Review of Feature Selection approaches based on Grouping of features

Cihan Kuzudisli Corresp., [1, 2], **Burcu Bakir-Gungor** [3], **Nurten Bulut** [3], **Bahjat Qaqish** [4], **Malik Yousef** Corresp. [5, 6]

[1] Department of Computer Engineering, Hasan Kalyoncu University, Gaziantep, Turkey

[2] Department of Electrical and Computer Engineering, Abdullah Gul University, Kayseri, Turkey

[3] Department of Computer Engineering, Abdullah Gul University, Kayseri, Turkey

[4] Department of Biostatistics, University of North Carolina at Chapel Hill, North Carolina, Chapell Hill, United States

[5] Department of Information Systems, Zefat Academic College, Zefat, Israel

[6] Galilee Digital Health Research Center, Zefat Academic College, Zefat, Israel

Corresponding Authors: Cihan Kuzudisli, Malik Yousef
Email address: cihan.kuzudisli@hku.edu.tr, malik.yousef@gmail.com

With the rapid development in technology, large amounts of high-dimensional data have been generated. This high dimensionality including redundancy and irrelevancy poses a great challenge in data analysis and decision making. Feature selection (FS) is an effective way to reduce dimensionality by eliminating redundant and irrelevant data. Most traditional FS approaches score and rank each feature individually; and then perform FS either by eliminating lower ranked features or by retaining highly-ranked features. In this review, we discuss an emerging approach to FS that is based on initially grouping features, then scoring groups of features rather than scoring individual features. Despite the presence of reviews on clustering and FS algorithms, to the best of our knowledge, this is the first review focusing on FS techniques based on grouping. The typical idea behind FS through grouping is to generate groups of similar features with dissimilarity between groups, then select representative features from each cluster. Approaches under supervised, unsupervised, semi supervised and integrative frameworks are explored. The comparison of experimental results indicates the effectiveness of sequential, optimization-based (fuzzy or evolutionary), hybrid and multi-method approaches. When it comes to biological data, involvement of external biological sources can improve analysis results. We hope this work’s findings can guide effective design of new FS approaches using feature grouping.

Review of Feature Selection Approaches Based on Grouping of Features

Cihan Kuzudisli[1,2], Burcu Bakir Gungor[3], Nurten Bulut[3], Bahjat F. Qaqish[4], Malik Yousef[5,6]

[1] Department of Computer Engineering, Hasan Kalyoncu University, Gaziantep, Turkey
[2] Department of Electrical and Computer Engineering, Abdullah Gul University, Kayseri, Turkey
[3] Department of Computer Engineering, Abdullah Gul University, Kayseri, Turkey
[4] Department of Biostatistics, University of North Carolina at Chapel Hill, NC, Chapell Hill, USA
[5] Department of Information Systems, Zefat Academic College, Zefat, Israel
[6] Galilee Digital Health Research Center, Zefat Academic College, Zefat, Israel

Corresponding Author:
Cihan Kuzudisli[1]
Gaziantep, 27010, Turkey
Email address: cihan.kuzudisli@hku.edu.tr
Malik Yousef[2]
Zefat, 13206, Israel
Email address: malik.yousef@gmail.com

Abstract

With the rapid development in technology, large amounts of high-dimensional data have been generated. This high dimensionality including redundancy and irrelevancy poses a great challenge in data analysis and decision making. Feature selection (FS) is an effective way to reduce dimensionality by eliminating redundant and irrelevant data. Most traditional FS approaches score and rank each feature individually; and then perform FS either by eliminating lower ranked features or by retaining highly-ranked features. In this review, we discuss an emerging approach to FS that is based on initially grouping features, then scoring groups of features rather than scoring individual features. Despite the presence of reviews on clustering and FS algorithms, to the best of our knowledge, this is the first review focusing on FS techniques based on grouping. The typical idea behind FS through grouping is to generate groups of similar features with dissimilarity between groups, then select representative features from each cluster. Approaches under supervised, unsupervised, semi supervised and integrative frameworks are explored. The comparison of experimental results indicates the effectiveness of sequential, optimization-based (fuzzy or evolutionary), hybrid and multi-method approaches. When it comes to biological data, involvement of external biological sources can improve analysis results. We hope this work’s findings can guide effective design of new FS approaches using feature

grouping.

Introduction

In the current digital era, the data produced by many applications in fields such as image processing, pattern recognition, machine learning and network communication grow exponentially in both dimension and size. Due to this high-dimensionality, the search space is widening and extraction of valuable knowledge from the data becomes a challenging task [1]. Also, utilizing all features in a dataset is unlikely to develop a predictive model with high accuracy. Existence of irrelevant and redundant features may weaken the generalizability of the model and decrease the overall precision of a classifier [2]. Hence, reducing the number of input variables is highly desired as it lowers the computational cost of model construction and allows improving model performance. As such, feature selection (FS) becomes an inevitable step for domain experts and data analysts.

FS is the process of selecting the minimally sized feature subset from the original set that is optimal for the target concept. It plays a crucial role in removing irrelevant and redundant features while keeping relevant and non-redundant ones [3]. Irrelevant features do not alter the target concept in any way and redundant features do not contribute to the target concept [4]. These features may contain a considerable amount of noise which can be misleading, resulting in significant computational overhead and poor predictor performance. Contrary to other dimensionality reduction techniques, FS preserves the data semantics as it does not distort the original feature representation and hence provides straightforward data interpretation for data scientists. Additionally, reduction in dimension by FS prevents overfitting that can lead to undesired validation results.

Although various FS techniques have been developed, traditional approaches to FS neglect structures of features during the selection process. Another issue is that retention and elimination of features on an individual basis ignores dependence among them. Because of these reasons, correlation between features may not be detected efficiently resulting in irrelevant or redundant features in the final subset. Some studies grouped samples (observations) for improving classification performance but these studies were not concerned with feature reduction at all [5,6].

On the other hand, FS based on grouping is an effective technique for reducing feature redundancy and enhancing classifier learning. By grouping the features, the search space is reduced substantially. Moreover, it can reduce estimator variance [7], improve stability, and reinforce generalization capability of the model. Although there are reviews of clustering methods [8] and of FS techniques [1,9], to the best of our knowledge, this is the first paper reviewing the literature on approaches to FS based on grouping. In this procedure, the process of grouping features into clusters is generally performed as the initial step, aiming to have maximal intra-class similarity (similarity in between the objects of the same cluster) and minimal inter-class similarity (i.e., objects in a cluster are more similar to those in another one) between features. These feature groups can be created by K-Means, fuzzy c-mean (FCM), hierarchical clustering, graph theory and other methods [10-12]. After cluster formation, features within each cluster are scored and selected using various techniques or metrics.

The remainder of this paper is organized as follows: We will give a concise overview of different FS methods in Section 2. In Section 3, we will present different works carried out in FS using feature grouping following the summary of traditional approaches. Then, in Section 4, we will review different studies which benefited from Recursive Cluster Elimination based on Support Vector Machine (SVM-RCE). Next, in Section 5, we will address FS techniques involving both feature grouping and incorporating domain knowledge. We discuss the advantages and disadvantages of the presented methods in Section 6. Lastly, in Section 7, we conclude our review with further discussions and future directions.

Rationale behind the review and intended audience

Nowadays, the advancements in different technologies resulted in the generation of high dimensional data in many different fields, which makes data analysis a challenging issue. Existence of irrelevant and redundant features makes it hard to infer meaningful conclusions from data, degrades model performance and leads to computational overhead. Especially in the field of molecular biology, the advancements in high throughput technologies have induced the emergence of a wealth of -omics data produced by different studies, such as genomics, transcriptomics, epigenomics, proteomics, meta-genomics, meta-transcriptomics, meta-proteomics, metabolomics, etc. For instance, high-dimensional RNA-sequencing data can be used for cancer subtype identification in order to ease cancer diagnosis and discover effective treatments. However, only a subset of features (mRNAs) carries information associated with the cancer subtype. Furthermore, this kind of biological data often involves redundant and irrelevant features which can mislead the learning procedure in modeling and can cause overfitting. As another example, in metagenomics studies the number of features (taxa) is much higher than the number of samples. This phenomenon is known as the curse of dimensionality. In this respect, some metagenomics studies focus on the FS process rather than focusing on classification [13]. Hence, FS has become a real prerequisite in the biological domain [14–17]. Due to these reasons, FS became an indispensable preprocessing step in different fields dealing with high dimensional data. Traditional approaches evaluate features without considering the correlation among them, and also this evaluation is performed on an individual basis. Furthermore, these methods generally fail to scale on a large space.

On the other hand, FS based on feature groping is a powerful approach due to the following reasons: i) it enables the discovery of correlations among features, ii) search space is significantly diminished, iii) it relieves computational burden. Although some grouping-based FS methods are proposed in the literature, to the best of our knowledge, none of the existing papers evaluate these existing approaches in detail as a review. For these reasons, compared to current literature, we believe that this review will be more guiding and suggestive for those learning the above-mentioned methods, for those working to derive such methods, and for those who want to apply this approach into their data analysis.

Survey Methodology

Our main focus in this review is to examine FS approaches via grouping. In this context, we reviewed Web of Science, Scopus, and Google Scholar on January 10, 2022 using the following query: "feature clustering" OR "feature grouping" OR "clustering based feature selection" OR "grouping based feature selection" OR "cluster based feature selection" OR "group based feature selection". We excluded those studies grouping samples (observations) or features as the final outcome and those concerned with feature extraction. We particularly focused on grouping of features as the preprocessing step followed by extraction of a reduced subset of features by a certain procedure which is subsequently input into a classification or clustering process for validation. Other articles for context were added while writing the review. Studies of this paradigm under an unsupervised setting are on a limited scale compared to the supervised setting, due to lack of labels in the former. Even though it’s not known clearly, we think that inclusion of this approach may have emerged in late 90s. Recently, interest in this concept has grown rapidly in different forms as we point out in the following sections of this review. In fact, selection of significant features by removing irrelevant or redundant ones is just one aspect; ranking of these features in terms of being informative or having discriminative power, and stability of them for different models are other issues that are taken into consideration. Here, we examined different studies that are identified in literature mining, categorized them, and presented readers a versatile work in which we aimed at providing a robust basis on the topic.

2. Basics of Feature Selection

In this section, we present basic concepts in the FS field. According to their interaction with classification model, FS techniques can be classified into filter, wrapper, and embedded techniques [18]. Later in the literature, hybrid and ensemble techniques have emerged as variants of them. Hybrid approach combines two different methods to utilize the advantages of both approaches, where the common combination is filter and wrapper methods. Ensemble technique integrates an ensemble of feature subsets and then yields the result from the ensemble. The overview of the three main types of methods is shown in **Fig. 1**.

2.1. Filter Method

Filter type methods select features by assessing intrinsic properties of data based on statistical measures instead of cross-validation performance. They are easily scalable to high-dimensional datasets, independent of the learning algorithm; they are simple and computationally fast; and they are resistant to overfitting. In this method, each feature is assigned a score determined by the selected statistical method. Afterwards, all features are ranked in descending order and those with low scores are removed using a threshold value. The remaining features comprise the feature subset and are then fed into the classification model. Consequently, FS is carried out once and then various classifiers can be employed. Disadvantages of this technique are i) features are selected irrespective of the classifier, and ii) feature dependencies are ignored. Some common statistical measures used in this technique are Information Gain (IG), Pearson's Correlation (PS), Chi Square (χ^2), Mutual Information (MI), and Symmetrical Uncertainty (SU).

2.1.1. Information Gain

Information gain (IG) [19] is an entropy-based FS method and used to measure how much information a feature carries about the target variable. IG of a feature X in a data group D with n class labels, $IG(X)$, is calculated using

$$IG(X) = E(D) - \sum_{i=1}^{n} \frac{D_i}{D} E(D_i) \qquad (1)$$

where $E(D)$ denotes the general entropy belonging to class labels, $\frac{D_i}{D}$ is the ratio of number of occurrences of each value on feature X, and $E(D_i)$ specifies the entropy of ith feature value calculated by splitting dataset D based on feature X. Entropy is a measurement of unpredictability or impurity of a data distribution and defined as

$$E(D) = -\sum_{i=1}^{n} p(i) \log_2 p(i) \qquad (2)$$

where $p(i)$ is the probability of class i in the data group D for n class labels. A feature is relevant to target variable if it has a high information gain. The way the features are selected is in a univariate way (features are selected independently), therefore, redundant features cannot be eliminated in this technique.

2.1.2. Pearson's correlation

Pearson's correlation is a measure of the dependency (similarity) of two variables and used for finding the relationship between continuous features and the target feature [20,21]. It produces the correlation coefficient r ranging between -1 to 1, where 1 shows a strong correlation and -1 means a total negative correlation. So, 0 value implies no correlation between the features. A positive correlation states that if one variable increases, so does the other variable, whereas a negative correlation implies that while one variable raises, another one decreases. This method can also be used to measure correlation between pairs of features. In this way redundant features can be identified. Pearson's correlation coefficient r can be found for feature X with values x and classes Y with values y where X, Y are random variables by the following equation:

$$r = \frac{\Sigma(x-\overline{x})(y-\overline{y})}{\sqrt{\Sigma(x-\overline{x})^2\Sigma(y-\overline{y})^2}} \tag{3}$$

where $\overline{x}$ and $\overline{y}$ are means of x and y, respectively. Note that Pearson's correlation is mainly covariance of two variables divided by product of their standard deviations.

2.1.3. Chi Square

Chi square (χ^2) [22] is a statistical method to test the independence of two events. It's a measurement of the degree of association between two categorical values. It measures the deviation from the expected frequency assuming the feature event is independent of the class label. This assumption is tested for a given feature with n class and m different feature values by the formula

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \tag{4}$$

where O_{ij} is the observed (actual) value and E_{ij} refers to the expected value suggested by the null hypothesis. E_{ij} is calculated as

$$E_{ij} = \frac{(O_{*j}O_{i*})}{O} \tag{5}$$

where O_{*j} means the number of samples in class m, and O_{i*} indicates the number of samples with the ith feature value for the feature under study. Higher value of χ^2 shows rejection to the null hypothesis, namely, higher dependency between the feature and the class label.

2.1.4. Mutual Information

Mutual information (MI) [23] is another statistical method used to assess the mutual dependence between the two variables. MI quantifies the amount of information that one random variable includes in the other random variable. MI between two continuous random variables X and Y with their joint probability functions $p(x,y)$, and their marginal probability density functions $p(x)$ and $p(y)$, respectively is given by

$$I(X;Y) = \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \qquad (6)$$

For discrete random variables, the double integral is substituted by a summation as

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \qquad (7)$$

We can also define the conditional mutual information (CMI) of two random variables X and Y given a third variable Z as

$$I(X;Y \,\middle|\, Z) = \iiint p(x,y,z)\log\frac{p(x,y|z)}{p(x,z)p(y|z)}dxdydz \qquad (8)$$

It can be interpreted as the amount of information X includes in Y which is not shared by Z.

2.1.5. Symmetrical Uncertainty

This is one of the techniques that are used to measure redundancy between two random variables [24]. It is obtained by normalizing MI to the entropies of two variables and limiting it to the range of [0,1]. It's able to circumvent inherent bias of MI toward features with a wide range of different values. Symmetrical Uncertainty (SU) is defined as

$$SU(X,Y) = \frac{2MI(X,Y)}{H(X)+H(Y)} \qquad (9)$$

where $H(X)$ and $H(Y)$ are entropy of variable X and Y, respectively. A value 1 between a pair of features indicates that knowledge of feature value can fully predict the values of other and 0 value shows that X and Y are not correlated.

Based on SU, C-Relevance between a feature and a target variable C, and F-Correlation between feature pair can be defined as follows [25]:

C-Relevance: SU between feature $F_i \in F$ and target variable C, denoted by $SU_{i,c}$.

F-Correlation: SU between any feature pair F_i and F_j ($i \neq$ j), denoted by $SU_{i,j}$.

2.2. Wrapper Method

In this methodology, a search strategy for possible subsets of features is defined, and the learning algorithm is trained using these subsets in an iterative manner. Unlike filter methods, wrapper methods are in interaction with the classifier, however, the evaluation of feature subsets is obtained using a specific classification model which makes this method specific to a learning model. Several possible combinations of features are evaluated in the model by wrapping the search algorithm around it [26]. This method provides suboptimal feature subsets for training the model since evaluating all possible subsets is computationally not practical, and generally gives better predictive accuracy than filter methods but is computationally intensive due to searching overhead and learner dependence.

The search for generating subsets can be performed with schemes such as Forward Selection, Backward Elimination, Stepwise Selection or a heuristic search [27]. Forward selection is a repetitive technique where no feature is considered at the onset. Initially, the feature with the best performance is added. Then another most significant feature giving the best performance together with the previously added feature is selected. This process proceeds until the inclusion of a new feature does not improve the classifier performance. In backward elimination, the algorithm starts with all the features available and discards the most insignificant feature from the model recursively. This elimination process is repeated until removal of features does not enhance the performance of the model. For stepwise selection, this technique is a combination of both forward selection and backward elimination. It starts with an empty set and the most significant feature is added at each iteration. While adding a new feature, previously selected features are removed if any of them has become insignificant. Heuristic search is concerned with optimization and aims at optimizing the objective function in evaluation of different subsets [28].

Support Vector Machines with Recursive Feature Elimination (SVM-RFE) [29] is a popular example of wrapper methods. The idea is mainly to train the classifier by the given data and assign a rank by SVM for each feature as its weight. Then, features with the smallest weights are removed by a specific rate determined by the user. This procedure is repeated until reaching a predefined number of features.

2.3. Embedded Method

This method includes advantages of filter and wrapper methods and performs FS and model construction at the same time. Just like wrapper techniques, they are specific to a learning model but they have less computational complexity than wrapper methods [30]. One technique of this type of FS is regularization that adds a penalty to the coefficients to overcome overfitting in the model. As an example, Lasso [31] is an embedded method that uses L_1 norm of the coefficient of a linear classifier **w** and penalty term (φ) is defined as

$$\varphi(\mathrm{w}) = \sum_{i=1}^{k} |\mathrm{w}_i| \tag{10}$$

and

$$\hat{\mathbf{w}} = \min_{\mathbf{w}} c(\mathbf{w}, X) + \alpha\varphi \tag{11}$$

where $c(.)$ is the objective function for classification, φ is a regularization term, k is the number of features, α is the regularization parameter controlling the trade-off between the objective function and the penalty. These coefficients may even be reduced to 0 for features that do not contribute to the model. Features with non-zero coefficients are retained and those with low or zero coefficient are excluded [32]. Another technique to integrate FS in model creation is decision trees. These tree-based methods are non-parametric models that consider features as nodes. Tree-based strategies used by random forests accumulate various numbers of decision trees and rank the nodes (features) by decrease in the impurity (Gini impurity) over all the trees, e.g. CART [33].

3. Feature Selection Approaches

Broadly speaking, FS algorithms conducted in many studies can be categorized into the following two classes: i) traditional FS, ii) FS based on grouping. Traditional approaches generally consider all features contingent on “singularity” during the selection process. To put it another way, they comprise inclusion or elimination of features based on some statistical measures or classifying capacity at a singular level. On the other hand, grouping-based methods detect relevant features by grouping them into clusters; and then remove redundant ones which lead to reduced search space.

3.1. Traditional Feature Selection

Different FS methods exist in abundance in the literature, including filters based on distinct criteria (dependency, information, distance and consistency [34]), and wrapper and embedded methods employing different induction algorithms. Due to their simplicity, filter methods are often preferable in the context of high dimensional data; the absence of necessity for a search route and the interaction with a classifier makes them computationally efficient and practically feasible in applications. A comparative study on various filtering methods (mixture model, regression modeling and t-test) was presented in [35] where the authors outlined similar and dissimilar aspects of these methods. The authors noted that all the three methods employ two-sample t-test or its variation; but these methods vary in different significance levels and the number of detected features. Lazar et al. [36] also reviewed filter type FS algorithms used in gene expression data analysis and presented them as a top-bottom strategy in a taxonomy.

Wrapper methods carry the computational burden since they require navigation in the search domain and and since they interact with the predictor. However, they provide better accuracy than filter approaches due to their interaction with the learning algorithm. Talavera L. et al. [37] compared filter and wrapper approaches in clustering. They confirm the superiority of wrappers along with some of their problems and they suggest filter techniques as an alternative approach due to their computational efficiency. A recent study [38] overviewed existing wrapper techniques and evaluated the pros and cons of them. Embedded methods, like wrapper techniques, possess computational complexity when it comes to high-dimensional data. They are more efficient than wrappers and have less complexity. Applications of this approach in the bioinformatics domain have been reviewed in [39].

Hybrid methods combine two methods (such as filter and wrapper) to take advantage of both methods in order to increase efficiency and performance. Ensemble methods integrate different methods for FS, classification or both. In this approach, multiple feature selectors, induction algorithms, different subsets may be included according to the design scheme. A detailed discussion on hybrid methods and a good review on ensemble FS techniques can be found in [40] and [41] , respectively. In some studies, FS methods are divided into these five categories [42].

Traditional FS approaches have several shortcomings. For instance, filter methods evaluate the significance of each feature individually without considering the relationships and interactions between the features. Wrapper methods can provide the optimal feature subset but their complexity makes them imperfect, they are not preferable especially in combinatorial techniques such as in ensemble methods. In addition, they are not applicable to data with small number of samples due to overfitting. Embedded methods, like wrappers, are specific to the model hence may give a different feature subset for the same dataset. The main drawback behind such methods is their inability to remove redundant features and retain informative features efficiently.

3.2. Feature Selection Through Feature Grouping

In this section, we will categorize FS approaches based on feature grouping under supervised, unsupervised and semi-supervised context. Supervised FS utilizes data labels to measure importance and relevance of features. Unsupervised FS, on the other hand, assesses feature relevance by exploiting natural structure of the data without using the class label. Semi-supervised FS benefits from both labeled and unlabeled data. **Fig. 2** illustrates a taxonomy of grouping-based FS approaches covered in this study. A typical scenario in FS approaches based on grouping is that the features are first partitioned into clusters and then (a) representative feature(s) is (are) selected from each cluster according to a specific metric or technique as shown in **Fig. 3**.

3.2.1. Grouping-based Feature Selection under Supervised Setting

In the literature, there are many studies that conducted FS through feature grouping. The grouping of features is performed by various techniques including K-means [43], hierarchical clustering [44,45], affinity propagation [46], graph theories [47], information theory metrics [48], kernel density estimation [49], logistic regression [50] and regularization methods [51]. With the availability of class labels in datasets, this prevalence is increasing day by day, offering new approaches and gaining new insights into the field.

Several studies performed K-means or hierarchical clustering for grouping features and then they chose genes from each cluster. Sahu et al. [52] proposed an ensemble approach where K-means is applied first for feature grouping and then three different filter based ranking techniques (t-test, signal-to-noise ratio and SAM) are implemented for each cluster independently; and the feature in the front rank from each cluster is selected to form three distinct feature subsets. Afterwards, features in subsets are subject to additional elimination by checking the inclusion of each feature in other subsets. In other words, a feature is discarded if it is not available in other subsets. They obtain good accuracy for different combinations in general but this study ignores correlations between genes. Another study [53] applied information compression index to group features by hierarchical clustering and then sorted features within each cluster by Fisher criterion measuring the classification capacity of each feature in a cluster. Subsequently, the feature in the front rank is selected for each cluster to form the feature subset.

Regarding selection of features from groups, in addition to ranking, selection can also be performed sequentially. For instance, Zhu and Yang [54] group features into clusters by a modified affinity propagation algorithm, and then they apply sequential FS for each cluster. Later on, they gather selected features in clusters to acquire the reduced subset. Their experimental results show improvement in execution time and the accuracies are comparable with sequential FS. Alimoussa et al. [55] proposed a sequential FS method based on feature grouping mainly consisting of three steps. They first remove irrelevant features using Pearson correlation. Then, the same correlation metric is employed for grouping of features into clusters by considering intercorrelated features directly or indirectly (via other features). Finally, a feature from each cluster is selected sequentially and features belonging to the same cluster are removed in each round. Their proposed method gives better accuracy and reduction in size compared to filter and wrapper methods. However, despite their approach being fully filter-based, execution time of the proposed method is moderate due to the grouping procedure. In their other work for color texture classification [56], they incorporate a classifier into their previous work in order to measure accuracy when a feature is added at each step of their procedure, thereby determining the dimensionality of the feature subset. They show that combining several descriptor configurations performs better compared to a predefined configuration.

Au et al. [57] proposed an effective algorithm called ACA for gene expression data analysis. This algorithm uses an information measure to quantify correlation between features, and performs K-mode algorithm, similar to K-means, to cluster features. They defined mode of each cluster as the attribute (feature) with the highest sum of relevancy with others in each feature group. These modes constituted the final reduced subset. Their measure was also utilized to get good clustering configurations automatically. Chitsaz et al. [58] presented a fuzzy variant of this study which relies on the basic underlying idea in fuzzy clustering approaches, that each feature may belong to more than one group. Rather than considering association of each feature with a sole cluster, association with all features among the overall clusters is considered by assigning different grades of membership to features. Their extended work [59] integrates chi-square test to assess the dependency of each feature on the class labels during the FS process. In their method, objective function is computed by the following formula

$$J = \sum_{r=1}^{k} \sum_{i=1}^{p} u_{ri}^{m} R(A_i : \eta_r) \quad (12)$$

where k and p designate number of clusters and features, respectively and u_{ri} is membership degree of i^{th} feature in r^{th} cluster and m is a weighting exponent with η_r being the mode of r^{th} cluster which is essentially center of that cluster. R function denotes interdependence measure between feature A_i and mode η_r. Their experimental results achieve improvement in the accuracy of the classifier with significant reduction in selected feature size compared to the basic version.

Graph-based approaches are also common in studies involving FS through grouping. Song et al. [25] proposed an algorithm, called FAST, and benefited from minimum spanning trees (MST) to create feature clusters. They adopted symmetric uncertainty to determine relevance between any pair of features or between the feature and the target class. Finally, the feature with the highest correlation with the class label is selected from each cluster. Another study [60] under supervised framework similarly used MST for grouping and variation of information for relevance measure. Desired number of features and the pruning rate should be given as inputs in their algorithm. A recent study by Zheng et al. [61] builds the graph by interaction gain , makes use of MST to produce feature groups and probabilistic consistency measure for quality metric including two different techniques for FS: in the first one, they apply the conventional way of selecting representatives from each feature groups; and in the second they use harmony search as a metaheuristic search. The metaheuristic approach dominates their first proposed algorithm together with other search mechanisms. Quite recently, the study proposed by Wan et al [62] employs graph theory for feature grouping and selection in a fuzzy space. They initially construct the fuzzy space using NA-β-PFRS and then constitute feature groups using MST and acquire the final subset considering feature-to feature and feature-to class relevance in the space. They achieve slightly better results in accuracy with reduced number of features in comparison with other FS approaches and they also show robustness of their model.

Speaking of metaheuristic, García-Torres et al. [63] employed Markov blanket for clustering features and then these predominant groups are involved in Variable Neighborhood Search (VNS) metaheuristic. Their algorithm yields competitive results in classifier performance and exhibits effective results in terms of number of features and running time. Another optimization-based approach in [64] adopted a Scatter Search (SS) strategy based on feature grouping where GreedyPGG [63] is used to group features. In their metaheuristic approach, each solution generated by the search is enhanced with sequential forward selection for selection of the reduced set of features. Their experimental work shows comparable classification results with SS but a significant reduction in feature subset size. Song et al. [65] presents a three-step hybrid study for high dimensional data. Their work initially removes irrelevant features with SU by a predetermined threshold ρ_0 which is defined as

$$\rho_0 = \min\,(0.1 * \ SU_{max}, SU_{\lfloor D/\log D \rfloor - th}) \quad (13)$$

where SU_{max} is the maximal relevance value between a feature and class labels among all D features. Secondly, it constitutes feature groups using a SU-based clustering approach in which cluster centers are chosen at first and initial number of clusters is not required. As the third step, representative features are selected from clusters based on particle swarm optimization (PSO) with global search capability. Their proposed methodology yields comparative results with respect to accuracy and running time. García-Torres et al. extended their previous SS work in [66], integrating an additional stopping criterion into their algorithm along with hyperparameter tuning. Their experimental results present the effectiveness of the additional stopping condition with respect to the computing time, and also exhibit similar classifier performance with highly reduced size of feature subset among other evolutionary and popular approaches.

Although many studies focused their attention on discriminative power and redundancy removal of features, most of them neglect the stability of the selected features. Yu et al. addressed this issue in their two studies [49,67]. In [49], rather than relying on typical clustering algorithms, they applied kernel density estimation accompanied by an iterative mean shift procedure to find feature clusters. Subsequently, these feature clusters were evaluated according to relevance using F-statistic and a representative feature is selected within each cluster. The same authors extended this study in [67], where consensus feature groups were identified in an ensemble learning manner and features were extracted in the same way as their first study. The experiments conducted in both studies showed the stability of the selected features.

All the works mentioned until now are considered as global FS, i.e., finding a reduced subset of global features for the entire population. However, there are cases where these approaches are not applicable. For instance, take an image recognition task, where feature importance may alter since a set of relevant features may be important for identifying a specific object but insignificant for another object at a different position. This gap paved the way for a different technique, called Instance-wise FS that associates each feature's relationship to its labels by assigning a different selector for each instance. Interested readers to grouping and selection of features in this approach can refer to [68,69]. A summary of above-mentioned approaches under the supervised framework is outlined in **Table I**.

FS approaches based on grouping are not necessarily in the manner of grouping features into clusters and choosing representatives. Distinctly, selection of the features may happen with different cluster configurations. Moshlei et al. [71] initially implement K-means for clustering all samples for a given dataset and a sample from each cluster is chosen at random to acquire the samples with the greatest differences for the preliminary dataset. Subsequently, variances of all features on the determined samples are calculated and a predefined number of features with the highest variances are selected, thereby forming the primary dataset. Thereafter, remaining features are added gradually to this dataset and K- means clustering (with a predefined number of clusters) is applied iteratively in each step. Features causing changes in the structure of clusters are observed in a repetitive manner and considered as significant. Other features that don't lead to any alteration in clusters are eliminated.

Another work by Yousef et al. [72] introduced *"recursive cluster elimination*" term into the community and their approach was later adopted in many studies. Since this approach was widely employed by different studies, in Section 4 we elaborate this method in detail by reviewing its application areas and modified usages.

3.2.2. Grouping-based Feature Selection under Unsupervised Setting

As with the traditional methods in FS, many of feature grouping-based FS approaches belong to the supervised learning paradigm. Unsupervised FS is more challenging than supervised FS because of no prior knowledge about class labels and unknown number of clusters. Unsupervised FS methods typically involve i) maximization of clustering performance by some index or ii) selection of features based on dependency. Since this paper is about FS, first one is out of scope for this study. Many statistical dependency/distance measures are available in the literature including correlation coefficient, least square regression error, Euclidean distance, entropy, and variance. Selected features in unsupervised FS methods can be evaluated in terms of both classification performance and clustering performance. **Table II** summarizes works on unsupervised FS based on grouping.

Mitra et al. [73] proposed an unsupervised FS algorithm using feature similarity. A new similarity measure called *maximum information compression index* is introduced in their study. Also, they demonstrated use of representation entropy for measuring redundancy and information loss quantitatively. Features are partitioned into clusters using k-NN principle along with a similarity measure. Entropy metric is chosen as the FS criterion and applied to select a single feature from each cluster to constitute the reduced subset. To evaluate the effectiveness of selected features, the proposed method is compared with KNN, Naive Bayes and class separability (including Relief-F) for classification capability, and with entropy and fuzzy feature evaluation index for clustering performance. Their algorithm is rapid since no search is required and hence their study is one of the state of the art work in the literature.

Another example is the study of Li et al. [74], which uses the same similarity measure in [73] and employs a distance function to obtain clusters of features. A representative feature, having the shortest distance to others within a cluster, is selected from each cluster. Their approach is based on hierarchical clustering which enables them to choose feature subsets with different sizes by choosing from top clusters in the hierarchy. Their algorithm works for both unsupervised and supervised learning tasks. Moreover, they run clustering just one time in their algorithm. The authors presented their experimental results for both clustering and classification.

As stated previously, FS methods developed under unsupervised framework do not utilize class labels. As an example, Covões T.F. et al. [75] presents a comparative study of their approach with the algorithm proposed by Mitra et al [73]. Again, maximal information compression index is utilized to find clusters of features. Hereafter, they employed the simplified silhouette (SS) criterion to find optimum clusters, allowing to find the number of clusters as well. The computation for simplified silhouette depends only on obtained partitions, and it is not dependent on any clustering algorithm. Hence, this silhouette is, not only determines the number of clusters automatically, but also it is capable of evaluating partitions acquired by any clustering algorithms. They employed the k-medoids algorithm along with the silhouette method in order to achieve optimum clusters. Then the corresponding medoid for each cluster is selected as the representative feature. The prerequisite for number of clusters known a priori in this algorithm has been overcome by SS since one can implement this algorithm for different values of number of clusters, and then select the best clustering according to the maximum value obtained in SS.

Another study under unsupervised framework is suggested in [76], where maximal information coefficient and affinity propagation are exploited for selection of features. Features are chosen as the centroid of each cluster in the final step. Although they present competitive results in classification with typical classifiers, no comparison is made for clustering.

FS methods developed under supervised framework can be an inspiration to unsupervised studies. For instance, Zhou et al. [77] developed an attribute (feature) clustering algorithm along with an FS method in an unsupervised manner. They test their algorithm considering different FS methods with different classifiers and achieve slightly improved mean accuracies. The unsupervised FS algorithm proposed by Zhu et al. [78] groups features according to their SU similarities. In their clustering approach, cluster centers are firstly determined and the features are assigned to these centers subsequently. Then, the feature with the highest SU on average is selected from each cluster as a representative based on the following formula

$$AR(f,C) = \frac{\sum_{i=1}^{|C|} SU(f,f_i)}{|C|} \tag{14}$$

where $AR(f,C)$ is the average redundancy for a feature f in cluster C and $f_i \in C$. Their experiments showed that compared to other methods, the proposed algorithm performs more efficiently in terms of running time and in terms of the size of the reduced subset of features. Also, clustering performance of their algorithm surpasses the compared techniques for various clustering performance measurements. Apart from this, a recent hybrid work which is a combination of grouping and binary ant system can be found in [79].

More recently, Yuan et al. formulated this phenomenon as an optimization problem [80], where their optimization benefits from feature grouping and orthogonal constraints. Clustering performance of their algorithm shows better performance in general compared to other unsupervised FS methods.

3.2.3. Grouping-based Feature Selection under Semi-supervised Setting

There are cases when a significant amount of data is unlabeled and only few samples are labeled. In such a case, the learning problem is denominated as semi-supervised. Quinzán et al. [81] conducted a grouping-based FS study under this setting. In their study, the distance measure between each pair of features is computed by both conditional entropy and conditional mutual information. Next, hierarchical clustering is applied to attain feature clusters and the feature with the highest MI is selected as the representative inside each cluster. They test the performance of their algorithm for a different number of labeled samples with other algorithms and their results exhibit satisfactory performance when there is not enough labeled data. Semi-supervised FS techniques are common in the literature and reviewed in many studies [82–84].

4. Feature Grouping with Recursive Cluster Elimination

In the original framework [72], the first step in SVM-RCE is to group genes into clusters using K-means in which correlated gene clusters are identified. As the second step, SVM is used to score (rank) these clusters and finally clusters with low scores are eliminated. Remaining genes (features) in clusters are combined and then clustering along with SVM is applied iteratively until a predefined number of clusters are left. In each iteration, surviving genes are used for classification to measure the accuracy at each level. Interests in this method have grown rapidly over time and many studies conducted their research via integrating this approach.

Weis et al. [85] presented a SVM-RCE-like approach where they included assessment of clusters collaboratively rather than evaluating clusters individually. The study of Deshpande et al. [86] utilized SVM-RCE (although they call it RCE-SVM in their paper) with small modifications for brain state classification.

Another study by Luo [87] aimed to reduce the computational complexity of SVM-RCE. They apply infinite norm of weight coefficient vector from the SVM model to score each cluster instead of scoring clusters by cross-validation. Their results show considerable reduction in computation time while exhibiting comparative performance as SVM-RCE.

In the study associated with military service members, in addition to the statistical significance test, SVM-RCE is used to classify individuals between posttraumatic stress disorder (PTSD), postconcussion syndrome (PCS) + PTSD, and controls [88]. In their study, the features refer to the connectivity paths acquired from 125 brain regions. In their experimental works using SVM-RCE, they conclude that higher classification rate (by 4%) is achieved through imaging-based grouping than conventional grouping. Furthermore, imaging measures dominate non-imaging measures by 9% for both conventional and imaging-based groupings.

Jin et al. [89] conducted a similar study and adopted a modified version of SVM-RCE in their study of brain connectivity. In their study, the diagnostic label of a novel subject is tested whether it belongs to subjects with PTSD or healthy group. The connectivity features are measured from mean resting-state time series taken from 190 regions across the entire brain. They employ SVM-RCE in their experimental work to suggest that dynamic functional and effective connectivity gives higher classification results compared to their static counterparts.

Interestingly, Zhao et al. [90] applied SVM-RCE tool to the detection of expression profiles identifying microRNAs related to venous metastasis in hepatocellular carcinoma.

Chaitra et al. [91] conducted a study to identify biomarkers of autism spectrum disorder (ASD) using imaging datasets. They utilized SVM-RCE to assess the classification performance for three distinct feature sets consisting of connectivity features alone, complex network (graph) measures alone, and a feature set including both. Their accuracy results are not competitive; however, the emphasis is on assessing different feature sets, especially on the combined feature set.

5. Grouping Features with Biological Domain Knowledge

Aforementioned FS approaches typically use some statistics and computational algorithms to group and select the features without any domain knowledge. However, specifically in bioinformatics, integration of biological knowledge is influential to improve the process of gene selection, patient stratification, and disease diagnosis [92]. The general idea in integrating biological knowledge to FS is to first apply a biological function for grouping the genes, and then give each group a rank by scoring them using a machine learning algorithm. Finally, genes or biological entities in the top groups form the reduced subset of features. We would like to note that this section is especially designed for researchers working in the field of molecular biology, genetics, bioinformatics; and we believe that this section is especially informative for those with a biological background.

An integrative approach presented by Qi and Tang integrates Gene Ontology (GO) annotations into gene selection process, where they start by finding discriminative score for each gene (feature) applying Information Gain (IG) and eliminating those with a score of zero [93]. The next step is to annotate these genes with GO terms. After that, the score of each term is calculated as the mean of discriminative scores of associated genes involved in the respective term. The GO term with the highest score is determined and the most discriminative associated gene is selected and extracted. The steps including calculation of scores for GO terms and selection of next most informative gene is repeated until the final subset completion. Their comparative work with sole IG shows the effectiveness of GO integration in the gene selection process.

Another integrative approach, Support Vector Machines with Recursive Network Elimination (SVM-RNE), was proposed in [94], which was an extension of SVM-RCE. Similarly, genes are grouped into clusters by GXNA [95] and clusters with low scores are eliminated at each iteration. The algorithm terminates when some predefined constraints on the number of groups are met.

In SoFoCles [96], genes are initially ranked by typical filter methods such as information gain, Relief-F or χ^2 and then a reduced subset of genes is created by a given threshold. Next, for each gene in the reduced subset, semantically similar genes from GO are determined. Finally, top semantically similar genes are selected to enrich the reduced subset. Experimental works conducted using SoFoCles reveal enhancement in classification results by integrating biological knowledge into gene selection.

Mitra et al. [97] adopted CLARANS for gene (feature) clustering via gene ontology (GO) analysis. The final reduced feature subset is composed of genes which were medoids of biologically enriched clusters. In their experiments, incorporation of biological knowledge enhanced classifier performance and reduced computational complexity. The same authors subsequently made use of a fuzzy technique, FCLARANS, to obtain clusters and selected representative genes from clusters by fold change [98].

The study suggested by Fang et al.[99] includes a combination of both KEGG and GO terms with IG. IG is applied on the initial dataset as filtering and then GO and KEGG annotations are explored for the remaining genes. As the next step, association mining is applied to this annotation information and the interestingness of the frequent itemsets is determined by averaging the original discriminative scores (from IG) of the involved genes. The final gene set is attained via the selection of the highest ranked genes from the top n frequent itemsets. They assessed their method using GO, KEGG, and both against IG and study of [93]. Despite the lower rate of improvement in the overall accuracy, they are able to achieve it with a significant reduction in the number of genes.

Raghu et al. [100] utilize KEGG, DisGeNET and other genetic meta information in their integrated approach. Two metrics, gene importance and gene distance, are computed in their framework. Importance score for each gene is calculated using DisGeNET, which is a public platform containing gene collections associated with diseases. Distance between genes is computed based on their chromosomal locations and associations to the same diseases. Both scores are then employed to compose gene sets with maximum relevance and diversity. Compared to variance-based techniques, their method performs better in predictive modeling task on a small scale.

Perscheid et al. [101] makes a comparison between traditional and knowledge-based gene selection methods applied on gene expression data. Their approach produces gene rankings by integrating knowledge bases and each of these rankings are evaluated with a predefined number of selected genes. Finally, the ranking with the best performance is selected. Moreover, they proposed a framework allowing external knowledge utilization, gene selection and evaluation in an automatic fashion. Although the framework seems to be knowledge base dependent, their experimental results demonstrate that incorporating biological knowledge into gene selection process upgrades performance in classification, decreases computational runtime, and enhances stability of selected genes.

Yet another study developed maTE [102], where gene groups are produced based on the miRNA target information and then each group is ordered by cross-validation. The average accuracy after a specific number of iterations determines the rank of each cluster. Genes on the top m groups are selected as the reduced subset.

The integrative FS method through grouping proposed by Yousef et al. [103] benefits from the biological knowledge for ranking and classification steps. Their proposed framework, named CogNet, initially implements pathfindR [104] to group the genes for clustering. These cluster groups are actually enriched KEGG pathways as a result of enrichment analysis. Then, a new dataset involving genes for the specific pathway is created for each cluster (pathway). These datasets are scored through Monte Carlo cross-validation (MCCV) and pathways are ranked according to the assigned scores. Ultimately, genes found in top chosen pathways are taken as features and used for classification.

Another study, called miRcorrNet [105], finds gene groups on the basis of their correlation to miRNA expression. Afterwards, these groups are subject to a ranking function for classification. The results showed AUC scores above 95%, proving that miRcorrNet is capable of prioritizing pan-cancer-regulating high-confidence miRNAs.

Very recently Zhang et al. [106] proposed a method DCG-Net; where they quantify distance correlation gain between features to construct the biological network. In their algorithm, a greedy search method is applied to detect network modules. The edge with the highest weight is selected, then this edge is extended with respect to correlation metric to obtain the module in the network. This is done iteratively to extract modules and the module with the highest distance correlation is selected for analysis. Their experimental results showed effective results in terms of FS and classification accuracy.

6. Discussion

As stated previously, FS based on feature grouping is a powerful technique with important advantages. Next, one may wonder which FS technique is the best in this context. Surely, it's hard to answer this question because the concept of FS is not dependent only on one parameter. The intrinsic structure and size of the dataset, the learning model and the selected parameters are known as effective factors in the field. In this section we make a cross-comparison and share our deductions among the approaches we have examined in the literature.

We mentioned before that a typical approach in grouping-based FS is to select representative features from groups. However, selection of multiple representatives from groups may enhance the classifier performance as shown in [107], where selection of the most independent feature with the representative along with the representative in each group improved the accuracy results.

The superiority of feature grouping is apparent in sequential-based FS because once a feature is selected, features of the same cluster can be discarded at each iteration, thereby diminishing search complexity in total. We particularly want to emphasize here that sequential-based FS approaches generally employ wrapper models which cause huge running time. We motivate researchers for filter-based sequential FS techniques since such an approach benefits both from the strength of feature grouping and from the high speed of filter models as presented in [55,56]. Dominance of this approach over deep learning algorithms can be seen in [56]. As a result, sequential approaches are effective in the field since they consider interactivity between features and are also used during subset search in evolutionary approaches [63,66].

Fuzzy approaches for FS based on grouping are effective because features can belong to more than one cluster rather than typical assignment of a feature to a specific cluster, which can improve the subset quality and accuracy. We should also say that feature-class relevance is an important metric in supervised setting for fuzzy or other approaches and importance of its utilization is specified in [59]. On the other hand, evolutionary algorithms such as genetic algorithms can be implemented as subset search algorithms during the selection process [108]. These approaches outperform the conventional way of selecting representatives due to inclusion of inter-feature collaboration as shown in [61]. Challenge for these algorithms is mainly high computational cost. A comparison of fuzzy and evolutionary approaches is available in [70], where both methods obtain similar accuracies but the proposed fuzzy technique dominates others in terms of running time and subset quality.

Incorporating different techniques can increase the strength of an approach rather than sticking to a specific one alone. For instance, the study of [62] combines the advantages of fuzziness, graph theory and conditional mutual information, and acquires better results in general than graph based or fuzzy approaches.

As implied in Section 5, integrative gene selection is an important matter when biological data is considered since statistical methods lack the ability to identify the underlying biological processes. Effectiveness of integrating domain knowledge from external sources is reviewed in [101].

FS methods based on deep learning (DL) are common in the literature [109-111] but these methods adopt feature extraction, i.e., transformation of the original feature space into a reduced size of new features which leads to loss of original semantics of features. In short, they provide competitive class accuracies but are far from interpretability [112].

Despite the plenitude of FS techniques, there's still room for further progress in this field. The current studies are mostly based on pairwise interactions; whereas interactions of multiple features should be explored. In addition, running time is still a barrier, and especially for complex algorithms smart steps should be taken on it.

7. Conclusions

The advances in high-throughput technologies have generated large high-dimensional data sets in many applications. The inevitable presence of redundant and noisy features increases computational complexity and degrades classifier capability. Hence, FS has become a required pre-processing step in itself as a primary concern for a long time. Here we present works done in the literature regarding FS techniques through feature grouping. Feature grouping is a powerful and efficient concept; it reduces search space and complexity, is resistant to the variations of samples, gives lower levels internal redundancy and provides better generalization capability to the classifier. The form of feature grouping and selection of features out of groups are determined by different metrics or techniques as reviewed in this paper.

In FS based on feature grouping, the aim is to first keep similar features together within clusters while maximizing diversity between clusters followed by selection of features out of clusters. We can conclude that sequential and optimization-based (fuzzy and evolutionary) FS approaches are noteworthy in this context since they take feature interactivity into consideration during the selection phase. Hybrid approaches or utilizing a combination of different techniques are also effective because each method brings its advantage. In case of biological data, integrating external knowledge can yield better results in the overall analysis. In fact, availability of independent and relevant features, correlation between features, and feature correlation to the decision are important items to be taken into consideration. The models with the ability to take these factors into consideration are likely to be effective in FS.

In this study, our goal is to inform interested readers about the recent trends in FS by feature grouping. Despite the wealth of many techniques in this field, there is still need for enhancement and novelty in the area. We believe approaches mentioned here may provide new insights into designing new schemes for FS in terms of better efficiency, effectiveness, stability, generalization and discrimination.

References

[1] B. Venkatesh and J. Anuradha, "A Review of Feature Selection and Its Methods," *Cybern. Inf. Technol.*, vol. 19, no. 1, pp. 3–26, Mar. 2019, doi: 10.2478/cait-2019-0001.

[2] A. Jovic, K. Brkic, and N. Bogunovic, "A review of feature selection methods with applications," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, May 2015, pp. 1200–1205. doi: 10.1109/MIPRO.2015.7160458.

[3] Md. Mehedi Hassan, S. Mollick, and F. Yasmin, "An unsupervised cluster-based feature grouping model for early diabetes detection," *Healthc. Anal.*, vol. 2, p. 100112, Nov. 2022, doi: 10.1016/j.health.2022.100112.

[4] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant Features and the Subset Selection Problem," in *Machine Learning Proceedings 1994*, Elsevier, 1994, pp. 121–129. doi: 10.1016/B978-1-55860-335-6.50023-4.

[5] J. Wang, X. Wu, and C. Zhang, "Support vector machines based on K-means clustering for real-time business intelligence systems," *Int. J. Bus. Intell. Data Min.*, vol. 1, no. 1, pp. 54–64, Jan. 2005, doi: 10.1504/IJBIDM.2005.007318.

[6] L. Maokuan, C. Yusheng, and Z. Honghai, "Unlabeled data classification via support vector machines and k-means clustering," in *Proceedings. International Conference on Computer Graphics, Imaging and Visualization, 2004. CGIV 2004.*, Jul. 2004, pp. 183–186. doi: 10.1109/CGIV.2004.1323982.

[7] X. Shen and H.-C. Huang, "Grouping Pursuit Through a Regularization Solution Surface," *J. Am. Stat. Assoc.*, vol. 105, no. 490, pp. 727–739, Jun. 2010, doi: 10.1198/jasa.2010.tm09380.

[8] M. Mittal, L. M. Goyal, D. J. Hemanth, and J. K. Sethi, "Clustering approaches for high-dimensional databases: A review," *WIREs Data Min. Knowl. Discov.*, vol. 9, no. 3, May 2019, doi: 10.1002/widm.1300.

[9] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, Jan. 2014, doi: 10.1016/j.compeleceng.2013.11.024.

[10] Y. Dai, Z. Gao, Y. Zhu, W. Zhang, H. Li, Y. Wang, Z. Li, "Feature Grouping for No-reference Image Quality Assessment," in *2022 7th International Conference on Automation, Control and Robotics Engineering (CACRE)*, Xi'an, China, Jul. 2022, pp. 204–208. doi: 10.1109/CACRE54574.2022.9834184.

[11] Ravishanker, M. Sood, P. Angra, S. Verma, Kavita, and N. Z. Jhanjhi, "Efficient Feature Grouping for IDS Using Clustering Algorithms in Detecting Known/Unknown Attacks," in *Information Security Handbook*, CRC Press, 2022.

[12] A. N. M. B. Rashid, M. Ahmed, L. F. Sikos, and P. Haskell-Dowland, "Cooperative co-evolution for feature selection in Big Data with random feature grouping," *J. Big Data*, vol. 7, no. 1, p. 107, Dec. 2020, doi: 10.1186/s40537-020-00381-y.

[13] B. Bakir-Gungor, H. Hacılar, A. Jabeer, O. U. Nalbantoglu, O. Aran, and M. Yousef, “Inflammatory bowel disease biomarkers of human gut microbiota selected via different feature selection methods,” *PeerJ*, vol. 10, p. e13205, Apr. 2022, doi: 10.7717/peerj.13205.

[14] Y. Li, U. Mansmann, S. Du, and R. Hornung, “Benchmark study of feature selection strategies for multi-omics data,” *BMC Bioinformatics*, vol. 23, no. 1, p. 412, Oct. 2022, doi: 10.1186/s12859-022-04962-x.

[15] T. Bhadra, S. Mallik, N. Hasan, and Z. Zhao, “Comparison of five supervised feature selection algorithms leading to top features and gene signatures from multi-omics data in cancer,” *BMC Bioinformatics*, vol. 23, no. S3, p. 153, Mar. 2022, doi: 10.1186/s12859-022-04678-y.

[16] G. Manikandan and S. Abirami, “Feature Selection and Machine Learning Models for High-Dimensional Data: State-of-the-Art,” in *Computational Intelligence and Healthcare Informatics*, 1st ed., O. P. Jena, A. R. Tripathy, A. A. Elngar, and Z. Polkowski, Eds. Wiley, 2021, pp. 43–63. doi: 10.1002/9781119818717.ch3.

[17] B. Remeseiro and V. Bolon-Canedo, “A review of feature selection methods in medical applications,” *Comput. Biol. Med.*, vol. 112, p. 103375, Sep. 2019, doi: 10.1016/j.compbiomed.2019.103375.

[18] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” *Artif. Intell.*, vol. 97, no. 1–2, pp. 273–324, Dec. 1997, doi: 10.1016/S0004-3702(97)00043-X.

[19] M. A. Hall and L. A. Smith, *Practical feature subset selection for machine learning*, vol. Volume 20 No 1. Springer, 1998, pp. 181–191.

[20] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes 3rd Edition: The Art of Scientific Computing*, 3rd edition. Cambridge, UK ; New York: Cambridge University Press, 2007.

[21] D. Nettleton, *Commercial Data Mining: Processing, Analysis and Modeling for Predictive Analytics Projects*, 1st edition. Amsterdam: Morgan Kaufmann, 2014.

[22] Huan Liu and R. Setiono, “Chi2: feature selection and discretization of numeric attributes,” in *Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence*, Herndon, VA, USA, 1995, pp. 388–391. doi: 10.1109/TAI.1995.479783.

[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 1st ed. Wiley, 2005. doi: 10.1002/047174882X.

[24] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd edition. Burlington, MA: Morgan Kaufmann, 2011.

[25] Qinbao Song, Jingjie Ni, and Guangtao Wang, “A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data,” *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 1, pp. 1–14, Jan. 2013, doi: 10.1109/TKDE.2011.181.

[26] S. Visalakshi and V. Radha, “A literature review of feature selection techniques and applications: Review of feature selection in data mining,” in *2014 IEEE International Conference on Computational Intelligence and Computing Research*, Coimbatore, India, Dec. 2014, pp. 1–6. doi: 10.1109/ICCIC.2014.7238499.

[27] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Boston, MA: Springer US, 1998. doi: 10.1007/978-1-4615-5689-3.

[28] Huan Liu and Lei Yu, “Toward integrating feature selection algorithms for classification and clustering,” *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, Apr. 2005, doi: 10.1109/TKDE.2005.66.

[29] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene Selection for Cancer

Classification using Support Vector Machines," *Mach. Learn.*, vol. 46, no. 1/3, pp. 389–422, 2002, doi: 10.1023/A:1012487302797.

[30] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, H. Liu, "Feature Selection: A Data Perspective," *ACM Comput. Surv.*, vol. 50, no. 6, pp. 1–45, Nov. 2018, doi: 10.1145/3136625.

[31] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *J. R. Stat. Soc. Ser. B Methodol.*, vol. 58, no. 1, pp. 267–288, Jan. 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.

[32] J. Tang, S. Alelyani, and H. Liu, "Feature Selection for Classification: A Review," 2014.

[33] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification And Regression Trees*, 1st ed. Routledge, 2017. doi: 10.1201/9781315139470.

[34] M. Dash and H. Liu, "Feature selection for classification," *Intell. Data Anal.*, vol. 1, no. 1–4, pp. 131–156, 1997, doi: 10.1016/S1088-467X(97)00008-5.

[35] W. Pan, "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments," *Bioinformatics*, vol. 18, no. 4, pp. 546–554, Apr. 2002, doi: 10.1093/bioinformatics/18.4.546.

[36] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. D. Schaetzen, R. Duque, H. Bersini, A. Nowé, "A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 9, no. 4, pp. 1106–1119, Jul. 2012, doi: 10.1109/TCBB.2012.33.

[37] L. Talavera, "An Evaluation of Filter and Wrapper Methods for Feature Selection in Categorical Clustering," in *Advances in Intelligent Data Analysis VI*, vol. 3646, A. F. Famili, J. N. Kok, J. M. Peña, A. Siebes, and A. Feelders, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 440–451. doi: 10.1007/11552253_40.

[38] N. El Aboudi and L. Benhlima, "Review on wrapper feature selection approaches," in *2016 International Conference on Engineering & MIS (ICEMIS)*, Agadir, Morocco, Sep. 2016, pp. 1–5. doi: 10.1109/ICEMIS.2016.7745366.

[39] S. Ma and J. Huang, "Penalized feature selection and classification in bioinformatics," *Brief. Bioinform.*, vol. 9, no. 5, pp. 392–403, Apr. 2008, doi: 10.1093/bib/bbn027.

[40] D. Asir, S. Appavu, and E. Jebamalar, "Literature Review on Feature Selection Methods for High-Dimensional Data," *Int. J. Comput. Appl.*, vol. 136, no. 1, pp. 9–17, Feb. 2016, doi: 10.5120/ijca2016908317.

[41] V. Bolón-Canedo and A. Alonso-Betanzos, "Ensembles for feature selection: A review and future trends," *Inf. Fusion*, vol. 52, pp. 1–12, Dec. 2019, doi: 10.1016/j.inffus.2018.11.008.

[42] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 13, no. 5, pp. 971–989, Sep. 2016, doi: 10.1109/TCBB.2015.2478454.

[43] S. Chormunge and S. Jena, "Correlation based feature selection with clustering for high dimensional data," *J. Electr. Syst. Inf. Technol.*, vol. 5, no. 3, pp. 542–549, Dec. 2018, doi: 10.1016/j.jesit.2017.06.004.

[44] H. Liu, X. Wu, and S. Zhang, "Feature selection using hierarchical feature clustering," in *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*, Glasgow, Scotland, UK, 2011, p. 979. doi: 10.1145/2063576.2063716.

[45] C. H. Park, "A Feature Selection Method Using Hierarchical Clustering," in *Mining*

Intelligence and Knowledge Exploration, vol. 8284, R. Prasath and T. Kathirvalavakumar, Eds. Cham: Springer International Publishing, 2013, pp. 1–6. doi: 10.1007/978-3-319-03844-5_1.

[46] D. Harris and A. Van Niekerk, “Feature clustering and ranking for selecting stable features from high dimensional remotely sensed data,” *Int. J. Remote Sens.*, vol. 39, no. 23, pp. 8934–8949, Dec. 2018, doi: 10.1080/01431161.2018.1500730.

[47] S. Yang, L. Yuan, Y.-C. Lai, X. Shen, P. Wonka, and J. Ye, “Feature grouping and selection over an undirected graph,” in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’12*, Beijing, China, 2012, p. 922. doi: 10.1145/2339530.2339675.

[48] J. Martínez Sotoca and F. Pla, “Supervised feature selection by clustering using conditional mutual information-based distances,” *Pattern Recognit.*, vol. 43, no. 6, pp. 2068–2081, Jun. 2010, doi: 10.1016/j.patcog.2009.12.013.

[49] L. Yu, C. Ding, and S. Loscalzo, “Stable feature selection via dense feature groups,” in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, Las Vegas, Nevada, USA, 2008, p. 803. doi: 10.1145/1401890.1401986.

[50] R. A. Shah, Y. Qian, and G. Mahdi, “Group Feature Selection via Structural Sparse Logistic Regression for IDS,” in *2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, Sydney, Australia, Dec. 2016, pp. 594–600. doi: 10.1109/HPCC-SmartCity-DSS.2016.0089.

[51] S. Petry, C. Flexeder, and G. Tutz, “Pairwise Fused Lasso,” 2011, doi: 10.5282/UBM/EPUB.12164.

[52] B. Sahu, S. Dehuri, and A. K. Jagadev, “Feature selection model based on clustering and ranking in pipeline for microarray data,” *Inform. Med. Unlocked*, vol. 9, pp. 107–122, 2017, doi: 10.1016/j.imu.2017.07.004.

[53] Z. Shang and M. Li, “Feature Selection Based on Grouped Sorting,” in *2016 9th International Symposium on Computational Intelligence and Design (ISCID)*, Hangzhou, Dec. 2016, pp. 451–454. doi: 10.1109/ISCID.2016.1111.

[54] K. Zhu and J. Yang, “A cluster-based sequential feature selection algorithm,” in *2013 Ninth International Conference on Natural Computation (ICNC)*, Shenyang, China, Jul. 2013, pp. 848–852. doi: 10.1109/ICNC.2013.6818094.

[55] M. Alimoussa, A. Porebski, N. Vandenbroucke, R. Thami, and S. El Fkihi, “Clustering-based Sequential Feature Selection Approach for High Dimensional Data Classification:,” in *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, Online Streaming, --- Select a Country ---, 2021, pp. 122–132. doi: 10.5220/0010259501220132.

[56] M. Alimoussa, A. Porebski, N. Vandenbroucke, S. El Fkihi, and R. Oulad Haj Thami, “Compact Hybrid Multi-Color Space Descriptor Using Clustering-Based Feature Selection for Texture Classification,” *J. Imaging*, vol. 8, no. 8, p. 217, Aug. 2022, doi: 10.3390/jimaging8080217.

[57] Wai-Ho Au, K. C. C. Chan, A. K. C. Wong, and Yang Wang, “Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data,” *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 2, no. 2, pp. 83–101, Apr. 2005, doi: 10.1109/TCBB.2005.17.

[58] Chitsaz E., Taheri M., Katebi S.D., “A fuzzy approach to clustering and selecting features for classification of gene expression data”. In: Proc. World Congress of Engineering (WCE 2008), 2008, 1650–1655.

[59] Chitsaz, E., Taheri, M., Katebi, S.D., Jahromi, M.Z.: An Improved Fuzzy Feature Clustering and Selection based on Chi-Squared-Test. In: Proceedings of the International MultiConference of Engineers and Computer Scientists, IMECS 2009, Hong Kong, vol. I (2009)

[60] Q. Liu, J. Zhang, J. Xiao, H. Zhu, and Q. Zhao, “A Supervised Feature Selection Algorithm through Minimum Spanning Tree Clustering,” in *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*, Limassol, Cyprus, Nov. 2014, pp. 264–271. doi: 10.1109/ICTAI.2014.47.

[61] L. Zheng, F. Chao, N. M. Parthaláin, D. Zhang, and Q. Shen, “Feature grouping and selection: A graph-based approach,” *Inf. Sci.*, vol. 546, pp. 1256–1272, Feb. 2021, doi: 10.1016/j.ins.2020.09.022.

[62] J. Wan, H. Chen, T. Li, B. Sang, and Z. Yuan, “Feature Grouping and Selection With Graph Theory in Robust Fuzzy Rough Approximation Space,” *IEEE Trans. Fuzzy Syst.*, vol. 31, no. 1, pp. 213–225, Jan. 2023, doi: 10.1109/TFUZZ.2022.3185285.

[63] M. García-Torres, F. Gómez-Vela, B. Melián-Batista, and J. M. Moreno-Vega, “High-dimensional feature selection via feature grouping: A Variable Neighborhood Search approach,” *Inf. Sci.*, vol. 326, pp. 102–118, Jan. 2016, doi: 10.1016/j.ins.2015.07.041.

[64] M. García-Torres, F. Gómez-Vela, F. Divina, D. P. Pinto-Roa, J. L. V. Noguera, and J. C. M. Román, “Scatter search for high-dimensional feature selection using feature grouping,” in *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, Lille France, Jul. 2021, pp. 149–150. doi: 10.1145/3449726.3459481.

[65] X.-F. Song, Y. Zhang, D.-W. Gong, and X.-Z. Gao, “A Fast Hybrid Feature Selection Based on Correlation-Guided Clustering and Particle Swarm Optimization for High-Dimensional Data,” *IEEE Trans. Cybern.*, vol. 52, no. 9, pp. 9573–9586, Sep. 2022, doi: 10.1109/TCYB.2021.3061152.

[66] M. García-Torres, R. Ruiz, and F. Divina, “Evolutionary feature selection on high dimensional data using a search space reduction approach,” *Eng. Appl. Artif. Intell.*, vol. 117, p. 105556, Jan. 2023, doi: 10.1016/j.engappai.2022.105556.

[67] S. Loscalzo, L. Yu, and C. Ding, “Consensus group stable feature selection,” in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’09*, Paris, France, 2009, p. 567. doi: 10.1145/1557019.1557084.

[68] Q. Xiao, H. Li, J. Tian, and Z. Wang, “Group-Wise Feature Selection for Supervised Learning,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022, pp. 3149–3153. doi: 10.1109/ICASSP43922.2022.9746666.

[69] A. Masoomi, C. Wu, T. Zhao, Z. Wang, P. Castaldi, and J. Dy, “Instance-wise Feature Grouping,” in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 13374–13386. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/9b10a919ddeb07e103dc05ff523afe38-Paper.pdf

[70] R. Jensen, N. M. Parthalain, and C. Cornells, “Feature grouping-based fuzzy-rough feature selection,” in *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Beijing, China, Jul. 2014, pp. 1488–1495. doi: 10.1109/FUZZ-IEEE.2014.6891692.

[71] F. Moslehi and A. Haeri, “A novel feature selection approach based on clustering algorithm,” *J. Stat. Comput. Simul.*, vol. 91, no. 3, pp. 581–604, Feb. 2021, doi: 10.1080/00949655.2020.1822358.

[72] M. Yousef, S. Jung, L. C. Showe, and M. K. Showe, “Recursive Cluster Elimination (RCE) for classification and feature selection from gene expression data,” *BMC Bioinformatics*, vol. 8, no. 1, p. 144, Dec. 2007, doi: 10.1186/1471-2105-8-144.

[73] P. Mitra, C. A. Murthy, and S. K. Pal, “Unsupervised feature selection using feature similarity,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 301–312, Mar. 2002, doi: 10.1109/34.990133.

[74] Guangrong Li, Xiaohua Hu, Xiajiong Shen, Xin Chen, and Zhoujun Li, “A novel unsupervised feature selection method for bioinformatics data sets through feature clustering,” in *2008 IEEE International Conference on Granular Computing*, Hangzhou, Aug. 2008, pp. 41–47. doi: 10.1109/GRC.2008.4664788.

[75] T. F. Covões, E. R. Hruschka, L. N. de Castro, and Á. M. Santos, “A Cluster-Based Feature Selection Approach,” in *Hybrid Artificial Intelligence Systems*, vol. 5572, E. Corchado, X. Wu, E. Oja, Á. Herrero, and B. Baruque, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 169–176. doi: 10.1007/978-3-642-02319-4_20.

[76] X. Zhao, W. Deng, and Y. Shi, “Feature Selection with Attributes Clustering by Maximal Information Coefficient,” *Procedia Comput. Sci.*, vol. 17, pp. 70–79, 2013, doi: 10.1016/j.procs.2013.05.011.

[77] P.-Y. Zhou and K. C. C. Chan, “An unsupervised attribute clustering algorithm for unsupervised feature selection,” in *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Campus des Cordeliers, Paris, France, Oct. 2015, pp. 1–7. doi: 10.1109/DSAA.2015.7344857.

[78] X. Zhu, Y. Wang, Y. Li, Y. Tan, G. Wang, and Q. Song, “A new unsupervised feature selection algorithm using similarity-based feature clustering,” *Comput. Intell.*, vol. 35, no. 1, pp. 2–22, 2019, doi: 10.1111/coin.12192.

[79] Z. Manbari, F. AkhlaghianTab, and C. Salavati, “Hybrid fast unsupervised feature selection for high-dimensional data,” *Expert Syst. Appl.*, vol. 124, pp. 97–118, Jun. 2019, doi: 10.1016/j.eswa.2019.01.016.

[80] A. Yuan, J. Huang, C. Wei, W. Zhang, N. Zhang, and M. You, “Unsupervised Feature Selection via Feature-Grouping and Orthogonal Constraint,” in *2022 26th International Conference on Pattern Recognition (ICPR)*, Aug. 2022, pp. 720–726. doi: 10.1109/ICPR56361.2022.9956408.

[81] I. Quinzán, J. M. Sotoca, and F. Pla, “Clustering-Based Feature Selection in Semi-supervised Problems,” in *2009 Ninth International Conference on Intelligent Systems Design and Applications*, Pisa, Italy, 2009, pp. 535–540. doi: 10.1109/ISDA.2009.211.

[82] Z. Song, X. Yang, Z. Xu, and I. King, “Graph-Based Semi-Supervised Learning: A Comprehensive Review,” *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–21, 2022, doi: 10.1109/TNNLS.2022.3155478.

[83] G. Kostopoulos, S. Karlos, S. Kotsiantis, and O. Ragos, “Semi-supervised regression: A recent review,” *J. Intell. Fuzzy Syst.*, vol. 35, no. 2, pp. 1483–1500, Aug. 2018, doi: 10.3233/JIFS-169689.

[84] R. Sheikhpour, M. A. Sarram, S. Gharaghani, and M. A. Z. Chahooki, “A Survey on semi-supervised feature selection methods,” *Pattern Recognit.*, vol. 64, pp. 141–158, Apr. 2017, doi: 10.1016/j.patcog.2016.11.003.

[85] D. C. Weis, D. P. Visco, and J.-L. Faulon, “Data mining PubChem using a support vector machine with the Signature molecular descriptor: Classification of factor XIa inhibitors,” *J. Mol. Graph. Model.*, vol. 27, no. 4, pp. 466–475, Nov. 2008, doi: 10.1016/j.jmgm.2008.08.004.

[86] G. Deshpande, Z. Li, P. Santhanam, C. D. Coles, M. E. Lynch, S. Hamann, X. Hu, “Recursive Cluster Elimination Based Support Vector Machine for Disease State Prediction Using Resting State Functional and Effective Brain Connectivity,” *PLoS ONE*, vol. 5, no. 12, p. e14277, Dec. 2010, doi: 10.1371/journal.pone.0014277.

[87] Lin-Kai Luo, Deng-Feng Huang, Ling-Jun Ye, Qi-Feng Zhou, Gui-Fang Shao, and Hong Peng, “Improving the Computational Efficiency of Recursive Cluster Elimination for Gene Selection,” *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 8, no. 1, pp. 122–129, Jan. 2011, doi: 10.1109/TCBB.2010.44.

[88] D. Rangaprakash, G. Deshpande, T. A. Daniel, A. M. Goodman, J. L. Robinson, N. Salibi, J. S. Katz, T. S. Denney Jr., M. N. Dretsch, “Compromised hippocampus-striatum pathway as a potential imaging biomarker of mild-traumatic brain injury and posttraumatic stress disorder,” *Hum. Brain Mapp.*, vol. 38, no. 6, pp. 2843–2864, Jun. 2017, doi: 10.1002/hbm.23551.

[89] C. Jin, H. Jia, P. Lanka, D. Rangaprakash, L. Li, T. Liu, X. Hu, G. Deshpande, “Dynamic brain connectivity is a better predictor of PTSD than static connectivity: Dynamic Brain Connectivity,” *Hum. Brain Mapp.*, vol. 38, no. 9, pp. 4479–4496, Sep. 2017, doi: 10.1002/hbm.23676.

[90] X. Zhao, L. Wang, and G. Chen, “Joint Covariate Detection on Expression Profiles for Identifying MicroRNAs Related to Venous Metastasis in Hepatocellular Carcinoma,” *Sci. Rep.*, vol. 7, no. 1, p. 5349, Dec. 2017, doi: 10.1038/s41598-017-05776-1.

[91] N. Chaitra, P. A. Vijaya, and G. Deshpande, “Diagnostic prediction of autism spectrum disorder using complex network measures in a machine learning framework,” *Biomed. Signal Process. Control*, vol. 62, p. 102099, Sep. 2020, doi: 10.1016/j.bspc.2020.102099.

[92] M. Yousef, A. Kumar, and B. Bakir-Gungor, “Application of Biological Domain Knowledge Based Feature Selection on Gene Expression Data,” *Entropy*, vol. 23, no. 1, p. 2, Dec. 2020, doi: 10.3390/e23010002.

[93] J. Qi and J. Tang, “Integrating gene ontology into discriminative powers of genes for feature selection in microarray data,” in *Proceedings of the 2007 ACM symposium on Applied computing - SAC ’07*, Seoul, Korea, 2007, p. 430. doi: 10.1145/1244002.1244101.

[94] M. Yousef, M. Ketany, L. Manevitz, L. C. Showe, and M. K. Showe, “Classification and biomarker identification using gene network modules and support vector machines,” *BMC Bioinformatics*, vol. 10, no. 1, p. 337, Dec. 2009, doi: 10.1186/1471-2105-10-337.

[95] J. Wang, H. Li, Y. Zhu, M. Yousef, M. Nebozhyn, M. Showe, L. Showe, J. Xuan, R. Clarke, Y. Wang, “VISDA: an open-source caBIG^{TM} analytical tool for data clustering and beyond,” *Bioinformatics*, vol. 23, no. 15, pp. 2024–2027, Aug. 2007, doi: 10.1093/bioinformatics/btm290.

[96] G. Papachristoudis, S. Diplaris, and P. A. Mitkas, “SoFoCles: Feature filtering for microarray classification based on Gene Ontology,” *J. Biomed. Inform.*, vol. 43, no. 1, pp. 1–14, Feb. 2010, doi: 10.1016/j.jbi.2009.06.002.

[97] S. Mitra and S. Ghosh, “Feature Selection and Clustering of Gene Expression Profiles Using Biological Knowledge,” *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 42, no. 6, pp. 1590–1599, Nov. 2012, doi: 10.1109/TSMCC.2012.2209416.

[98] S. Ghosh and S. Mitra, “Gene selection using biological knowledge and fuzzy clustering,” in *2012 IEEE International Conference on Fuzzy Systems*, Brisbane, Australia, Jun. 2012, pp. 1–9. doi: 10.1109/FUZZ-IEEE.2012.6250797.

[99] O. H. Fang, N. Mustapha, and Md. N. Sulaiman, “An integrative gene selection with association analysis for microarray data classification,” *Intell. Data Anal.*, vol. 18, no. 4, pp. 739–758, Jun. 2014, doi: 10.3233/IDA-140666.

[100] V. K. Raghu, X. Ge, P. K. Chrysanthis, and P. V. Benos, “Integrated Theory-and Data-Driven Feature Selection in Gene Expression Data Analysis,” in *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, San Diego, CA, USA, Apr. 2017, pp. 1525–1532. doi: 10.1109/ICDE.2017.223.

[101] C. Perscheid, B. Grasnick, and M. Uflacker, “Integrative Gene Selection on Gene Expression Data: Providing Biological Context to Traditional Approaches,” *J. Integr. Bioinforma.*, vol. 16, no. 1, Dec. 2018, doi: 10.1515/jib-2018-0064.

[102] M. Yousef, L. Abdallah, and J. Allmer, “maTE: discovering expressed interactions between microRNAs and their targets,” *Bioinformatics*, vol. 35, no. 20, pp. 4020–4028, Oct. 2019, doi: 10.1093/bioinformatics/btz204.

[103] M. Yousef, E. Ülgen, and O. Uğur Sezerman, “CogNet: classification of gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis,” *PeerJ Comput. Sci.*, vol. 7, p. e336, Feb. 2021, doi: 10.7717/peerj-cs.336.

[104] E. Ulgen, O. Ozisik, and O. U. Sezerman, “pathfindR: An R Package for Comprehensive Identification of Enriched Pathways in Omics Data Through Active Subnetworks,” *Front. Genet.*, vol. 10, p. 858, Sep. 2019, doi: 10.3389/fgene.2019.00858.

[105] M. Yousef, G. Goy, R. Mitra, C. M. Eischen, A. Jabeer, and B. Bakir-Gungor, “miRcorrNet: machine learning-based integration of miRNA and mRNA expression profiles, combined with feature grouping and ranking,” *PeerJ*, vol. 9, p. e11458, May 2021, doi: 10.7717/peerj.11458.

[106] Y. Zhang, X. Lin, Z. Gao, and S. Bai, “A novel method for feature selection based on molecular interactive effect network,” *J. Pharm. Biomed. Anal.*, vol. 218, p. 114873, Sep. 2022, doi: 10.1016/j.jpba.2022.114873.

[107] T. F. Covões and E. R. Hruschka, “Towards improving cluster-based feature selection with a simplified silhouette filter,” *Inf. Sci.*, vol. 181, no. 18, pp. 3766–3782, Sep. 2011, doi: 10.1016/j.ins.2011.04.050.

[108] X. Lin, X. Wang, N. Xiao, X. Huang, and J. Wang, “A Feature Selection Method Based on Feature Grouping and Genetic Algorithm,” in *Intelligence Science and Big Data Engineering. Big Data and Machine Learning Techniques*, vol. 9243, X. He, X. Gao, Y. Zhang, Z.-H. Zhou, Z.-Y. Liu, B. Fu, F. Hu, and Z. Zhang, Eds. Cham: Springer International Publishing, 2015, pp. 150–158. doi: 10.1007/978-3-319-23862-3_15.

[109] M. R. Hassan, S. Huda, M. M. Hassan, J. Abawajy, A. Alsanad, and G. Fortino, “Early detection of cardiovascular autonomic neuropathy: A multi-class classification model based on feature selection and deep learning feature fusion,” Inf. Fusion, vol. 77, pp. 70–80, Jan. 2022, doi: 10.1016/j.inffus.2021.07.010.

[110] N. Hussain, M. A. Khan, U. Tariq, S. Kadry, M. A. E. Yar, A. M. Mostafa, A. A. Alnuaim, S. Ahmad, “Multiclass Cucumber Leaf Diseases Recognition Using Best Feature Selection,” *Comput. Mater. Contin.*, vol. 70, no. 2, pp. 3281–3294, 2022, doi: 10.32604/cmc.2022.019036.

[111] Krell, E., Kamangir, H., Friesand, J., Judge, J., Collins, W., King, S. A., & Tissot, P.

(2022). The influence of grouping features on explainable artificial intelligence for a complex fog prediction deep learning model.

[112] J. Figueroa Barraza, E. López Droguett, and M. R. Martins, “Towards Interpretable Deep Learning: A Feature Selection Framework for Prognostics and Health Management Using Deep Neural Networks,” *Sensors*, vol. 21, no. 17, p. 5888, Sep. 2021, doi: 10.3390/s21175888.

Table 1(on next page)

Applications of FS by Grouping under Supervised Context

Table I. Applications of FS by Grouping under Supervised Context

Grouping Method		FS Method (metric)	FS Strategy	Validation	Types of Data	Study
K-means		correlation	selection of features from front rank	classification accuracy	text and microarray	[43]
		SNR, SAM, t-test	checking existence of a feature in other subsets	LOOCV	microarray	[52]
Hierarchical		Fisher	selection of features from front rank	classification accuracy	miscellaneous	[53]
		average similarity	choosing representative in each group	cross validation	miscellaneous	[45]
Sequential	Correlation-based	trace criterion	features are added sequentially only when trace is maximum.	cross validation	color texture	[56]
	Modified Affinity Propagation	sequential feature selection	applying sequential search in each group and merging selected features	cross validation	miscellaneous	[54]

ACA		interdependence mesure	selection of mode of each cluster	classification accuracy	synthetic & gene expression	[57]
Fuzzy	Correlation	fuzzy-rough subset evaluation	selection of representative features among groups in the fuzzy environment	classification accuracy	miscellaneous	[70]
	Fuzzy ACA	fuzzy multiple interdependence redundancy		classification accuracy	miscellaneous	[59]
		fuzzy multiple interdependence redundancy		classification accuracy	microarray	[58]
Graph-based		neighborhood adaptive fuzzy mutual information	using feature-to-feature & feature-to-class relevance	cross validation	publicly available datasets	[62]*
		probabilistic consistency	i) choosing representative in each group ii) metaheuristic search	cross validation	miscellaneous	[61]

		variation of information	choosing representative in each group	silhoutte index & classification accuracy	miscellaneous	[60]
		symmetric uncertainty	choosing representative in each group	classification accuracy	miscellaneous	[25]
Evolutionary	GreedyPGG	SS	using SS to find subset of features	cross validation	gene expression & text-mining	[66]
	SU-based	PSO	adopting PSO to determine final subset	cross validation	miscellaneous	[65]
	GreedyPGG	SS	using SS to find subset of features	cross validation	biomedical datasets	[64]
	GreedyPGG	VNS	utilizing VNS to decide reduced subset	cross validation	microarray & text-mining	[63]

Table 2(on next page)

Applications of FS by Grouping under Unsupervised Context

1 **Table II. Applications of FS by Grouping under Unsupervised Context**

Grouping Method	FS Method (metric)	FS Strategy	Validation	Types of Data	Study
K-means	generalized incoherent regression model	grouping and selection of optimal features based on orthogonal constraints	unsupervised clustering accuracy (ACC) & normalized mutual information (NMI)	face image & biological datasets	[80]
Louvain community detection	binary ant system (BAS)	features in each group are sorted by modified BAS and best features are selected iteratively	classification error rate (CER)	real-world datasets	[79]
SU-based	SU	feature with the highest SU on average is chosen as representative in each cluster	scatter separability criterion, random adjust index, normalized mutual information, F-score	miscellaneous	[78]
K-mode	mode	selection of mode of each cluster	classification accuracy	miscellaneous	[77]

Affinity Propagation	MICAP	centroid of each cluster is selected for final subset	classification accuracy	miscellaneous	[76]
k-medoids	SSF	medoid of each cluster is chosen as the representative feature	classification accuracy	miscellaneous	[75]
hierarchical	FSFC	feature with the shortest distance to others is selected in each cluster	Minkowski Score	public gene datasets	[74]
kNN	entropy	a single feature from each cluster is chosen applying entropy	entropy, fuzzy feature evaluation index, classification accuracy	real life public domain	[73]

Figure 1

Three basic types of FS methods

'(A) Filter. (B) Wrapper. (C) Embedded.'

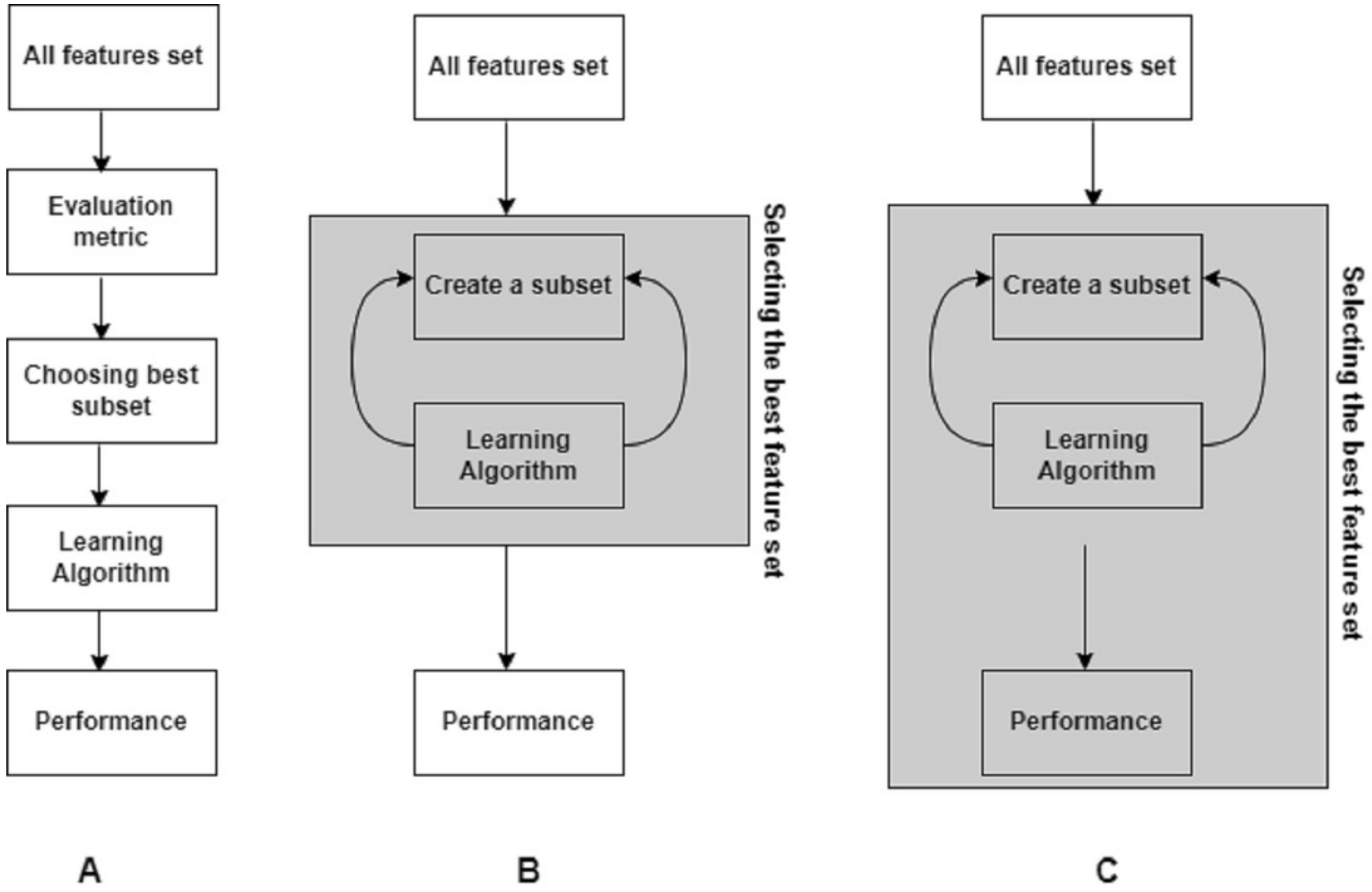


Figure 2

The representation of feature selection approaches based on grouping

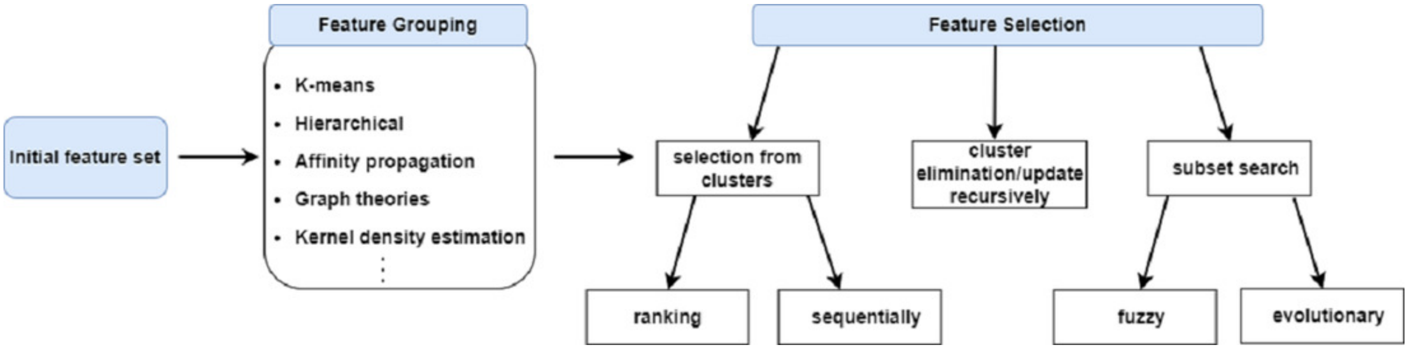
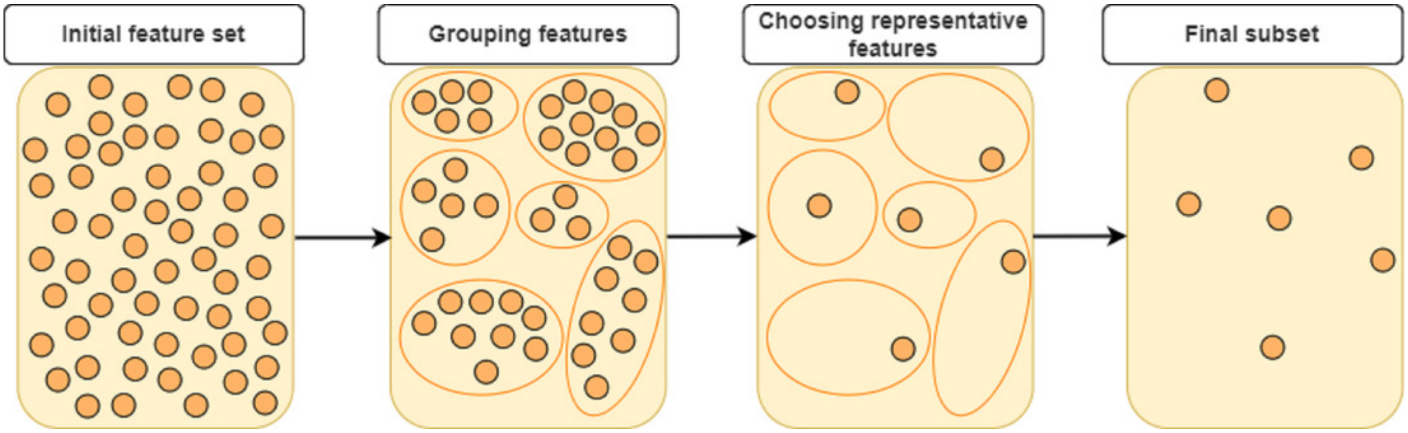


Figure 3

Typical approach for representative feature selection based on grouping


Initial feature set
Grouping features
Choosing representative features
Final subset