

A Review of Feature Selection Methods in Big Data

Hiba ALMarwi^{*1}, Ghaleb H. AL-Gaphari²

Submitted:27/03/2024 Revised: 09/05/2024 Accepted: 20/05/2024

Abstract: The exponential growth of available data has led to the emergence of high dimensional datasets characterized by hundreds of variables. Consequently, feature selection has emerged as a critical task in data mining and machine learning. Its primary objective is to mitigate the curse of dimensionality by reducing data dimensionality and enhancing algorithm performance, particularly in classification tasks. This paper aims to provide the reader with a comprehensive understanding of feature selection techniques beyond the traditional categorization into filter, wrapper, and embedded approaches. Instead, we present feature selection as a combinatorial optimization or search problem, leading to a novel categorization of methods into exhaustive search, heuristic search, metaheuristic search, embedded methods, and hybrid approaches. By examining the strengths and weaknesses of these methods, we shed light on their strengths and weaknesses. Furthermore, we address current problems and challenges in the field, paving the way for future research and identifying promising areas of investigation.

Keywords: data dimensionality reduction, evolutionary algorithms, feature selection, filter methods, high dimensional data, machine learning.

1. Introduction

Data are experiencing a rapid increase not only in terms of the abundance of patterns (data instances or tuples), but also the dimensionality of features (or data attributes). The term “big data” is used to deal with datasets that are so large and complex that it becomes very difficult to process them by using traditional data processing applications [1, 2]. Many of these big datasets contain misleading and redundant features that can confuse the classification process and significantly increase the time required for processing datasets during an exhaustive search in the solution space. Sometimes, the overwhelming volume of data even makes the data mining algorithms completely useless. The situation is particularly acute in bioinformatics [3-9]. Langley et al in [10, 11] pointed out that the predictive accuracy of the learning algorithms is reduced in the presence of irrelevant features. Fu and Wang [12] showed that deleting those irrelevant features can not only improve the classification accuracy but also reduce the structural complexity of the radial basis function (RBF) within a neural network to facilitate rule extraction. It is often advantageous to select a suitable subset of features that are relevant to target prediction for improving the prediction performance.

Microarray data consist of a number of genes, but in most cases, only a fraction of genes is important. Therefore, gene selection is necessarily required to identify the important genes which help classify samples effectively [13]. Microarray data pose a great challenge for computational

techniques because of their large dimensionality (up to several tens of thousands of genes) and the various gene expression scenarios, as well as the difficulties of validating distributional assumptions because of small sample sizes. Many nonparametric methods have been widely proposed as an attractive alternative method to improve distribution assumptions [14]. Classification of microarray data plays a significant role in the diagnosis and prediction of many microarray data. However, the high dimensionality of the data compared to the number of observations may result in poor classification accuracy.

As many machine learning algorithms were not originally designed to cope with large amounts of irrelevant features, it becomes necessary for those algorithms to be accompanied by a feature selection process [15, 16]. The aim of feature selection FS is to choose a subset of the original features that have a major influence on the classification method to achieve the best feature set according to a predetermined objective [17, 18]. It is important to note that the best feature set may simply be the original features in the absence of redundant or irrelevant features. However, in the presence of redundant or irrelevant features, there can be multiple subsets of features that are equally effective for a given objective. Consequently, FS may not necessarily result in a unique subset of features. Therefore, identifying the core representative features that contribute to the best prediction is a challenging task.

FS techniques have the advantage of enhancing accuracy, avoiding challenges, such as overfitting, and low interpretability of the final model, decreasing model complexity, and reducing training time. [19, 20] However, the advantages of feature selection techniques come at a

¹ Computer Science Department, Sanaa University, Sanaa, AL. Yemen
ORCID ID : 0009-0001-7811-5129

² Computer Science Department, Sanaa University, Sanaa, AL. Yemen
ORCID ID : 0000-0002-5145-3144

* Corresponding Author Email: Hebh.Almarwi@gmail.com

certain price, as the search for a subset of relevant features introduces an additional layer of complexity in the modelling task. Instead of just optimizing the parameters of the model for the full feature subset, we now need to find the optimal model parameters for the optimal feature subset, as there is no guarantee that the optimal parameters for the full feature set are equally optimal for the optimal feature subset [21]. As a result, the search in the model hypothesis space is augmented by another dimension: the one of finding the optimal subset of relevant features. During the last decade, FS techniques have shifted from being an illustrative example to becoming a real prerequisite for model building. Feature selection has been proven to be an effective way to improve the results of many classification algorithms [12, 18, 22-24].

The main aim of this paper is to introduce a basic taxonomy of feature selection techniques, provide guideline for the research community to figure out feature selection challenges, and highlight a new direction of the research in the context of big data feature selection. A summary of existing articles on feature selection techniques is shown in Table 1.

2. FEATURE SELECTION TECHNIQUES

FS techniques differ from each other in the way they incorporate the search for feature subsets in the feature space with the model hypothesis space. Based on this assumption, we can classify feature selection techniques into:

2.1. Filter Methods:

Filter based techniques have been used by many researchers to identify very few relevant features and eliminate redundant ones. These techniques measure the relevance of features by examining the intrinsic characteristics of the data and their relevance with dependent variables. They select features that have high score value according to the ranking criterion function [20, 25]. Several publications [26] have presented various definitions and measurements for the relevance of a feature. One definition is that “A feature can be regarded as irrelevant if it is conditionally independent of the class labels.” [27]. Another definition states that “if a feature is to be relevant, it can be independent of the input data but cannot be independent of the class labels.”

Filtering methods can be easily applied to large datasets as they are computationally less rigorous and faster than other competing methods. Although all filtering methods have the same mechanism, each one of them uses different ways to rank features, such as score function and classification results. Research work given in [28] proposes a measure named “relative importance” to evaluate and recognize useful features depending on how much they contributed in the development of classifying data points while

constructing many classification trees. Top ranked features usually correlate features and tend to get similar rankings. From a statistical point of view, these correlated features are redundant. To select the top nonredundant features and minimize the redundancy between sequentially selected features, Ding et. Al in [29, 30] used an algorithm called minimal redundancy maximal relevance (mRMR). This algorithm exploited a sequence of instinctive measures of relevance to the class concept and minimize the redundancy between sequentially selected features. However, this method used the greedy search, thus the global feature redundancy was not considered and the results were not optimal. To address this problem, research in [31] used the famous GRM algorithm to globally minimize the feature redundancy while maximizing the given feature ranking scores. Comparative studies comparing several algorithms for identifying important feature in large datasets are being conducted in [32]. Additionally, there are some excellent survey papers related to filter techniques for FS in gene expression microarray analysis. Researchers in [33, 34] introduced gene selection filter methods to select important genes for disease prediction. In [35, 36], filter methods are studied on large datasets, analyzing the predictive accuracy with respect to the number of features to select. In [37, 38], filter methods are compared with respect to classification accuracy based on microarray datasets. In [39, 40], several filter methods that are based on mutual information are compared. In [39], the accuracy and the run time of the methods are analyzed separately. The authors also examine the proportions of correctly detected features on artificial data and consider theoretical properties. In [40], the authors analyze the classification accuracy with respect to the number of selected features and search for pareto optimal methods considering the accuracy and feature selection stability. In [41], an extensive comparison of 22 filter methods on 16 large or high dimensional data sets is conducted with respect to both classification accuracy and run time jointly.

Filter methods can be also categorized according to the type of output into: feature subset selection and feature ranking [42]. Feature subset selection methods generate a subset of features that collectively have good predictive capability. They are based on the assumption that a given feature may have better predictive power when combined with some other attributes, compared to when used alone. Feature subset selection provides one optimal subset of features, and thus users need not to consider the number of features to be selected (i. e., cutoff). On the other hand, feature ranking methods rank features by scoring each feature according to a particular method, then selecting features based on their scores. Feature ranking provides the rank of features and this requires the user’s decision about the cutoff, but this enables feature ranking to have multiple choices for constructing a prediction model depending on the cutoff

instead. Usually, feature subset selection has a higher time complexity than feature ranking because searching through feature subsets in the subset generation step is required $O(2^d)$. Although there exist various heuristic search strategies such as greedy sequential search, best first search, and genetic algorithm, most of them still incur $O(d^2)$, which prevents them from scaling well to datasets containing tens of thousands of features. Filter methods can be further classified into:

2.1.1. Univariate:

Due to the burden of time and memory complexity, univariate feature ranking methods are widely used in feature selection [43]. However, most of them are not that accurate because they only consider the relevance of features to the target without considering the redundancy among features. Each feature is evaluated separately, thereby ignoring feature dependencies. Examples of the univariate filter method include mutual information [44-46], information gain [44, 45, 47, 48], Pearson correlation coefficient [49], relative entropy [48], and Fisher score [20]. The idea of this filter method is to remove features that only consist of noise and therefore have very little variation. This filter only makes sense for datasets where the features are measured on the same scale and have not been scaled to unit variance. These methods aim to keep features that have a strong linear association with the target variable. However, univariate filtering may overlook features that are relevant only when interacting with other features. Moreover, the number of overlooked but important features may be even higher when only top-ranked features are selected. Fisher score is a filter-based supervised feature selection method with feature weights [50]. High-quality features in the same class should have the same value and different value in different classes. Based on this intuition, the score for the i -th feature S_i will be calculated by the Fisher Score as:

$$S_i = \frac{\sum_{k=1}^K n_j (\mu_{ij} - \mu_i)^2}{\sum_{k=1}^K n_j p_{ij}^2} \quad (1)$$

where μ_{ij} and p_{ij} are the mean and the variance of the i -th feature in the j -th class, respectively, n_j is the number of instances in the j -th class, and μ_i is the mean of the i -th feature. Fisher score models have various benefits [51], including fewer calculations and greater accuracy, which effectively reduce time-space complexity. Fisher score was used by Guyon et al [52] for gene classification. Fisher score based on binary classification was utilized by Chen and Lin [53] to pick features with high scores, but the characteristics between features were not taken into account. ReliefF and Fisher score was combined by Hancer in [54] to select the highest scoring features while considering the correlation between features and labels. Despite the fact that the fact that these methods could successfully remove irrelevant features, some redundant features were still selected. Therefore, the maximum information coefficient with the

Fisher score was combined in [51]. Mutual information (MI) is another univariate filter method. The mutual information (MI) is information theoretic ranking criteria to measure how much knowledge one feature has about another feature [55]. This term is helpful for feature selection since it provides a mechanism to measure the significance of a feature subset with respect to the output vector C . Formally, the MI is defined as follows:

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

In [56], the author develops a feature ranking criteria based on conditional mutual information for binary data. By using a conditional MI, two variables' information can be measured in relation to a third one, which is defined as follows:

$$I(x; y|z) = \sum_{i=1}^n p(z(i)) I(x; y|z = z(i)) \quad (3)$$

Although conditional MI is a good measure, it does not measure the information among the three features. Multiinformation is an interesting extension of MI, proposed by McGill [57], which allows measuring the interaction among more than two variables. In order to overcome univariate problems, a number of multivariate filter techniques were introduced.

2.1.2. Multivariate:

Multivariate filtering methods employ two measures in order to obtain a small set of nonredundant features. The first measure assesses the relationship between the features and the decision variable, whereas the second measure evaluates the interdependencies among predictor variables within the feature set. The RELIEF algorithm [58-60] is a filter-based method that ranks the features according to a feature relevance criterion. A threshold is used to choose a subset of the features. However, one of the drawbacks of the RELIEF algorithm is the challenge of selecting a proper threshold and the possibility of obtaining a subset that is redundant rather than optimal. This problem of redundant vs. relevant variables is addressed in [16]. Important features that are informative when paired with others but are less informative on their own could be disregarded in feature ranking. The author in [61] used the improved ReliefF algorithm to effectively eliminate irrelevant and redundant genes. The irrelevant genes were filtered using the threshold, and the optimal gene subset was determined through an iterative improvement process. Finally, the gene was classified and recognized using the classical classification algorithm.

2.2 Wrapper Methods

Wrapper methods rely on the predictive performance of a predefined learning algorithm to assess the quality of selected features. For each of the subsets of the set of all features, a predictive model is fitted, and the subsets are

evaluated. The evaluation of a specific subset of features is obtained by training and testing a specific classification model [62]. Some examples of wrapper methods include recursive feature elimination [63], sequential feature selection algorithms [64], and genetic algorithms. A classic wrapper method involves two steps: first, the feature set search component creates a subset of features. In the first phase, different search algorithms may be used, such as complete search methods like branch and bound [65] or incomplete search methods like local search methods, greedy search, or metaheuristics [66, 67]. Then the learning algorithm acts as a black box to assess the quality of these features based on the learning performance. Second, the feature subset with the highest learning performance is returned as the selected features. Researchers in [6, 68, 69] described several intelligent techniques for identifying single nucleotide polymorphisms (SNP) from large microarray dataset. Wrapper methods such as stepwise regression iteratively perform feature selection in the direction of improving the quality criteria of a particular model, such as accuracy and error rate [70], and may include heuristic search methods, such as genetic algorithms and particle swarm optimization, to find an optimal subset of features. Among the large number of classifiers available, it will be sensible to choose a good classifier, such as random forests [71], support vector machines (SVMs) [72]. In [73], filter and wrapper methods are compared with respect to the correctness of the selected features. Additionally, the authors in [74] conduct comparisons with respect to classification accuracy on real datasets. The authors in [75, 76] conducted a comprehensive comparison with respect to classification accuracy and run time, considering each of the two objectives separately. The results generally indicate that wrapper methods provide better accuracy, but they are computationally more expensive than filter methods. In addition, many classifiers are prone to overlearning and show sensitiveness to initialization. Wrapper methods can be classified based on the search strategies into:

2.2.1 Exhaustive Search:

Searching for the truly optimal subset(s) of features is usually computationally expensive and has been shown to be NP-hard. [77]. Hence, in most practical situations, exhaustively searching for the truly optimal subset(s) of features cannot be accomplished. Basically, in order to find the very best feature subset(s), one would need to exhaustively try out all possible M feature combinations of the N original features, with $M = 1; 2; \dots N$. This “combinatorial explosion” for an exhaustive search leads to a computational load that grows exponentially as the total number of features increases. In practical terms, this becomes impossible even for the most powerful computers if there are more than 30 features to be searched. Thus, exhaustive search is used in a limited number of research studies [78-81].

Wang et al [80] attempted at finding the smallest set of genes that can ensure highly accurate classification of cancers from microarray data. Their simple yet effective approach consists of the following two steps: In the first step, some important genes are chosen using a feature importance ranking measure. In the second step, an exhaustive search is carried out within the top-ranked genes, evaluating the classification capability of each simple (i.e., 1-gene, 2-gene, 3-gene) combination of those important genes using a good classifier. This approach leads to very high accuracy and can significantly reduce the number of genes required for highly reliable diagnosis. To determine the most informative traits, a rigorous multidimensional technique was developed by [82]. This method analyzes exhaustively the relationship between the descriptor and decision variables.

2.2.2 Heuristic search:

As the space of feature subsets grows exponentially with the number of features, and to avoid an exhaustive search, different heuristic search methods are used to guide the search for an optimal subset. In general, the problem-dependent stochastic techniques are called heuristic techniques. These methods constantly seek to get the full benefits of the nature of the problem, which makes them extremely greedy. Sometimes, because of this greed, these methods are trapped in local optima, and therefore, the global optimal solution is not reach. Several heuristic algorithms have been used to get an optimal set of features, including but not limited to greedy hill climbing algorithm [83, 84], branch and bound method, beam search, and best first algorithm. The greedy hill climbing algorithm aims to find the best solution by making local changes to the current solution and continuing this process until no better solution is found. SFS (Sequential Forward Selection) [85, 86] and SBS (Sequential Backward Selection) [87, 88] are two kinds of hill climbing methods [89]. The forward search method starts with an empty set of features and, at each step of the algorithm, adds one of the informative features to increase the classifier performance, whereas SBS starts with the full set of features and greedily removes a feature at each step according to the classifier performance. However, these algorithms not only require high computation cost but are also likely to fall into a local optimum. The greediness of this method may lead to low-accurate solutions, and in the case of highly correlated data sets, a redundant subset of features may be selected. Nesting effect is the main drawback of those algorithms, which means that while a change is considered positive, there is no chance of re-evaluating that feature. To reduce the effect of the nesting problem, the “plus l-take away-r” method was proposed [90], but determining the correct values between l and r remains the main drawback of this method. In order to prevent this problem, SFFS (Sequential Forward Floating Selection) and SBFS (Sequential Backward Floating Selection) are proposed.

Selection) were proposed, aiming to overcome the nesting effect problem [89].

2.2.3 Meta-Heuristic search:

Metaheuristics refer to stochastic optimization techniques that improve the performance of heuristic techniques. These techniques are able to examine the search space extensively and can therefore escape from local optima. The metaheuristic optimization algorithm is indispensable in the wrapper algorithm [91, 92] for the following reasons: they can apply the program easily, escape the local best, and obtain global optimal value. These search methods can be divided into two classes: deterministic and randomized search algorithms. In practice, well-designed heuristic search is likely to outperform random search, although this cannot be guaranteed since, in some cases, random search could quickly stumble on a global optimum by chance. The actual performance of a metaheuristics search depends largely on the design of the heuristic function as well as the problem at hand. Furthermore, good heuristic search algorithms should find a global optimum in a reasonable amount of time. This time decrease is caused by the fact that metaheuristic approaches are population-based methods, where the solution is determined by the size of the population and a number of generations. In the last few decades, more than a hundred nature-inspired metaheuristic algorithms have been designed, including simulated annealing [93, 94], genetic algorithm [81, 95-97], ant colony optimization (ACO)[98], particle swarm optimization (PSO)[99], bat algorithm (BA)[100, 101], cuckoo Search (CS)[102-104], and whale Optimization Algorithm (WOA) [92, 105]. Several researchers have tried to solve the feature selection problem by means of swarm algorithms [106, 107]. For example, Abd Elaziz and et al in 2020 proposed the opposition-based moth flame optimization to generate an optimal initial population in the differential evolution algorithm for feature selection [108]. Based on the algorithm's behavior, metaheuristic search algorithms can be further classified into the evolutionary algorithm and the swarm algorithm. Among existing evolutionary algorithms, teaching learning-based optimization (TLBO) has been recently proposed for solving feature selection problem [109]. The TLBO algorithm is a human-based approach that works on the teaching and learning phenomena where a teacher influences the output of the learners in a class. To improve the performance of TLBO, researchers in [110] proposed a neighbor teaching learning-based optimization (NTLBO). The experimental results demonstrate that TLBO outperformed the existing metaheuristic methods, despite getting stuck in local optima when they resolve complex numerical problems. Recently, a new physician algorithm called the gravitational search algorithm (GSA) was introduced. GSA received much attention in research because of its superior global search capabilities and faster convergence in bioinformatics domains. Moreover, the

leading advantages of GSA are less computational effort and very few control parameters compared to other existing evolutionary algorithms (EAs). In [111, 112], the authors proposed a new strategy method called FSS-MGSA for feature subset selection, which has been successfully applied to high-dimensional search space. Swarm algorithms are among the most popular optimization techniques used to improve the performance of feature selection. One of the nature-inspired algorithm is the binary ABC method, which utilizes the benefits of bees to resolve the gene selection problem. The differential evolution algorithm is another nature-inspired algorithm used to solve feature selection problems. For example, researchers in [113] proposed a new self-adaptive binary variant of a differential evolution algorithm for solving the discrete problems. Researchers in [114] proposed a feature selection approach based on adaptive particle swarm optimization (ABPSO). The logistic equation of a chaotic system was used to initialize the particle swarm. This value dynamically adjusts the level of convergence of particles and the diversity within the particle swarm. Based on the objective function, heuristic search algorithms can be further classified into:

2.2.2.1 Single objective:

Over the past few years, most of the natural inspired algorithms mentioned in the above section have been used as efficient methods for feature selection problem. However, they have been used as a single objective, either to reduce the number of features or maximize classification accuracy. Recently, multiobjective optimization problems (MOOPs) of these techniques have been adapted to solve the problem of feature selection.

2.2.2.2 Multiobjective:

In the last few years, multiobjective optimization problems (MOOPs) have attracted much attention since real-world problems have two conflicting goals, namely, maximizing the performance of classification and minimizing the number of features. Unlike a single objective optimization problem (SOOP), in which there is only one optimal point, there exists a set of solutions in a typical MOOP. In MOOP, the solution is superior to the rest of the solutions in the search space when all objectives are considered, but it may be inferior to other solutions in the space in one or more objectives. Gumus et al [115] used a multi-objective optimization technique called pareto optimal for selecting SNP subsets that offer both high classification accuracy and correlation between genomic and geographical distances. The discriminatory power of an SNP was determined using mutual information. They demonstrated their method using the Human Genome Diversity Project (HGDP) SNP dataset. [116] proposed the multiobjective PSO method, which ranks features based on their frequency in a set of archives. The authors in [117] proposed a hybrid mutation operator

for an improved multiobjective PSO. The experimental findings indicate that the selected features provide better efficiency measurement compared to standard PSO and other related algorithms. [118] proposed a multiobjective approach based on the bat algorithm for gene selection and tumor classification. [119] proposed a binary multiobjective method based on the Grey Wolf algorithm for cervix lesion classification. MPSO and NSGA-II are used in [120] as multiobjective feature selection method to predict the dose of warfarin. Mutation operator and local search were utilized to enhance PSO. The experiments demonstrated that the proposed algorithm performs better than another traditional method. Authors in [121] identified cancer microRNA makers by incorporating SVM classifier with a multiobjective genetic algorithm approach. The experimental findings demonstrated the efficiency of the proposed method compared to state-of-the-art methods in terms of classification performance. The authors in [122] proposed two multiobjective frameworks based on the NSGAII and SPESA2 evolutionary algorithms for feature selection. The proposed frameworks were evaluated, and the findings indicate that the proposed method outperformed the state-of-the-art method in terms of increasing classification performance and reducing the number of features. The experiment demonstrates that in multiobjective methods, the number of features is always less than in a single objective method, which means that it has efficient capability to discover the search space. Multiobjective algorithms generate a solution that contains a reduced number of features and attains high classification performance.

2.3 Embedded Methods:

Embedded methods use internal information of the classification model to identify the features that have a great impact on the accuracy of the model during the model fitting process. The search for an optimal subset of features is built into the classifier construction and can be seen as a search in the combined space of feature subsets and hypotheses. Several researchers have tried to solve the feature selection problem using embedded methods. For example, researchers in [52] proposed a support vector machine method based on recursive feature elimination (RFE). [123] used SVM-RFE to select features for predicting essential proteins and remove the features that share biological meaning with other features according to Pearson Correlation Coefficients (PCC). To address the problem that the SVM-RFE is highly sensitive to the filter-out factor, i. e. , how many genes are removed at one step, Tang et al [124] proposed a two-stage SVM-RFE algorithm for microarray gene selection. The algorithm effectively eliminated most of the irrelevant, redundant, and noisy genes while keeping information loss small at the first stage. A fine selection for the final gene subset was then performed at the second stage. The two-stage SVM-RFE overcome the instability problem

of the SVM-RFE and achieve improved results. Pang et al [125] used the random forests to identify a set of prognostic genes. They showed that their method incorporates multivariate correlations and is advantageous over single gene-based approaches. These embedded FS methods are suboptimal since training of the objective evaluator is not guaranteed to reach global optima. Although embedded methods are usually less computationally expensive than wrapper methods, they are still much slower than filter approaches. Additionally, the selected features in embedded methods are dependent on the learning machine. In [126], the classification accuracy of different filter, wrapper, and embedded methods on several artificial datasets was analyzed.

2.4 Hybrid Methods:

Most of the techniques that have been reported still possess numerous shortcomings, including high execution time and being trapped in local optima. These methods often fail to achieve satisfactory classification accuracy due to their failure to consider minimizing the size of the selected target feature. Additionally, they require a maximum fitness evaluation value and parameter adjustment [13]. To overcome these shortcomings, researchers have investigated many hybrid evolutionary algorithms. Instead of choosing one particular FS method and accepting its outcome as the final subset, different FS methods can be combined. These hybrid approaches are aimed at sharing the strength of each other to expand the capability of exploitation and reduce the chances of dropping in local optima. Based on the evidence that there is often not a single universally optimal feature selection technique [127], and due to the possible existence of more than one subset of features that discriminates the data equally well [26], model combination approaches such as boosting have been adapted to improve the robustness and stability of final discriminative methods [97, 128]. Various hybrid methods have been used in the literature. For example, researchers in [129] developed a hybrid method based on ABC and DE. To obtain very accurate solutions for the classification problem, [136] proposed a hybrid evolutionary approach based on Binary Black Hole Algorithm (BPSO) and Binary Particle Swarm Optimization (BBHA) for gene selection [130].

In the literature, PSO has been hybridized with other metaheuristic algorithms. [87, 131] used PSO to search for optimal feature subsets, together with three classifiers, i.e., pattern network, decision tree, and Navies Bayes. The approach was shown to achieve high accuracy. [138,139] hybrid PSO with GA (PSOGA). To decrease the high computational cost associated with feature selection algorithms, Wang et al [132] have presented the hybrid GSA-PSO for improving the solution quality obtained by GSA for solving function optimization problems. In order to solve the global optimization problem, Wang et al.

[133] hybridized PSO and GSA that combined the assets of both particle swarm optimization and GSA. Similarly, to PSOGSA [134, 135], PSO was integrated to GSA for solving a continuous problem.

GA has gained much attention in the hybrid metaheuristics field. In order to escape from local optimum of simulated annealing (SA), Gheyas et al. [136] combined both simulated annealing (SA) and GA. To improve the search ability of GA, [137] hybridized GA with SA. The performance was evaluated, and the experimental results show good classification accuracy considering the selected number of attributes. In another study, Nemati et al proposed a new hybrid algorithm of GA and Ant-cuckoo colon (ACO) in order to use the advantages of both algorithms [109][146-147]. In another attempt, Zhu et al [138] proposed a new algorithm that is a combination of a genetic algorithm and a local search method. Although finding a good feature subset using heuristic search may not be as time-consuming as an exhaustive search, it may still take a long time. An alternative approach is to use a two-stage algorithm, where the filter method is used to evaluate the “importance” of each feature individually, and then a greedy search with the help of a classifier is used to select the most important features. For the same purpose, the authors in [139] used the GA algorithm to select the optimal gene, which was provided by mRMR. In another study, Li and Yin [140] proposed a multiobjective binary biogeography-based optimization (MOBBBO) to select small subsets of informative genes, which were provided by a Fisher Markov selector. The final feature subset cannot contain genes that are not given by the filter algorithm. As a result, the final output can be locally optimal. In order to solve these drawbacks, the filter method must supply a variety of candidate feature subsets from different perspectives. [148] combined a genetic algorithm (GA) and the k-Nearest Neighbor (KNN) method to identify genes that can jointly discriminate between different classes of samples. Alba in [151] compared the use of PSO and genetic algorithm (GA), both augmented with SVM, as the classifier for high-dimensional microarray data. In comparison to GA, ACO, and PSO, the ABC algorithm has demonstrated comparable performance [141, 142].

2.5 Ensemble

The ensemble feature selection method is adopted based on an assumption indicating that a combination of several feature selection methods obtains more robust results than any individual feature selection method. It is also required to set a threshold to acquire a functional subset of features. Accordingly, when carrying out ensemble feature selection, a combinational method should be used to combine rankings of features from diverse algorithms into an individual rank for each feature. Ensemble techniques

might be used to improve the validity of feature selection techniques. Ensemble learning builds on the assumption that combining the output of multiple experts is better than using the output of any single expert [143, 144]. Creating a feature selection ensemble involves two essential steps: the first step requires applying different feature selection methods, each providing their output, whereas the second step combines the outputs of the single models. Combining different rankings has been considered in the literature, from basic methods such as Min, Max, and Mean to more complicated ones like Complete Linear Aggregation CLA. [145]. It is also necessary to set a threshold value to obtain an empirical subset of features. However, it is often unsure what the threshold is. A large feature set will increase the risk of including noisy, spare, and irrelevant features, which may reduce the learning performance. On the other hand, it is not acceptable to include a small number of selected features since some relevant features may be eliminated [146]. In the literature, classifiers is usually used to search the feature sets and pick the number that has the best learning performance, but the whole process is computationally expensive [147]. Researchers in [147] introduced a novel ensemble approach for feature selection called Automatic Thresholding Feature Selection (ATFS). ATFS consists of three sequential steps. Firstly, diversity is introduced by applying multiple rankers to each dataset, resulting in unique feature rankings. Secondly, a fast nondominated sorting technique is employed to combine the output rankings from individual selectors, thereby improving the ensemble's automatic thresholding capability. Lastly, optimal feature sets are generated to identify the most suitable subset of features for the given task. To select the optimal subsets of genes from an ensemble approach, Emmanuelle et al [148] have presented an ensemble gene selection approach using particle swarm optimization, ant colony optimization, and genetic algorithm methods.

3. Conclusion:

In light of the abundance of high-dimensional datasets resulting from the availability of data with numerous variables, feature selection has become a crucial task in data mining and machine learning. Its primary objective is to reduce data dimensionality and enhance algorithm performance, particularly in classification tasks. This paper aims to raise awareness among interested readers regarding the possibilities of feature selection and provides a comprehensive discussion of various feature selection models.

In contrast to the commonly used categorization into filter, wrapper, and embedded approaches, we propose formulating feature selection as a combinatorial optimization or search problem. Consequently, we categorize feature selection methods into exhaustive search,

heuristic search, and hybrid methods and evaluate their respective strengths and weaknesses. The filter model, despite its significantly faster execution time compared to other models, is prone to accuracy problems. We further classify the filter approach into univariate and multivariate techniques. The univariate approach considers the relevance of features to the target variable without accounting for feature redundancy. To address the limitations of the univariate approach, several multivariate filter techniques have been introduced, which assess the interdependencies among features and their relationship with the decision variable.

The wrapper model, while yielding the highest accuracy, comes at the cost of increased computational time. We classify wrapper methods based on their search strategies into exhaustive search, heuristic search, and meta-heuristic search. Exhaustive search, although capable of finding optimal solutions by exploring the entire search space, is computationally expensive. To mitigate the computational burden, various heuristic search methods guide the search for an optimal feature subset. Depending on the objective function, heuristic search algorithms can be further classified into single objective and multi-objective techniques.

In addition, the embedded model is adopted to address the limitations of the aforementioned models. However, many reported techniques still possess shortcomings such as high execution time and susceptibility to local optima. To overcome these limitations, researchers have investigated numerous hybrid evolutionary algorithms. Finally, this paper addresses the existing problems in feature selection and highlights the significant contributions made in the field of feature selection.

Author contributions

HA: took on the main role, performed the literature, designed, and wrote the paper. **GH:** supervised the research and co-wrote the paper.

Conflicts of interest

The authors declare no conflicts of interest.

Table 1A summary of existing articles on feature selection techniques

NO.	year	Reference	Algorithms	contribution	No.DS
1	2020	[149]	binary Jaya algorithm	This study presents a novel hybrid feature selection approach that combines a multi-attribute decision-making method with the binary Jaya algorithm. The objective of this method is to identify an optimal subset of features for the classification of Parkinson's disease.	One dataset
2	2018	[150]	GA	proposes a hybrid genetic algorithm with a wrapper-embedded feature approach (HGAW). The algorithm combines a genetic algorithm for global search with embedded regularization approaches for local search.	5Datasets AML, DLBCL, Lymphoma, Prostate, Lung cancer
3	2020	[111]	Teaching Learning Based Optimization algorithm with Gravitational	In this study, a novel hybrid wrapper approach named TLBOGSA is introduced to tackle the challenges associated with gene selection for cancer classification. TLBOGSA combines the teaching learning-based algorithm (TLBO) with the gravitational search algorithm (GSA) and incorporates an innovative encoding strategy to convert the continuous search space into a binary form. The proposed approach initiates with the utilization of the minimum redundancy maximum relevance (mRMR) feature selection technique, which aims to identify informative genes from gene expression datasets. Subsequently, a wrapper method is employed to select informative genes from the reduced dataset generated by mRMR. To enhance the search capability throughout the evolutionary process, the gravitational search mechanism is integrated into the teaching phase.	Seven datasets
4	2004	[151]	Hybrid GA	The paper presents a novel hybrid genetic algorithm for feature selection, incorporating local search operations to improve the results of GA	11 datasets Glass Vowel Wine Letter Vehicle Segmentation WDBC Ionosphere Satellite Sonar Numeral
5	2016	[152]	PSO	The paper introduces a novel hybrid feature selection algorithm called HPSO-LS, which combines particle swarm optimization with a local search strategy. The algorithm aims to select a subset of less correlated and salient features by considering their correlation information	12 datasets Glass Vowel Wisconsin Breast Cancer (WBC) Wine Heart

					Segment Two norm Sonar Arrhythmia LSVT Voice Rehabilitation Colon cancer Lymphoma Leukemia
6	2014	[153]	PSO	proposes the use of particle swarm optimization (PSO) as an efficient technique. It introduces three new initialization strategies and three new personal best and global best updating mechanisms in PSO to improve feature selection performance	Fourteen datasets Wine Vehicle WBCD Lung Movement Musk1 Madelon Zoo German Ionosphere Sonar Hillvalley Arrhythmia Isolet5
7	2011	[154]	Binary PSO	It introduces a novel optimization algorithm called Catfish Binary Particle Swarm Optimization (CatfishBPSO), In this proposed algorithm, the worst particle is replaced by extreme points in the search space, providing that gbest remains constant for several iterations. This replacement, which is known as the Catfish effect, improved the results.	Ten Datasets
8	2019	[155]	binary version of hybrid GWO and PSO	The paper introduces a binary version of the hybrid Grey Wolf Optimization (GWO) and Particle Swarm Optimization (PSO) algorithm for optimal feature selection. To achieve this purpose, they utilized GWO's exploration ability to improve the exploitation of PSO.	18 standard benchmark datasets
9	2017	[156]	ABC	This paper proposes a novel hybrid algorithm called AC-ABC Hybrid, combining the characteristics of ACO and Artificial Bee Colony (ABC) algorithms for feature selection optimization. The algorithm aims to eliminate stagnation behavior and improve initial solution search efficiency. Ants exploit the solutions generated by bees, while bees adapt the feature subsets obtained by ants	Thirteen datasets: Heart-Cleveland Dermatology Hepatitis Lung Cancer Lymphography Pima Indian Diabetes Iris Wisconsin Breast Cancer Diabetes Heart-Stalog Thyroid Sonar Gene

10	2018	[157]	ACO	This paper propose an improved feature selection algorithm called FACO, which combines ant colony optimization and feature selection. The algorithm includes a designed fitness function and an optimized updating rule to eliminate redundant features and prevent local optima	KDD CUP99
11	2016	[158]	FOA	This article proposes a discrete version of FOA was presented for solving feature selection as a discrete search space problem	11 datasets.
12	2018	[159]	WOA	This paper introduces a hybrid feature subset selection algorithm called MPMDIWOA, which combines a filter algorithm called MPMD based on Pearson's correlation coefficient and correlation distance, with a modified whale optimization algorithm acting as a wrapper algorithm	10 benchmark datasets
13	2017	[156]	ACO and ABC algorithms	This paper presents a novel Swarm-based hybrid algorithm called AC-ABC Hybrid, which combines the characteristics of Ant Colony and Artificial Bee Colony (ABC) algorithms to optimize feature selection. The algorithm aims to eliminate stagnation behavior in ants and reduce the time-consuming global search for initial solutions in bees	Thirteen
14	2017	[160]	WOA and Simulated Annealing (SA)	This study introduces two hybrid models that combine the Whale Optimization Algorithm (WOA) and Simulated Annealing (SA) algorithms. In the first model, SA is utilized to enhance the best solution found by the WOA algorithm after each iteration. In the second model, SA is employed to improve the exploitation capability of the WOA algorithm.	18 standard benchmark datasets
15	2007	[161]	Markov Blanket and GA	The researchers introduced a hybrid approach that incorporated the concept of Markov Blanket into the Genetic Algorithm (GA). The Markov Blanket was integrated into the GA to enhance its solution by adding or deleting features from a genetic algorithm (GA) solution	11 publicly available datasets
16	2009	[162]	SVM+ mRMR	The researchers enhance support vector machine recursive feature elimination (SVM-RFE) method by integrating a minimum-redundancy maximum-relevancy (MRMR) filter, which considers both the mutual information among genes and class labels for relevancy and the mutual information among genes for redundancy.	Four microarray gene expression cancer datasets
17	2010	[163]	correlation-based feature selection+GA	In this paper, the authors present a comprehensive evaluation of hybrid feature selection models that integrate correlation-based feature selection (CFS), Gain Ratio (GR), and Information Gain (INFO) as filter methods, and employ Support Vector Machine (SVM) classifier in conjunction with Greedy Search (GS) and Genetic Algorithm (GA) as wrapper	Three public gene expression datasets.

				methods. The performance assessment involves the analysis of precision, recall, F-measure, and accuracy rate, demonstrating that the CFSSVMGA model exhibits superior performance compared to other models across three distinct gene expression datasets.	
18	2011	[139]	mRMR+GA	In this paper, a novel two-stage selection algorithm is presented, which integrates the Minimum Redundancy–Maximum Relevance (MRMR) technique and Genetic Algorithm (GA) for the analysis of genomic data. The proposed methodology involves the utilization of MRMR in the initial stage to effectively filter out genes that are noisy and redundant. Subsequently, GA is employed in the second stage, leveraging the accuracy of the classifier as a fitness function to identify genes that exhibit high discriminative power	five open datasets
19	2014	[139]	PSO+ Decision tree	This paper presents a novel approach for gene selection in the context of cancer identification. The proposed method combines particle swarm optimization with a decision tree classifier to identify a concise subset of informative genes from a large pool of available genes.	11 gene expression cancer datasets
20	2013	[164]	Harmony Search+ Markov Blanket	In this paper, the authors introduce a novel approach named HSA-MB that integrates the Harmony Search Algorithm (HSA) and the Markov Blanket (MB) for gene selection purposes. The HSA-MB hybrid methodology utilizes HSA as a wrapper technique to generate new harmonies, subsequently enhanced by the MB algorithm acting as a filter mechanism. The MB algorithm employs gene ranking information to add or remove operators, refining the harmony and optimizing the search space accordingly.	Ten microarray datasets
21	2018	[165]	correlation-based filter+ iBPSO	In this paper, a novel two-phase hybrid model is introduced for cancer classification. The model combines Correlation-based Feature Selection (CFS) with an enhanced version of Binary Particle Swarm Optimization (iBPSO). The objective of this model is to identify a reduced-dimensional set of prognostic genes that can effectively classify biological samples in both binary and multi-class cancer scenarios.	11 microarray datasets
22	2002	[166]	Multi-objective Genetic algorithm	The author explores the practical implementation of genetic algorithms for the purpose of feature selection. The unique aspect of this approach is the utilization of multi-objective genetic algorithms that integrate sensitivity analysis and neural networks.	One dataset
23	2010	[167]	NSGA-II	This paper presents a novel approach to address the task of multiobjective feature selection. The proposed method is built upon the NSGA-II optimization approach, which is modified to	One dataset

				select local feature subsets of various sizes. Additionally, the method incorporates a nondominated solution search technique to identify global nondominated feature subsets.	
24	2018	[168]	ABC multi-objective	This paper presents a feature selection approach that combines a newly developed multi-objective artificial bee colony algorithm with non-dominated sorting procedures and genetic operators. The proposed approach is implemented in two different forms: one utilizing binary representation and the other utilizing continuous representation	12 datasets
25	2019	[169]	Filter-Wrapper based on NSGA-II	This research presents a novel hybrid tri-objective evolutionary algorithm that seeks to optimize two filter objectives, specifically the number of features and the mutual information, along with a wrapper objective that is associated with accuracy. The algorithm divides the population into separate non-dominated fronts and then concentrates on enhancing the feature subsets within the first (best) front.	twelve datasets
26	2020	[170]	Multi-objective Grey Wolf optimizer (MOGWO)	In recent times, the Multi-objective Grey Wolf optimizer (MOGWO) was introduced as a solution for multi-objective optimization problems. However, MOGWO was specifically designed for continuous optimization problems, making it unsuitable for directly addressing multi-objective feature selection problems, which are inherently discrete in nature. Consequently, this study presents a binary variant of MOGWO, referred to as BMOGW-S, which employs a sigmoid transfer function. The purpose of BMOGW-S is to optimize feature selection problems within a binary context.	15 standard benchmark datasets
27	2021	[171]	forest optimization algorithm (FOA)	This research paper presents a novel multi-objective feature selection algorithm that utilizes the forest optimization algorithm (FOA) in conjunction with the concepts of archive, grid, and region-based selection. To achieve this objective, the research effort involves the development of two distinct versions of the algorithm, utilizing both continuous and binary representations	nine UCI datasets and two microarray datasets

References

- [1] Qasim, O.S. and Z.Y. Algamal, Feature selection using different transfer functions for binary bat algorithm. *International Journal of Mathematical, Engineering and Management Sciences*, 2020. 5(4): p. 697.
- [2] Li, J. and H. Liu, Challenges of feature selection for big data analytics. *IEEE Intelligent Systems*, 2017. 32(2): p. 9-15.
- [3] Saeys, Y., I. Inza, and P. Larranaga, A review of feature selection techniques in bioinformatics. *bioinformatics*, 2007. 23(19): p. 2507-2517.
- [4] Mitra, S., R. Das, and Y. Hayashi, Genetic networks and soft computing. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2009. 8(1): p. 94-107.
- [5] Phan, J.H., C.F. Quo, and M.D. Wang, Cardiovascular genomics: a biomarker identification pipeline. *IEEE Transactions on Information Technology in Biomedicine*, 2012. 16(5): p. 809-822.
- [6] Chen, C.C., et al., Methods for identifying SNP interactions: a review on variations of Logic Regression, Random Forest and Bayesian logistic regression. *IEEE/ACM transactions on computational biology and bioinformatics*, 2011. 8(6): p. 1580-1591.
- [7] Kourou, K., et al., Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 2015. 13: p. 8-17.
- [8] Braga-Neto, U., Fads and fallacies in the name of small-sample microarray classification-a highlight of misunderstanding and erroneous usage in the applications of genomic signal processing. *IEEE Signal Processing Magazine*, 2007. 24(1): p. 91-99.
- [9] Liang, M.P., et al., Computational functional genomics. *IEEE Signal Processing Magazine*, 2004. 21(6): p. 62-69.
- [10] Langley, P. and S. Sage, Induction of selective Bayesian classifiers, in *Uncertainty Proceedings 1994*. 1994, Elsevier. p. 399-406.
- [11] Langley, P., *Elements of machine learning*. 1996: Morgan Kaufmann.
- [12] Fu, X. and L. Wang, Data dimensionality reduction with application to simplifying RBF network structure and improving classification performance. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2003. 33(3): p. 399-409.
- [13] Ben-Dor, A., et al. Tissue classification with gene expression profiles. in *Proceedings of the fourth annual international conference on Computational molecular biology*. 2000.
- [14] Troyanskaya, O.G., et al., Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, 2002. 18(11): p. 1454-1461.
- [15] Liu, H. and L. Yu, Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering*, 2005. 17(4): p. 491-502.
- [16] Guyon, I. and A. Elisseeff, An introduction to variable and feature selection. *Journal of machine learning research*, 2003. 3(Mar): p. 1157-1182.
- [17] Qasim, O.S., N.A. Al-Thanoon, and Z.Y. Algamal, Feature selection based on chaotic binary black hole algorithm for data classification. *Chemometrics and Intelligent Laboratory Systems*, 2020. 204: p. 104104.
- [18] Wang, L. and X. Fu, *Data mining with computational intelligence*. 2005: Springer Science & Business Media.
- [19] Shukla, A.K., P. Singh, and M. Vardhan, A hybrid framework for optimal feature subset selection. *Journal of Intelligent & Fuzzy Systems*, 2019. 36(3): p. 2247-2259.
- [20] Paul, S. and S. Das, Simultaneous feature selection and weighting—an evolutionary multi-objective optimization approach. *Pattern Recognition Letters*, 2015. 65: p. 51-59.
- [21] Daelemans, W., et al. Combined optimization of feature selection and algorithm parameters in machine learning of language. in *Machine Learning: ECML 2003: 14th European Conference on Machine Learning*, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. *Proceedings 14*. 2003. Springer.
- [22] Raman, B. and T.R. Ioerger, Enhancing learning using feature and example selection. *Journal of Machine Learning Research* (submitted for publication), 2003.
- [23] Halgamuge, S.K. and L. Wang, Classification and clustering for knowledge discovery. Vol. 4. 2005: Springer Science & Business Media.
- [24] Wang, L., Y. Wang, and Q. Chang, Feature selection methods for big data bioinformatics: A survey from the search perspective. *Methods*, 2016. 111: p. 21-31.
- [25] Haar, L., et al. Comparison between Supervised and Unsupervised Feature Selection Methods. in *ICPRAM*. 2019.
- [26] Guyon, I. and A. Elisseeff, An introduction to feature extraction, in *Feature extraction: foundations and applications*. 2006, Springer. p. 1-25.
- [27] Law, M.H., M.A. Figueiredo, and A.K. Jain, Simultaneous feature selection and clustering using

- mixture models. *IEEE transactions on pattern analysis and machine intelligence*, 2004. 26(9): p. 1154-1166.
- [28] Marczyk, M., et al., Adaptive filtering of microarray gene expression data based on Gaussian mixture decomposition. *BMC bioinformatics*, 2013. 14: p. 1-12.
- [29] Ding, C. and H. Peng, Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 2005. 3(02): p. 185-205.
- [30] Reinelt, W., A. Garulli, and L. Ljung, Comparing different approaches to model error modeling in robust identification. *Automatica*, 2002. 38(5): p. 787-803.
- [31] Wang, D., F. Nie, and H. Huang, Feature selection via global redundancy minimization. *IEEE transactions on Knowledge and data engineering*, 2015. 27(10): p. 2743-2755.
- [32] Ultsch, A., et al. A comparison of algorithms to find differentially expressed genes in microarray data. in *Advances in Data Analysis, Data Handling and Business Intelligence: Proceedings of the 32nd Annual Conference of the Gesellschaft für Klassifikation eV, Joint Conference with the British Classification Society (BCS) and the Dutch/Flemish Classification Society (VOC), Helmut-Schmidt-University, Hamburg, July 16-18, 2008*. 2010. Springer.
- [33] Duch, W., et al. Comparison of feature ranking methods based on information entropy. in *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)*. 2004. IEEE.
- [34] Lazar, C., et al., A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM transactions on computational biology and bioinformatics*, 2012. 9(4): p. 1106-1119.
- [35] Liu, Y., A comparative study on feature selection methods for drug discovery. *Journal of chemical information and computer sciences*, 2004. 44(5): p. 1823-1828.
- [36] Peng, H., F. Long, and C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 2005. 27(8): p. 1226-1238.
- [37] Bolón-Canedo, V., et al., A review of microarray datasets and applied feature selection methods. *Information sciences*, 2014. 282: p. 111-135.
- [38] Inza, I., et al., Filter versus wrapper gene selection approaches in DNA microarray domains. *Artificial intelligence in medicine*, 2004. 31(2): p. 91-103.
- [39] Meyer, P.E., C. Schretter, and G. Bontempi, Information-theoretic feature selection in microarray data using variable complementarity. *IEEE Journal of Selected Topics in Signal Processing*, 2008. 2(3): p. 261-274.
- [40] Brown, G., et al., Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *The journal of machine learning research*, 2012. 13: p. 27-66.
- [41] Bommert, A., et al., Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis*, 2020. 143: p. 106839.
- [42] Bolón-Canedo, V., N. Sánchez-Marono, and A. Alonso-Betanzos. *Feature Selection for High-Dimensional Data. in Artificial Intelligence: Foundations, Theory, and Algorithms*. 2015.
- [43] Lê Cao, K.-A., A. Bonnet, and S. Gadat, Multiclass classification and gene selection with a stochastic algorithm. *Computational Statistics & Data Analysis*, 2009. 53(10): p. 3601-3615.
- [44] Steuer, R., et al., The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, 2002. 18(suppl_2): p. S231-S240.
- [45] Liu, X., A. Krishnan, and A. Mondry, An entropy-based gene selection method for cancer classification using microarray data. *BMC bioinformatics*, 2005. 6(1): p. 1-14.
- [46] Apiletti, D., et al. The painter's feature selection for gene expression data. in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2007. IEEE.
- [47] Chuang, L.-Y., et al., A two-stage feature selection method for gene expression data. *OMICS A journal of Integrative Biology*, 2009. 13(2): p. 127-137.
- [48] Yan, X., et al., Detecting differentially expressed genes by relative entropy. *Journal of theoretical biology*, 2005. 234(3): p. 395-402.
- [49] Van't Veer, L.J., et al., Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 2002. 415(6871): p. 530-536.
- [50] Fix, E. and J.L. Hodges, Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 1989. 57(3): p. 238-247.
- [51] Wu, D. and S. Guo, An improved Fisher Score feature selection method and its application. *J. Liaoning Tech. Univ.(Nat. Sci.)*, 2019. 38: p. 472-479.
- [52] Guyon, I., et al., Gene selection for cancer classification using support vector machines. *Machine learning*, 2002. 46: p. 389-422.

- [53] Chen, Y.-W. and C.-J. Lin, Combining SVMs with various feature selection strategies. *Feature extraction: foundations and applications*, 2006: p. 315-324.
- [54] Hancer, E., B. Xue, and M. Zhang, Differential evolution for filter feature selection based on information theory and feature ranking. *Knowledge-Based Systems*, 2018. 140: p. 103-119.
- [55] Thomas, J.A., *Elements of information theory*. 1991.
- [56] Fleuret, F., Fast binary feature selection with conditional mutual information. *Journal of Machine learning research*, 2004. 5(9).
- [57] McGill, W., Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory*, 1954. 4(4): p. 93-111.
- [58] Kira, K. and L.A. Rendell. The feature selection problem: Traditional methods and a new algorithm. in *Proceedings of the tenth national conference on Artificial intelligence*. 1992.
- [59] Acuna, E., F. Coaquira, and M. Gonzalez. A comparison of feature selection procedures for classifiers based on kernel density estimation. in *Proc. of the int. conf. on computer, communication and control technologies, CCCT*. 2003.
- [60] Yang, P., et al., Gene-gene interaction filtering with ensemble of filters. *BMC bioinformatics*, 2011. 12(1): p. 1-10.
- [61] Spolaôr, N., et al. ReliefF for multi-label feature selection. in *2013 Brazilian Conference on Intelligent Systems*. 2013. IEEE.
- [62] Huang, X., et al., Feature clustering based support vector machine recursive feature elimination for gene selection. *Applied Intelligence*, 2018. 48: p. 594-607.
- [63] Yan, K. and D. Zhang, Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sensors and Actuators B: Chemical*, 2015. 212: p. 353-363.
- [64] Jain, A. and D. Zongker, Feature selection: Evaluation, application, and small sample performance. *IEEE transactions on pattern analysis and machine intelligence*, 1997. 19(2): p. 153-158.
- [65] Somol, P., P. Pudil, and J. Kittler, Fast branch & bound algorithms for optimal feature selection. *IEEE Transactions on pattern analysis and machine intelligence*, 2004. 26(7): p. 900-912.
- [66] Liu, L., et al. A comparative study on unsupervised feature selection methods for text clustering. in *2005 International Conference on Natural Language Processing and Knowledge Engineering*. 2005. IEEE.
- [67] Sharma, M. and P. Kaur, A comprehensive analysis of nature-inspired meta-heuristic techniques for feature selection problem. *Archives of Computational Methods in Engineering*, 2021. 28: p. 1103-1127.
- [68] Ruiz, R., J.C. Riquelme, and J.S. Aguilar-Ruiz, Incremental wrapper-based gene selection from microarray data for cancer classification. *Pattern Recognition*, 2006. 39(12): p. 2383-2392.
- [69] Wang, A., et al., Wrapper-based gene selection with Markov blanket. *Computers in biology and medicine*, 2017. 81: p. 11-23.
- [70] Kohavi, R. and G.H. John, Wrappers for feature subset selection. *Artificial intelligence*, 1997. 97(1-2): p. 273-324.
- [71] Breiman, L., Random forests. *Machine learning*, 2001. 45: p. 5-32.
- [72] Cortes, C. and V. Vapnik, Support-vector networks. *Machine learning*, 1995. 20: p. 273-297.
- [73] Wah, Y.B., et al., Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy. *Pertanika Journal of Science & Technology*, 2018. 26(1).
- [74] Xue, B., M. Zhang, and W.N. Browne, A comprehensive comparison on evolutionary feature selection approaches to classification. *International Journal of Computational Intelligence and Applications*, 2015. 14(02): p. 1550008.
- [75] Wu, G., R. Mallipeddi, and P.N. Suganthan, Ensemble strategies for population-based optimization algorithms—A survey. *Swarm and evolutionary computation*, 2019. 44: p. 695-711.
- [76] Shukla, A.K., Identification of cancerous gene groups from microarray data by employing adaptive genetic and support vector machine technique. *Computational Intelligence*, 2020. 36(1): p. 102-131.
- [77] Bins, J. and B.A. Draper. Feature selection from huge feature sets. in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*. 2001. IEEE.
- [78] Liu, H. and H. Motoda, *Feature selection for knowledge discovery and data mining*. Vol. 454. 2012: Springer Science & Business Media.
- [79] Blum, A.L. and P. Langley, Selection of relevant features and examples in machine learning. *Artificial intelligence*, 1997. 97(1-2): p. 245-271.
- [80] Wang, L., F. Chu, and W. Xie, Accurate cancer classification using expressions of very few genes. *IEEE/ACM Transactions on computational biology and bioinformatics*, 2007. 4(1): p. 40-53.
- [81] Wang, L., S. Arunkumar, and W. Gu. Genetic algorithms for optimal channel assignment in mobile communications. in *Proceedings of the 9th International Conference on Neural Information Processing*, 2002. ICONIP'02. 2002. IEEE.

- [82] Mnich, K. and W.R. Rudnicki, All-relevant feature selection using multidimensional filters with exhaustive search. *Information Sciences*, 2020. 524: p. 277-297.
- [83] Selman, B. and C.P. Gomes, Hill-climbing search. *Encyclopedia of cognitive science*, 2006. 81: p. 82.
- [84] Selman, B. and H. Levesque. Mitchell: A new method for solving hard satisfiability problems. in *Proc. of the Tenth National Conference on Artificial Intelligence*.
- [85] Hatamikia, S., K. Maghooli, and A.M. Nasrabadi, The emotion recognition system based on autoregressive model and sequential forward feature selection of electroencephalogram signals. *Journal of medical signals and sensors*, 2014. 4(3): p. 194.
- [86] Whitney, A.W., A direct method of nonparametric measurement selection. *IEEE transactions on computers*, 1971. 100(9): p. 1100-1103.
- [87] Xue, Y., L. Zhang, and B. Wang. Dissimilarity-based sequential backward feature selection algorithm for fault diagnosis. in *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14–18, 2017, Proceedings, Part IV 24. 2017. Springer*.
- [88] Marill, T. and D. Green, On the effectiveness of receptors in recognition systems. *IEEE transactions on Information Theory*, 1963. 9(1): p. 11-17.
- [89] Pudil, P., J. Novovičová, and J. Kittler, Floating search methods in feature selection. *Pattern recognition letters*, 1994. 15(11): p. 1119-1125.
- [90] Min, F., Q. Hu, and W. Zhu, Feature selection with test cost constraint. *International Journal of Approximate Reasoning*, 2014. 55(1): p. 167-179.
- [91] Diao, R. and Q. Shen, Nature inspired feature selection meta-heuristics. *Artificial Intelligence Review*, 2015. 44: p. 311-340.
- [92] Mirjalili, S. and A. Lewis, The whale optimization algorithm. *Advances in engineering software*, 2016. 95: p. 51-67.
- [93] Kirkpatrick, S., C.D. Gelatt Jr, and M.P. Vecchi, Optimization by simulated annealing. *science*, 1983. 220(4598): p. 671-680.
- [94] Wang, L., S.N.S. Lee, and W.Y. Hing. Solving channel assignment problems using local search methods and simulated annealing. in *Independent Component Analyses, Wavelets, Neural Networks, Biosystems, and Nanoengineering IX. 2011. SPIE*.
- [95] Macal, C.M., Agent-based modeling and artificial life. *Complex Social and Behavioral Systems: Game Theory and Agent-Based Models*, 2020: p. 725-745.
- [96] Zhu, M. and L. Wang. Intelligent trading using support vector regression and multilayer perceptrons optimized with genetic algorithms. in *The 2010 international joint conference on neural networks (IJCNN). 2010. IEEE*.
- [97] Wang, L., L. Zhou, and W. Liu. FPGA segmented channel routing using genetic algorithms. in *2005 IEEE Congress on Evolutionary Computation. 2005. IEEE*.
- [98] Li, B., L. Wang, and W. Song. Ant colony optimization for the traveling salesman problem based on ants with memory. in *2008 fourth international conference on natural computation. 2008. IEEE*.
- [99] Wang, L. and G. Si. Optimal location management in mobile computing with hybrid genetic algorithm and particle swarm optimization (GA-PSO). in *2010 17th IEEE International Conference on Electronics, Circuits and Systems. 2010. IEEE*.
- [100] Alsalibi, B., I. Venkat, and M.A. Al-Betar, A membrane-inspired bat algorithm to recognize faces in unconstrained scenarios. *Engineering Applications of Artificial Intelligence*, 2017. 64: p. 242-260.
- [101] Yang, B., et al., Feature selection based on modified bat algorithm. *IEICE TRANSACTIONS on Information and Systems*, 2017. 100(8): p. 1860-1869.
- [102] Zheng, Y., et al., A novel hybrid algorithm for feature selection. *Personal and Ubiquitous Computing*, 2018. 22: p. 971-985.
- [103] Aziz, M.A.E. and A.E. Hassanien, Modified cuckoo search algorithm with rough sets for feature selection. *Neural Computing and Applications*, 2018. 29: p. 925-934.
- [104] Gunavathi, C. and K. Premalatha, Cuckoo search optimisation for feature selection in cancer classification: a new approach. *International journal of data mining and bioinformatics*, 2015. 13(3): p. 248-265.
- [105] Tharwat, A., T. Gabel, and A.E. Hassanien. Classification of toxicity effects of biotransformed hepatic drugs using optimized support vector machine. in *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2017. 2018. Springer*.
- [106] Xue, B., et al., A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on evolutionary computation*, 2015. 20(4): p. 606-626.
- [107] Brezočnik, L., I. Fister Jr, and V. Podgorelec, Swarm intelligence algorithms for feature selection: a review. *Applied Sciences*, 2018. 8(9): p. 1521.
- [108] Abd Elaziz, M., et al., Opposition-based moth-flame optimization improved by differential evolution for feature selection. *Mathematics and Computers in*

Simulation, 2020. 168: p. 48-75.

- [109] Nemati, S., et al., A novel ACO–GA hybrid algorithm for feature selection in protein function prediction. *Expert systems with applications*, 2009. 36(10): p. 12086-12094.
- [110] Shukla, A.K., P. Singh, and M. Vardhan, Neighbour teaching learning based optimization for global optimization problems. *Journal of Intelligent & Fuzzy Systems*, 2018. 34(3): p. 1583-1594.
- [111] Shukla, A.K., P. Singh, and M. Vardhan, Gene selection for cancer types classification using novel hybrid metaheuristics approach. *Swarm and Evolutionary Computation*, 2020. 54: p. 100661.
- [112] Han, X., et al., Feature subset selection by gravitational search algorithm optimization. *Information Sciences*, 2014. 281: p. 128-146.
- [113] Banitalebi, A., M.I. Abd Aziz, and Z.A. Aziz, A self-adaptive binary differential evolution algorithm for large scale binary optimization problems. *Information Sciences*, 2016. 367: p. 487-511.
- [114] Song, X.-f., et al., Feature selection using bare-bones particle swarm optimization with mutual information. *Pattern Recognition*, 2021. 112: p. 107804.
- [115] Gumus, E., Z. Gormez, and O. Kursun, Multi objective SNP selection using pareto optimality. *Computational biology and chemistry*, 2013. 43: p. 23-28.
- [116] Amoozegar, M. and B. Minaei-Bidgoli, Optimizing multi-objective PSO based feature selection method using a feature elitism mechanism. *Expert Systems with Applications*, 2018. 113: p. 499-514.
- [117] Yong, Z., G. Dun-wei, and Z. Wan-qiu, Feature selection of unreliable data using an improved multi-objective PSO algorithm. *Neurocomputing*, 2016. 171: p. 1281-1290.
- [118] Dashtban, M., M. Balafar, and P. Suravajhala, Gene selection for tumor classification using a novel bio-inspired multi-objective approach. *Genomics*, 2018. 110(1): p. 10-17.
- [119] Sahoo, A. and S. Chandra, Multi-objective grey wolf optimizer for improved cervix lesion classification. *Applied Soft Computing*, 2017. 52: p. 64-80.
- [120] Zhang, Y., et al., A PSO-based multi-objective multi-label feature selection method in classification. *Scientific reports*, 2017. 7(1): p. 376.
- [121] Mukhopadhyay, A. and U. Maulik, An SVM-wrapped multiobjective evolutionary feature selection approach for identifying cancer-microRNA markers. *IEEE transactions on nanobioscience*, 2013. 12(4): p. 275-281.
- [122] Xue, B., et al., Multi-objective evolutionary algorithms for filter based feature selection in classification. *International Journal on Artificial Intelligence Tools*, 2013. 22(04): p. 1350024.
- [123] Zhong, J., et al., A feature selection method for prediction essential protein. *Tsinghua Science and Technology*, 2015. 20(5): p. 491-499.
- [124] Tang, Y., Y.-Q. Zhang, and Z. Huang, Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2007. 4(3): p. 365-381.
- [125] Pang, H., et al., Gene selection using iterative feature elimination random forests for survival outcomes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2012. 9(5): p. 1422-1431.
- [126] Bolón-Canedo, V., N. Sánchez-Marroño, and A. Alonso-Betanzos, A review of feature selection methods on synthetic data. *Knowledge and information systems*, 2013. 34: p. 483-519.
- [127] Kennedy, J. and R.C. Eberhart. A discrete binary version of the particle swarm algorithm. in 1997 IEEE International conference on systems, man, and cybernetics. *Computational cybernetics and simulation*. 1997. IEEE.
- [128] Abdelaziz, A.Y. and A. Fathy, A novel approach based on crow search algorithm for optimal selection of conductor size in radial distribution networks. *Engineering Science and Technology, an International Journal*, 2017. 20(2): p. 391-402.
- [129] Zorarpacı, E. and S.A. Özel, A hybrid approach of differential evolution and artificial bee colony for feature selection. *Expert Systems with Applications*, 2016. 62: p. 91-103.
- [130] Raweh, A.A., M. Nassef, and A. Badr, A hybridized feature selection and extraction approach for enhancing cancer prediction based on DNA methylation. *IEEE Access*, 2018. 6: p. 15212-15223.
- [131] Fong, S., et al., Feature selection in life science classification: metaheuristic swarm search. *IT Professional*, 2014. 16(4): p. 24-29.
- [132] Wang, J.-S. and J.-D. Song, A Hybrid Algorithm Based on Gravitational Search and Particle Swarm Optimization Algorithm to Solve Function Optimization Problems. *Engineering Letters*, 2017. 25(1).
- [133] Mirjalili, S., G.-G. Wang, and L.d.S. Coelho, Binary optimization using hybrid particle swarm optimization and gravitational search algorithm. *Neural Computing and Applications*, 2014. 25: p. 1423-1435.
- [134] Mirjalili, S. and S.Z.M. Hashim. A new hybrid PSOGSA algorithm for function optimization. in 2010 international conference on computer and information application. 2010. IEEE.

- [135] Talbi, E.-G., et al. Comparison of population based metaheuristics for feature selection: Application to microarray data classification. in 2008 IEEE/ACS International Conference on Computer Systems and Applications. 2008. IEEE.
- [136] Gheyas, I.A. and L.S. Smith, Feature subset selection in large dimensionality domains. *Pattern recognition*, 2010. 43(1): p. 5-13.
- [137] Mafarja, M. and S. Abdullah, Investigating memetic algorithm in solving rough set attribute reduction. *International Journal of Computer Applications in Technology*, 2013. 48(3): p. 195-202.
- [138] Zhu, W., et al., Neighborhood effective information ratio for hybrid feature subset evaluation and selection. *Neurocomputing*, 2013. 99: p. 25-37.
- [139] El Akadi, A., et al., A two-stage gene selection scheme utilizing MRMR filter and GA wrapper. *Knowledge and Information Systems*, 2011. 26: p. 487-500.
- [140] Li, X. and M. Yin, Multiobjective binary biogeography based optimization for feature selection using gene expression data. *IEEE Transactions on NanoBioscience*, 2013. 12(4): p. 343-353.
- [141] El-Abd, M. A cooperative approach to the artificial bee colony algorithm. in *IEEE congress on evolutionary computation*. 2010. IEEE.
- [142] Akbari, R., et al., A multi-objective artificial bee colony algorithm. *Swarm and Evolutionary Computation*, 2012. 2: p. 39-52.
- [143] Kuncheva, L.I., *Combining pattern classifiers: methods and algorithms*. 2014: John Wiley & Sons.
- [144] Kuncheva, L.I. and C.J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 2003. 51: p. 181-207.
- [145] Abeel, T., et al., Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 2010. 26(3): p. 392-398.
- [146] Li, J., et al., Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 2017. 50(6): p. 1-45.
- [147] Abasabadi, S., et al., Automatic ensemble feature selection using fast non-dominated sorting. *Information Systems*, 2021. 100: p. 101760.
- [148] Santana, L.E.A.d.S. and A.M. de Paula Canuto, Filter-based optimization techniques for selection of feature subsets in ensemble systems. *Expert Systems with Applications*, 2014. 41(4): p. 1622-1631.
- [149] Chaudhuri, A. and T. Sahu, PROMETHEE-based hybrid feature selection technique for high-dimensional biomedical data: application to Parkinson's disease classification. *Electronics Letters*, 2020. 56(25): p. 1403-1406.
- [150] Liu, X.-Y., et al., A hybrid genetic algorithm with wrapper-embedded approaches for feature selection. *IEEE Access*, 2018. 6: p. 22863-22874.
- [151] Oh, I.-S., J.-S. Lee, and B.-R. Moon, Hybrid genetic algorithms for feature selection. *IEEE Transactions on pattern analysis and machine intelligence*, 2004. 26(11): p. 1424-1437.
- [152] Moradi, P. and M. Gholampour, A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. *Applied Soft Computing*, 2016. 43: p. 117-130.
- [153] Xue, B., M. Zhang, and W.N. Browne, Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms. *Applied soft computing*, 2014. 18: p. 261-276.
- [154] Chuang, L.-Y., S.-W. Tsai, and C.-H. Yang, Improved binary particle swarm optimization using catfish effect for feature selection. *Expert Systems with Applications*, 2011. 38(10): p. 12699-12707.
- [155] Al-Tashi, Q., et al., Binary optimization using hybrid grey wolf optimization for feature selection. *Ieee Access*, 2019. 7: p. 39496-39508.
- [156] Shunmugapriya, P. and S. Kanmani, A hybrid algorithm using ant and bee colony optimization for feature selection and classification (AC-ABC Hybrid). *Swarm and evolutionary computation*, 2017. 36: p. 27-36.
- [157] Peng, H., et al., An improved feature selection algorithm based on ant colony optimization. *Ieee Access*, 2018. 6: p. 69203-69209.
- [158] Ghaemi, M. and M.-R. Feizi-Derakhshi, Feature selection using forest optimization algorithm. *Pattern Recognition*, 2016. 60: p. 121-129.
- [159] Zheng, Y., et al., A novel hybrid algorithm for feature selection based on whale optimization algorithm. *Ieee Access*, 2018. 7: p. 14908-14923.
- [160] Mafarja, M.M. and S. Mirjalili, Hybrid whale optimization algorithm with simulated annealing for feature selection. *Neurocomputing*, 2017. 260: p. 302-312.
- [161] Zhu, Z., Y.-S. Ong, and M. Dash, Markov blanket-embedded genetic algorithm for gene selection. *Pattern Recognition*, 2007. 40(11): p. 3236-3248.
- [162] Mundra, P.A. and J.C. Rajapakse, SVM-RFE with MRMR filter for gene selection. *IEEE transactions on nanobioscience*, 2009. 9(1): p. 31-37.
- [163] Saengsiri, P., et al. Comparison of hybrid feature selection models on gene expression data. in 2010 Eighth International Conference on ICT and Knowledge Engineering. 2010. IEEE.

- [164] Shreem, S.S., S. Abdullah, and M.Z.A. Nazri, Hybridising harmony search with a Markov blanket for gene selection problems. *Information Sciences*, 2014. 258: p. 108-121.
- [165] Jain, I., V.K. Jain, and R. Jain, Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. *Applied Soft Computing*, 2018. 62: p. 203-215.
- [166] Oliveira, L.S., et al. Feature selection using multi-objective genetic algorithms for handwritten digit recognition. in *2002 International Conference on Pattern Recognition*. 2002. IEEE.
- [167] Huang, B., B. Buckley, and T.-M. Kechadi, Multi-objective feature selection by using NSGA-II for customer churn prediction in telecommunications. *Expert Systems with Applications*, 2010. 37(5): p. 3638-3646.
- [168] Hancer, E., et al., Pareto front feature selection based on artificial bee colony optimization. *Information Sciences*, 2018. 422: p. 462-479.
- [169] Hammami, M., et al., A multi-objective hybrid filter-wrapper evolutionary approach for feature selection. *Memetic Computing*, 2019. 11: p. 193-208.
- [170] Al-Tashi, Q., et al., Binary multi-objective grey wolf optimizer for feature selection in classification. *IEEE Access*, 2020. 8: p. 106247-106263.
- [171] Nouri-Moghaddam, B., M. Ghazanfari, and M. Fathian, A novel multi-objective forest optimization algorithm for wrapper feature selection. *Expert Systems with Applications*, 2021. 175: p. 114737.