensional classification: a literature survey, Int. J. Comput. 33 (1) (2019) 69–83.
[7] B. Seijo-Pardo, I. Porto-Díaz, V. Bolón-Canedo, A. Alonso-Betanzos, Ensemble feature selection: homogeneous and heterogeneous approaches, Knowl. Base Syst. 118 (2017) 124–139.
[8] G. Manikandan, S. Abirami, Feature selection is important: state-of-the-art methods and application domains of feature selection on high-dimensional data, in: Applications in Ubiquitous Computing, Springer, 2021, pp. 177–196.
[9] T. Almutiri, F. Saeed, Review on feature selection methods for gene expression data classification, in: International Conference of Reliable Information and Communication Technology, Springer, 2019, pp. 24–34.
[10] A.K. Shukla, D. Tripathi, B.R. Reddy, D. Chandramohan, A Study on Metaheuristics Approaches for Gene Selection in Microarray Data: Algorithms, Applications and Open Challenges, Evolutionary Intelligence, 2019, pp. 1–21.
[11] A. Alonso-Betanzos, V. Bolón-Canedo, L. Morán-Fernández, B. Seijo-Pardo, Feature selection applied to microarray data, in: Microarray Bioinformatics, Springer, 2019, pp. 123–152.
[12] N. Sánchez-Maroño, O. Fontenla-Romero, B. Pérez-Sánchez, Classification of microarray data, in: Microarray Bioinformatics, Springer, 2019, pp. 185–205.
[13] R.K. Singh, M. Sivabalakrishnan, Feature selection of gene expression data for cancer classification: a review, Procedia Computer Science 50 (2015) 52–57.
[14] N.A. Zamri, B. Thangavel, N.A. Ab Aziz, N.H.A. Aziz, Review on the usage of swarm intelligence in gene expression data, in: International Conference for Innovation in Biomedical Engineering and Life Sciences, Springer, 2017, pp. 153–160.
[15] N. Almugren, H. Alshamlan, A survey on hybrid feature selection methods in microarray gene expression data for cancer classification, IEEE Access 7 (2019) 78533–78548.
[16] J.C. Ang, A. Mirzal, H. Haron, H.N.A. Hamed, Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection, IEEE ACM Trans. Comput. Biol. Bioinf 13 (5) (2015) 971–989.
[17] V. Bolón-Canedo, A. Alonso-Betanzos, I. López-de Ullibarri, R. Cao, Challenges and future trends for microarray analysis, in: Microarray Bioinformatics, Springer, 2019, pp. 283–293.
[18] S. Vanjimalar, D. Ramyachitra, P. Manikandan, A review on feature selection techniques for gene expression data, in: 2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), IEEE, 2018, pp. 1–4.
[19] S.D. Bharathi, S. Sudha, A survey on gene selection for microarray cancer classification based on soft computing techniques, in: 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE, 2018, pp. 304–309.
[20] A. Jović, K. Brkić, N. Bogunović, A review of feature selection methods with applications, in: 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO), Ieee, 2015, pp. 1200–1205.
[21] K.P. Shroff, H.H. Maheta, A comparative study of various feature selection techniques in high-dimensional data set to improve classification accuracy, in: 2015 International Conference on Computer Communication and Informatics (ICCCI), IEEE, 2015, pp. 1–6.
[22] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, Feature selection in dna microarray classification, in: Feature Selection for High-Dimensional Data, Springer, 2015, pp. 61–94.
[23] Z. Mungloo-Dilmohamud, Y. Jaufeerally-Fakim, C. Peña-Reyes, A meta-review of feature selection techniques in the context of microarray data, in: International Conference on Bioinformatics and Biomedical Engineering, Springer, 2017, pp. 33–49.
[24] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, An ensemble of filters and classifiers for microarray data classification, Pattern Recogn. 45 (1) (2012) 531–539.
[25] J. Li, K. Cheng, S. Wang, F. Morstatter, R.P. Trevino, J. Tang, H. Liu, Feature selection: a data perspective, ACM Comput. Surv. 50 (6) (2017) 1–45.
[26] D. Cai, C. Zhang, X. He, Unsupervised feature selection for multi-cluster data, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010, pp. 333–342.
[27] R. Xu, S. Damelin, B. Nadler, D.C. Wunsch II, Clustering of high-dimensional gene expression data with feature filtering methods and diffusion maps, Artif. Intell. Med. 48 (2–3) (2010) 91–98.
[28] J. Zhao, K. Lu, X. He, Locality sensitive semi-supervised feature selection, Neurocomputing 71 (10–12) (2008) 1842–1849.
[29] S. Huang, Supervised feature selection: a tutorial, Artif. Intell. Res. 4 (3) (2015), https://doi.org/10.5430/air.v4n2p22.
[30] L. Fu, T. Zhu, K. Zhu, Y. Yang, Condition monitoring for the roller bearings of wind turbines under variable working conditions based on the Fisher score and permutation entropy, Energies 12 (16) (2019) 3085.
[31] M.A. Sulaiman, J. Labadin, Feature selection based on mutual information, in: 2015 9th International Conference on IT in Asia, CITA), 2015, pp. 1–6, https://doi.org/10.1109/CITA.2015.7349827.
[32] X. He, D. Cai, P. Niyogi, Laplacian score for feature selection, Adv. Neural Inf. Process. Syst. 18 (2005).
[33] L.-X. Zhang, J.-X. Wang, Y.-N. Zhao, Z.-H. Yang, A novel hybrid feature selection algorithm: using relieff estimation for ga-wrapper search, in: Proceedings of the 2003 International Conference on Machine Learning and Cybernetics, vol. 1, IEEE, 2003, pp. 380–384. IEEE Cat. No. 03EX693.
[34] M. Mandal, A. Mukhopadhyay, An improved minimum redundancy maximum relevance approach for feature selection in gene expression data, Procedia Technology 10 (2013) 20–27.
[35] Y. Saeys, I. Inza, P. Larranaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.
[36] I.A. Gheyas, L.S. Smith, Feature subset selection in large dimensionality domains, Pattern Recogn. 43 (1) (2010) 5–13.

[37] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Optimization by simulated annealing, Science 220 (4598) (1983) 671–680.
[38] M. Abdel-Basset, W. Ding, D. El-Shahat, A hybrid harris hawks optimization algorithm with simulated annealing for feature selection, Artif. Intell. Rev. 54 (1) (2021) 593–637.
[39] M. Mitchell, An Introduction to Genetic Algorithms, 1996. Cambridge.
[40] J. Kennedy, R. Eberhart, Ieee, particle swarm optimization, in: 1995 Ieee International Conference on Neural Networks Proceedings vol. 1, 1948, p. 61995.
[41] M. Dorigo, L. Optimization, N. Algorithms, Politecnico di milano, EU, 1992. Italy.
[42] S. Tabakhi, A. Najafi, R. Ranjbar, P. Moradi, Gene selection for microarray data classification using a novel ant colony optimization, Neurocomputing 168 (2015) 1024–1036.
[43] B.E. Boser, I.M. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, 1992, pp. 144–152.
[44] B. Scholkopf, S. Mika, C.J. Burges, P. Knirsch, K.-R. Muller, G. Ratsch, A.J. Smola, Input space versus feature space in kernel-based methods, IEEE Trans. Neural Network. 10 (5) (1999) 1000–1017.
[45] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learnin, 2009, p. 33. Cited on.
[46] K. Vembandasamy, R. Sasipriya, E. Deepa, Heart diseases detection using naive bayes algorithm, Int. J. Innov. Sci. Eng. Technol. 2 (9) (2015) 441–444.
[47] D.W. Aha, D. Kibler, M.K. Albert, Instance-based learning algorithms, Mach. Learn. 6 (1) (1991) 37–66.
[48] P. Mohapatra, S. Chakravarty, Modified pso based feature selection for microarray data classification, in: 2015 IEEE Power, Communication and Information Technology Conference (PCITC), IEEE, 2015, pp. 703–709.
[49] Y. Wang, I.H. Witten, Induction of Model Trees for Predicting Continuous Classes, 1996.
[50] J. Quinlan, C4. 5, programs for machine learning, in: Proc. Of 10th International Conference on Machine Learning, 1993, pp. 252–259.
[51] M. Daoud, M. Mayo, A survey of neural network-based cancer prediction models from microarray data, Artif. Intell. Med. 97 (2019) 204–214.
[52] C. Kang, Y. Huo, L. Xin, B. Tian, B. Yu, Feature selection and tumor classification for microarray data using relaxed lasso and generalized multi-class support vector machine, J. Theor. Biol. 463 (2019) 77–91.
[53] H. Zhu, N. Bi, J. Tan, D. Fan, An embedded method for feature selection using kernel parameter descent support vector machine, in: Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Springer, 2018, pp. 351–362.
[54] S. Maldonado, J. López, Dealing with high-dimensional class-imbalanced datasets: embedded feature selection for svm classification, Appl. Soft Comput. 67 (2018) 94–105.
[55] S. Mishra, D. Mishra, Svm-bt-rfe: an improved gene selection framework using bayesian t-test embedded in support vector machine (recursive feature elimination) algorithm, Karbala Int. J. Modern Sci. 1 (2) (2015) 86–96.
[56] L. Zhang, X. Huang, Multiple svm-rfe for multi-class gene selection on dna microarray data, in: 2015 International Joint Conference on Neural Networks (IJCNN), IEEE, 2015, pp. 1–6.
[57] I. Jain, V.K. Jain, R. Jain, An improved binary particle swarm optimization (ibpso) for gene selection and cancer classification using dna microarrays, in: 2018 Conference on Information and Communication Technology (CICT), IEEE, 2018, pp. 1–6.
[58] P. Moradi, M. Gholampour, A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy, Appl. Soft Comput. 43 (2016) 117–130.
[59] C. Garibay, G. Sanchez-Ante, L.E. Falcon-Morales, H. Sossa, Modified binary inertial particle swarm optimization for gene selection in dna microarray data, in: Mexican Conference on Pattern Recognition, Springer, 2015, pp. 271–281.
[60] K.-H. Chen, K.-J. Wang, K.-M. Wang, M.-A. Angelia, Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data, Appl. Soft Comput. 24 (2014) 773–780.
[61] N. Almugren, H. Alshamlan, Ff-svm: new firefly-based gene selection algorithm for microarray cancer classification, in: 2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE, 2019, pp. 1–6.
[62] P. Jinthanasatian, S. Auephanwiriyakul, N. Theera-Umpon, Microarray data classification using neuro-fuzzy classifier with firefly algorithm, in: 2017 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2017, pp. 1–6.
[63] T. Ragunthar, S. Selvakumar, A wrapper based feature selection in bone marrow plasma cell gene expression data, Cluster Comput. 22 (6) (2019) 13785–13796.
[64] M.S. Pratiwi, A. Aditsania, et al., Cancer detection based on microarray data classification using genetic bee colony (gbc) and conjugate gradient backpropagation with modified polak ribiere (mbp-cgp), in: 2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA), IEEE, 2018, pp. 163–168.
[65] M.A. Tawhid, A.M. Ibrahim, Feature selection based on rough set approach, wrapper approach, and binary whale optimization algorithm, Int. J. Machine Learn. Cybernet. 11 (3) (2020) 573–602.
[66] A. Zakeri, A. Hokmabadi, Efficient feature selection method using real-valued grasshopper optimization algorithm, Expert Syst. Appl. 119 (2019) 61–72.
[67] K. Chatra, V. Kuppili, D.R. Edla, A.K. Verma, Cancer data classification using binary bat optimization and extreme learning machine with a novel fitness function, Med. Biol. Eng. Comput. 57 (12) (2019) 2673–2682.
[68] M. Ghosh, S. Begum, R. Sarkar, D. Chakraborty, U. Maulik, Recursive memetic algorithm for gene selection in microarray data, Expert Syst. Appl. 116 (2019) 172–185.
[69] M. Allam, M. Nandhini, Optimal Feature Selection Using Binary Teaching Learning Based Optimization Algorithm, Journal of King Saud University-Computer and Information Sciences, 2018.
[70] A. Sharma, K.K. Paliwal, S. Imoto, S. Miyano, A feature selection method using improved regularized linear discriminant analysis, Mach. Vis. Appl. 25 (3) (2014) 775–786.
[71] M. Alweshah, S. Al Khalaileh, B.B. Gupta, A. Almomani, A.I. Hammouri, M.A. Al-Betar, The monarch butterfly optimization algorithm for solving feature selection problems, Neural Comput. Appl. (2020) 1–15.
[72] Y. Arshak, A. Eesa, A new dimensional reduction based on cuttlefish algorithm for human cancer gene expression, in: 2018 International Conference on Advanced Science and Engineering (ICOASE), IEEE, 2018, pp. 48–53.
[73] M.S.R. Nalluri, T. SaiSujana, K.H. Reddy, V. Swaminathan, An efficient feature selection using artificial fish swarm optimization and svm classifier, in: 2017 International Conference on Networks & Advances in Computational Technologies (NetACT), IEEE, 2017, pp. 407–411.
[74] K. Kavitha, A. Prakasan, P. Dhrishya, Score-based feature selection of gene expression data for cancer classification, in: 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), IEEE, 2020, pp. 261–266.
[75] W. Ke, C. Wu, Y. Wu, N.N. Xiong, A new filter feature selection based on criteria fusion for gene microarray data, IEEE Access 6 (2018) 61065–61076.
[76] A. Rouhi, H. Nezamabadi-pour, Filter-based feature selection for microarray data using improved binary gravitational search algorithm, in: 2018 3rd Conference on Swarm Intelligence and Evolutionary Computation (CSIEC), IEEE, 2018, pp. 1–6.
[77] B. Liao, Y. Jiang, W. Liang, W. Zhu, L. Cai, Z. Cao, Gene selection using locality sensitive laplacian score, IEEE ACM Trans. Comput. Biol. Bioinf 11 (6) (2014) 1146–1156.
[78] J. Tang, S. Zhou, A new approach for feature selection from microarray data based on mutual information, IEEE ACM Trans. Comput. Biol. Bioinf 13 (6) (2016) 1004–1015.
[79] K. Umamaheswari, M. Dhivya, D-mbpso, An unsupervised feature selection algorithm based on pso, in: Innovations in Bio-Inspired Computing and Applications, Springer, 2016, pp. 359–369.
[80] F. Al-Obeidat, A. Tubaishat, B. Shah, Z. Halim, et al., Gene encoder: a feature selection technique through unsupervised deep learning-based clustering for large gene expression data, Neural Comput. Appl. (2020) 1–23.
[81] M. Ghosh, S. Adhikary, K.K. Ghosh, A. Sardar, S. Begum, R. Sarkar, Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods, Med. Biol. Eng. Comput. 57 (1) (2019) 159–176.
[82] K. Yan, H. Lu, Evaluating ensemble learning impact on gene selection for automated cancer diagnosis, in: International Workshop on Health Intelligence, Springer, 2019, pp. 183–186.
[83] C. Peng, X. Wu, W. Yuan, X. Zhang, Y. Li, Mgrfe: multilayer recursive feature elimination based on an embedded genetic algorithm for cancer classification, IEEE ACM Trans. Comput. Biol. Bioinf 18 (2) (2019) 621–632, https://doi.org/10.1109/TCBB.2019.2921961.
[84] H. Lu, J. Chen, K. Yan, Q. Jin, Y. Xue, Z. Gao, A hybrid feature selection algorithm for gene expression data classification, Neurocomputing 256 (2017) 56–62.
[85] M. Dashtban, M. Balafar, Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts, Genomics 109 (2) (2017) 91–107.
[86] H. Salem, G. Attiya, N. El-Fishawy, Classification of human cancer diseases by gene expression profiles, Appl. Soft Comput. 50 (2017) 124–134.
[87] H. Djellali, S. Guessoum, N. Ghoualmi-Zine, S. Layachi, Fast correlation based filter combined with genetic algorithm and particle swarm on feature selection, in: 2017 5th International Conference on Electrical Engineering-Boumerdes (ICEE-B), IEEE, 2017, pp. 1–6.
[88] S. Thangavelu, S. Akshaya, K. Naetra, K.S. AC, V. Lasya, Feature selection in cancer genetics using hybrid soft computing, in: 2019 Third International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), IEEE, 2019, pp. 734–739.
[89] S. Begum, A.A. Ansari, S. Sultan, R. Dam, A hybrid model for optimum gene selection of microarray datasets, in: Recent Developments in Machine Learning and Data Analytics, Springer, 2019, pp. 423–430.
[90] I. Sangaiah, A.V.A. Kumar, Improving medical diagnosis performance using hybrid feature selection via relieff and entropy based genetic search (rf-ega) approach: application to breast cancer prediction, Cluster Comput. 22 (3) (2019) 6899–6906.
[91] X.-Y. Liu, Y. Liang, S. Wang, Z.-Y. Yang, H.-S. Ye, A hybrid genetic algorithm with wrapper-embedded approaches for feature selection, IEEE Access 6 (2018) 22863–22874.
[92] A.K. Shukla, P. Singh, M. Vardhan, A two-stage gene selection method for biomarker discovery from microarray data for cancer classification, Chemometr. Intell. Lab. Syst. 183 (2018) 47–58.
[93] A.K. Shukla, P. Singh, M. Vardhan, A hybrid gene selection method for microarray recognition, Biocybernet. Biomed. Eng. 38 (4) (2018) 975–991.
[94] K. Yan, H. Lu, An extended genetic algorithm based gene selection framework for cancer diagnosis, in: 2018 9th International Conference on Information Technology in Medicine and Education (ITME), IEEE, 2018, pp. 43–47.
[95] P. Singh, A. Shukla, M. Vardhan, Hybrid approach for gene selection and classification using filter and genetic algorithm, in: 2017 International Conference on Inventive Computing and Informatics (ICICI), IEEE, 2017, pp. 832–837.

[96] A. Hasnat, A.U. Molla, Feature selection in cancer microarray data using multi-objective genetic algorithm combined with correlation coefficient, in: 2016 International Conference on Emerging Technological Trends (ICETT), IEEE, 2016, pp. 1–6.

[97] H. Salem, G. Attiya, N. El-Fishawy, Gene expression profiles based human cancer diseases classification, in: 2015 11th International Computer Engineering Conference (ICENCO), IEEE, 2015, pp. 181–187.

[98] E. Bonilla-Huerta, A. Hernandez-Montiel, R. Morales-Caporal, M. Arjona-Lopez, Hybrid framework using multiple-filters and an embedded approach for an efficient selection and classification of microarray data, IEEE ACM Trans. Comput. Biol. Bioinf 13 (1) (2015) 12–26.

[99] R. Luque-Baena, D. Urda, M.G. Claros, L. Franco, J.M. Jerez, Robust gene signatures from microarray data using genetic algorithms enriched with biological pathway keywords, J. Biomed. Inf. 49 (2014) 32–44.

[100] X. Gao, X. Liu, A novel effective diagnosis model based on optimized least squares support machine for gene microarray, Appl. Soft Comput. 66 (2018) 50–59.

[101] M. Abinash, V. Vasudevan, A study on wrapper-based feature selection algorithm for leukemia dataset, in: Intelligent Engineering Informatics, Springer, 2018, pp. 311–321.

[102] R. Alzubi, N. Ramzan, H. Alzoubi, A. Amira, A hybrid feature selection method for complex diseases snps, IEEE Access 6 (2017) 1292–1301.

[103] H.M. Alshamlan, Co-abc: correlation artificial bee colony algorithm for biomarker gene discovery using gene expression profile, Saudi J. Biol. Sci. 25 (5) (2018) 895–903.

[104] R.A. Musheer, C. Verma, N. Srivastava, Novel machine learning approach for classification of high-dimensional microarray data, Soft Computing 23 (24) (2019) 13409–13421.

[105] A. Alkuhlani, M. Nassef, I. Farag, Multistage feature selection approach for high-dimensional cancer data, Soft Computing 21 (22) (2017) 6895–6906.

[106] A.K. Shukla, P. Singh, M. Vardhan, Gene selection for cancer types classification using novel hybrid metaheuristics approach, Swarm Evolut. Comput. 54 (2020) 100661.

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

[108] X.H. Han, D.A. Li, L. Wang, A hybrid cancer classification model based recursive binary gravitational search algorithm in microarray data, Procedia Computer Science 154 (2019) 274–282.

[109] S.A.A. Vijay, P. GaneshKumar, Fuzzy expert system based on a novel hybrid stem cell (hsc) algorithm for classification of micro array data, J. Med. Syst. 42 (4) (2018) 1–12.

[110] F.V. Sharbaf, S. Mosafer, M.H. Moattar, A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization, Genomics 107 (6) (2016) 231–238.

[111] I. Jain, V.K. Jain, R. Jain, Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification, Appl. Soft Comput. 62 (2018) 203–215.

[112] A. Chinnaswamy, R. Srinivasan, Hybrid feature selection using correlation coefficient and particle swarm optimization on microarray gene expression data, in: Innovations in Bio-Inspired Computing and Applications, Springer, 2016, pp. 229–239.

[113] M. Liu, L. Xu, J. Yi, J. Huang, A feature gene selection method based on relieff and pso, in: 2018 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), IEEE, 2018, pp. 298–301.

[114] M. Mollaee, M.H. Moattar, A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification, Biocybernet. Biomed. Eng. 36 (3) (2016) 521–529.

[115] W. Zhou, J.A. Dickerson, A novel class dependent feature selection method for cancer biomarker discovery, Comput. Biol. Med. 47 (2014) 66–75.

[116] E. Pashaei, M. Ozen, N. Aydin, Gene selection and classification approach for microarray data based on random forest ranking and bbha, in: 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), IEEE, 2016, pp. 308–311.

[117] A. Dabba, A. Tari, S. Meftali, Hybridization of moth flame optimization algorithm and quantum computing for gene selection in microarray data, J. Ambi. Intell. Humanized Comput. (2020) 1–20.

[118] S.K. Baliarsingh, S. Vipsita, B. Dash, A new optimal gene selection approach for cancer classification using enhanced jaya-based forest optimization algorithm, Neural Comput. Appl. 32 (12) (2020) 8599–8616.

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

[120] E. Bonilla-Huerta, R. Morales-Caporal, M.A. Arjona-López, Exploration and exploitation of high dimensional biological datasets using a wrapper approach based on strawberry plant algorithm, in: International Conference on Intelligent Computing, Springer, 2018, pp. 307–317.

[121] O.A. Alomari, A.T. Khader, M.A. Al-Betar, M.A. Awadallah, A novel gene selection method using modified mrmr and hybrid bat-inspired algorithm with β-hill climbing, Appl. Intell. 48 (11) (2018) 4429–4447.

[122] M.M. Mufassirin, R.G. Ragel, A novel filter-wrapper based feature selection approach for cancer data classification, in: 2018 IEEE International Conference on Information and Automation for Sustainability (ICIAfS), IEEE, 2018, pp. 1–6.

[123] R. Alanni, J. Hou, H. Azzawi, Y. Xiang, New gene selection method using gene expression programing approach on microarray data sets, in: International Conference on Computer and Information Science, Springer, 2018, pp. 17–31.

[124] O.A. Alomari, A.T. Khader, M.A. Al-Betar, Z.A.A. Alyasseri, A hybrid filter-wrapper gene selection method for cancer classification, in: 2018 2nd International Conference on BioSignal Analysis, Processing and Systems (ICBAPS), IEEE, 2018, pp. 113–118.

[125] Z.Y. Algamal, M.H. Lee, A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification, Adv. Anal. Class. 13 (3) (2019) 753–771.

[126] A.B. Brahim, M. Limam, A hybrid feature selection method based on instance learning and cooperative subset search, Pattern Recogn. Lett. 69 (2016) 28–34.

[127] B. Seijo-Pardo, V. Bolón-Canedo, A. Alonso-Betanzos, Using data complexity measures for thresholding in feature selection rankers, in: Conference of the Spanish Association for Artificial Intelligence, Springer, 2016, pp. 121–131.

[128] V. Bolón-Canedo, N. Sánchez-Marono, A. Alonso-Betanzos, Data classification using an ensemble of filters, Neurocomputing 135 (2014) 13–20.

[129] B. Seijo-Pardo, V. Bolón-Canedo, A. Alonso-Betanzos, On developing an automatic threshold applied to feature selection ensembles, Inf. Fusion 45 (2019) 227–245.

[130] M. Abdulla, M.T. Khasawneh, G.- forest, An ensemble method for cost-sensitive feature selection in gene expression microarrays, Artif. Intell. Med. 108 (2020) 101941.

[131] S. Sayed, M. Nassef, A. Badr, I. Farag, A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets, Expert Syst. Appl. 121 (2019) 233–243.

[132] A.B. Brahim, M. Limam, Ensemble feature selection for high dimensional data: a new method and a comparative study, Adv. Data Anal. Class. 12 (4) (2018) 937–952.

[133] A.K. Shukla, P. Singh, M. Vardhan, Dna gene expression analysis on diffuse large b-cell lymphoma (dlbcl) based on filter selection method with supervised classification method, in: Computational Intelligence in Data Mining, Springer, 2019, pp. 783–792.

[134] B. Pes, Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains, Neural Comput. Appl. (2019) 1–23.

[135] N.A. Firdausanti, et al., On the comparison of crazy particle swarm optimization and advanced binary ant colony optimization for feature selection on high-dimensional data, Procedia Computer Science 161 (2019) 638–646.

[136] S. Liang, A. Ma, S. Yang, Y. Wang, Q. Ma, A review of matched-pairs feature selection methods for gene expression data analysis, Comput. Struct. Biotechnol. J. 16 (2018) 88–97.

[137] V. Bolón-Canedo, D. Rego-Fernández, D. Peteiro-Barral, A. Alonso-Betanzos, B. Guijarro-Berdiñas, N. Sánchez-Maroño, On the scalability of feature selection methods on high-dimensional data, Knowl. Inf. Syst. 56 (2) (2018) 395–442.

[138] B. Seijo-Pardo, V. Bolón-Canedo, A. Alonso-Betanzos, Testing different ensemble configurations for feature selection, Neural Process. Lett. 46 (3) (2017) 857–880.

[139] C.A. Kumar, M. Sooraj, S. Ramakrishnan, A comparative performance evaluation of supervised feature selection algorithms on microarray datasets, Procedia Comput. Sci. 115 (2017) 209–217.

[140] C. Fahy, S. Ahmadi, A. Casey, A comparative analysis of ranking methods in a hybrid filter-wrapper model for feature selection in dna microarrays, in: International Conference on Innovative Techniques and Applications of Artificial Intelligence, Springer, 2015, pp. 387–392.

[141] P. Drotár, J. Gazda, Z. Smékal, An experimental comparison of feature selection methods on two-class biomedical datasets, Comput. Biol. Med. 66 (2015) 1–10.

[142] S. Akila, S.A. Christe, An experimental analysis of gene feature selection and classification methods for cancer microarray, in: Applied Computer Vision and Image Processing, Springer, 2020, pp. 204–211.

[143] B. Chandra, Gene selection methods for microarray data, in: Applied Computing in Medicine and Health, Elsevier, 2016, pp. 45–78.

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

[145] M. Amrane, S. Oukid, T. Ensari, N. Benblidia, Z. Orman, Microarray lung cancer data classification using similarity based feature selection, in: 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), Ieee, 2019, pp. 1–4.

[146] S.H. Bouazza, K. Auhmani, A. Zeroual, N. Hamdi, Selecting significant marker genes from microarray data by filter approach for cancer diagnosis, Procedia Computer Science 127 (2018) 300–309.

[147] P. Singh, A. Shukla, M. Vardhan, A novel filter approach for efficient selection and small round blue-cell tumor cancer detection using microarray gene expression data, in: 2017 International Conference on Inventive Computing and Informatics (ICICI), IEEE, 2017, pp. 827–831.

[148] M. Babu, K. Sarkar, A comparative study of gene selection methods for cancer classification using microarray data, in: 2016 Second International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), IEEE, 2016, pp. 204–211.

[149] T.M. Fahrudin, I. Syarif, A.R. Barakbah, Ant colony algorithm for feature selection on microarray datasets, in: 2016 International Electronics Symposium (IES), IEEE, 2016, pp. 351–356.

[150] B. Srivastava, R. Srivastava, M. Jangid, Filter vs. wrapper approach for optimum gene selection of high dimensional gene expression dataset: an analysis with cancer datasets, in: 2014 International Conference on High Performance Computing and Applications (ICHPCA), IEEE, 2014, pp. 1–6.

[151] S.P. Potharaju, M. Sreedevi, Distributed feature selection (dfs) strategy for microarray gene expression data to improve the classification performance, Clin. Epidemiol. Global Health 7 (2) (2019) 171–176.

[152] M.K. Ebrahimpour, M. Eftekhari, Distributed feature selection: a hesitant fuzzy correlation concept for microarray high-dimensional datasets, Chemometr. Intell. Lab. Syst. 173 (2018) 51–64.
[153] M.R. Alhamidi, D.M. Arsa, M.F. Rachmadi, W. Jatmiko, 2-dimensional homogeneous distributed ensemble feature selection, in: 2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS), IEEE, 2018, pp. 367–372.
[154] L. Morán-Fernández, V. Bolón-Canedo, A. Alonso-Betanzos, Centralized vs. distributed feature selection methods based on data complexity measures, Knowl. Base Syst. 117 (2017) 27–45.
[155] L. Venkataramana, S.G. Jacob, R. Ramadoss, D. Saisuma, D. Haritha, K. Manoja, Improving classification accuracy of cancer types using parallel hybrid feature selection on microarray gene expression data, Genes Genom. 41 (11) (2019) 1301–1313.
[156] D. Kečo, A. Subasi, J. Kevric, Cloud computing-based parallel genetic algorithm for gene selection in cancer classification, Neural Comput. Appl. 30 (5) (2018) 1601–1610.
[157] R.B. Ray, M. Kumar, S.K. Rath, Fast in-memory cluster computing of sizeable microarray using spark, in: 2016 International Conference on Recent Trends in Information Technology (ICRTIT), IEEE, 2016, pp. 1–6.
[158] A. Boucheham, M. Batouche, Massively parallel feature selection based on ensemble of filters and multiple robust consensus functions for cancer gene identification, in: Science and Information Conference, Springer, 2014, pp. 93–108.
[159] R.B. Ray, M. Kumar, S.K. Rath, Fast computing of microarray data using resilient distributed dataset of Apache spark, in: Recent Advances in Information and Communication Technology 2016, Springer, 2016, pp. 171–182.
[160] L. Venkataramana, S.G. Jacob, R. Ramadoss, A parallel multilevel feature selection algorithm for improved cancer classification, J. Parallel Distr. Comput. 138 (2020) 78–98.
[161] Y. Lokeswari, S.G. Jacob, Prediction of child tumours from microarray gene expression data through parallel gene selection and classification on spark, in: Computational Intelligence in Data Mining, Springer, 2017, pp. 651–661.
[162] M. Kumar, N.K. Rath, S.K. Rath, Analysis of microarray leukemia data using an efficient mapreduce-based k-nearest-neighbor classifier, J. Biomed. Inf. 60 (2016) 395–409.
[163] A. Kourid, M. Batouche, Biomarker discovery based on large-scale feature selection and mapreduce, in: IFIP International Conference on Computer Science and its Applications, Springer, 2015, pp. 81–92.
[164] M. Kumar, N.K. Rath, A. Swain, S.K. Rath, Feature selection and classification of microarray data using mapreduce based anova and k-nearest neighbor, Procedia Computer Science 54 (2015) 301–310.
[165] M. Kumar, S.K. Rath, Classification of microarray using mapreduce based proximal support vector machine classifier, Knowl. Base Syst. 89 (2015) 584–602.

Esra'a Alhenawi (esra_a_2008@live.com) is currently a PhD candidate, The University of Jordan, King Abdullah II School for Information Technology, Department of Computer Science.

Rizik Al-Sayyed (r.alsayyed@ju.edu.jo), is a professor in computer networks, cloud computing, databases systems, and simulation, The University of Jordan, King Abdullah II School for Information Technology, Department of Information Technology.

Amjad Hudaib (ahudaib@ju.edu.jo), is a professor in Software Engineering, The University of Jordan, King Abdullah II School for Information Technology, Department of Computer Information Systems.

Seyedali Mirjalili (ali.mirjalili@gmail.com), is currently an Associate Professor and the director of the Centre for Artificial Intelligence Research and Optimization at Torrens University Australia. He is internationally well recognized in Swarm Intelligence and Optimization.