International Journal of Theoretical and Applied Computational Intelligence
https://ijtaci.com

Review Article

Feature Selection Methods in Big Medical Databases: A Comprehensive Survey

Mohsen Karimi[1*], Zahra Karimi[1], Mahsa Khosravi[2], Zeinab Delaram[3], Mostafa Habibi Dehsheikhim[3], Somayeh Arab Najafabadi[4], Mohammadreza Alizadeh Aliabadi[5] and Nakisa Tavakoli [6]

[1] Department of Bioelectric and Biomedical Engineering, School of Advanced Technologies in Medicine Isfahan University of Medical Sciences, Isfahan, Iran

[2] Department of Industrial and Manufacturing Systems Engineering at Iowa State University, Iowa, United States

[3] Computer Engineering, University of Texas at Dallas, Dallas, United States.

[4] Department of Computer Engineering, Najaf Abad Branch, Islamic Azad University, Isfahan, Iran.

[5] Department of Electrical Engineering, Ministry of Science, Research and Technology Non-governmental and Non-profit Higher Education, Institute Mehr Astan University, Guilan, Iran.

[6] Department of Computer Engineering, Faculty of Computer, Farsan Branch, Islamic Azad University, Farsan, Iran

[*]Corresponding Author: Mohsen Karimi. Email: mohsenkarimi@amt.mui.ac.ir

https://orcid.org/0000-0002-1645-5347

Received: 29/7/2025; Accepted: 30/8/2025; Published: 26/9/2025

https://doi.org/10.65278/IJTACI.2025.21

Abstract: Medical science is rapidly evolving alongside the continuous growth of medical data recording systems, resulting in massive and diverse datasets. While these data offer valuable opportunities, their high volume and complexity create significant challenges in processing, particularly in tasks such as classification and clustering. Feature selection has emerged as a critical strategy to overcome these issues by improving efficiency, accuracy, interpretability, and scalability. This article presents a comprehensive survey of feature selection methods with a dedicated focus on medical data, providing a structured categorization into filter, wrapper, and embedded approaches. It further reviews evaluation criteria, highlights strengths and limitations of each category, and discusses their relevance through examples from real-world medical applications. By combining theoretical perspectives with practical insights, this work contributes a clear roadmap for researchers in healthcare informatics, emphasizing that effective feature selection can substantially enhance medical data analysis and support future advancements in the field.

Keywords: Medical databases; Features selection; Machine learning; Evaluation criteria; Medical informatics

1. Introduction

Feature selection is a process that is commonly used in machine learning to solve the dimensionality reduction problem. This process selects a subset of the most important and efficient features. In this process, irrelevant features, extensions and noise are removed in order to show the data more simply and concisely. The selected features also play a decisive role in improving the performance of the bundles [1].

Although features selection is now successfully used in many expert systems such as text processing or text mining [2], image processing and machine learning [3], medical information [4, 5] industrial applications[6, 7], medical signal processing [8, 9], etc. Attention has been paid. In fields such as signal processing, image processing, data processing and data mining, new information and data are produced every day. Increasing computational costs, excessive training, and low efficiency are some of the disadvantages of processing large volumes of data. These voluminous data usually have duplicate information that even with the extraction of various features, duplicate features can still be seen in abundance [6]. The existence of redundant and Overfitting features makes the performance of the classifier suffer from over-regulation [10]. Feature selection in pattern recognition and machine learning presents a significant challenge due to the presence of a large number of features, often in the thousands [11-13]. By filtering out redundant and irrelevant features, it is possible to reduce overfitting, minimize computational costs, and ultimately improve the efficiency of the classification process. Consequently, efficient and effective feature selection algorithms that both reduce data dimensionality and enhance performance are essential. Feature selection is a combinatorial optimization problem that plays a crucial role in various applied fields and research areas [14].

Features represent the characteristics of objects, and the key to identifying and classifying these objects lies in the effective selection of a relevant subset of features. The performance of a classifier heavily depends on the quality of these features [15]. However, extracting appropriate features is challenging due to several factors, including noise, feature correlation, and irrelevant or redundant data. This leads to the fundamental question: which features are most relevant, and how can we identify the subset that best represents the data?[16]. In detection and classification systems, the goal of feature selection is to identify a subset of features that maximizes performance while minimizing computational complexity [17]. This process ensures that only the most informative features are used, improving both the accuracy of the model and its efficiency. By carefully selecting features, it is possible to not only enhance the classification outcomes but also reduce the computational burden, making the system more scalable and faster. Moreover, feature selection algorithms can be broadly categorized into three main types: filter methods, wrapper methods, and embedded methods. Filter methods assess the relevance of features based on their statistical properties before the learning algorithm is applied. Wrapper methods, on the other hand, evaluate subsets of features based on the performance of the model, iteratively selecting the best-performing subsets. Embedded methods perform feature selection during the training of the model, thus integrating the feature selection process directly with the learning algorithm itself. Each method has its advantages and challenges, depending on the specific context and type of data [18].

In real-world applications, selecting an appropriate feature selection strategy requires considering multiple aspects, including dataset dimensionality, heterogeneity of samples, computational constraints, and the intended clinical or analytical objectives. With the continuous expansion of medical datasets, feature selection is no longer just a step to improve efficiency, but rather a fundamental process for tailoring

models to complex healthcare environments. By identifying the most informative variables, these techniques help adapt machine learning systems to clinical needs such as early disease detection, prognosis monitoring, and personalized treatment planning.[19].Below are several key points that highlight the importance and challenges of feature selection in medical data:

- **Feature dependence and complexity**: Often, detection systems are presented with many features. However, these features may not be independent; they can be highly correlated or dependent on each other. The presence of irrelevant or redundant features can severely affect the system's performance. Additionally, using more features increases the system's complexity, but this does not necessarily result in better diagnostic accuracy. Therefore, selecting a subset of efficient, informative features is critical to optimizing performance and minimizing unnecessary complexity [20].
- **Learning algorithms and feature Selection**: During the training phase, a learning algorithm selects relevant features to build a model that represents the training data [21]. If too many features are selected, the model becomes more complex, which may lead to overfitting. According to the Minimum Description Length Principle (MDLP), simpler models are preferable as they are less likely to overfit and more likely to generalize well. The Minimum Description Length principle is an information-theoretic approach that selects models by minimizing the total description length of the data and the model itself [22]. A complex model may also be sensitive to noise in the training data, leading to poor performance when tested on new, unseen data [23].
- **Computational efficiency and real-time applications**: Reducing the number of features can significantly decrease the computational burden, which is especially important for real-time applications. In medical data processing, where the speed of diagnosis can be critical, reducing computational costs without sacrificing accuracy is essential. In some cases, reducing the feature set may actually improve classification accuracy by focusing on the most relevant and powerful features, thereby reducing the risk of overfitting and improving the model's generalization capability [16]. The goal is to select features that provide the best separation between classes and enhance predictive performance [24].

- ***Related research in medical data feature selection***

  Several studies have focused on feature selection in medical data, exploring various methods and strategies for improving classification performance [25]. Despite this, there remains a lack of comprehensive research that thoroughly covers feature selection in the context of medical data across a range of applications. This article seeks to fill this gap by providing an overview of feature selection methods tailored specifically for medical data. The innovations of this paper are as follows:

  - **Definition and categorization of feature selection methods**: A comprehensive collection of definitions and the different types of feature selection methods commonly used in medical data.
  - **Comparison of feature selection methods**: A comparative analysis of various feature selection methods, highlighting their advantages and limitations in the context of medical data.
  - **Review of existing research**: An overview of the existing body of research on feature selection in medical data, identifying trends, challenges, and research gaps.

This article is structured as follows: Section 2: Provides a detailed explanation of different definitions and concepts related to feature selection. Section 3: Categorizing of feature selection methods. Section 4: discussions on paper content and finally. Section 5: Concludes the article with key takeaways and recommendations for future research.

2. Definition of feature selection

Various definitions have been proposed for feature selection [26]. In one definition, selecting a subset of features with the best classification performance result is called feature selection [27]. In this definition, Redundancy features or unrelated features, which often have noise in the data, which cause the classifier to make mistakes and degrade the classification performance, are removed. In addition, reducing the number of features helps the human expert to focus on a subset of relevant features, which gives a better view of the process described by the data [28].

Feature selection is a critical process in machine learning that addresses the problem of high-dimensional data by selecting a subset of the most relevant features and eliminating unnecessary or noisy ones [29]. This process helps simplify and clarify the data, improving model efficiency and performance. The benefits of feature selection can be categorized into several key areas:

- **Advantages of feature selection:**
  1. **Reduced execution time**: Feature selection helps decrease the time needed for learning processes by removing irrelevant, redundant, or excessive features. This makes the training process faster and more efficient.
  2. **Improved model accuracy**: By removing noise, redundant features, and irrelevant data, the learning algorithm can focus on the key aspects of the data. This enables the creation of simpler and more accurate data models, which ultimately improves the classification performance.
  3. **Generalization and simplicity**: Feature selection can also lead to a more general and simpler model. By focusing on the most important features, the resulting model is less likely to overfit the training data and is more likely to generalize well to new, unseen data [30].

2.1. Feature selection vs. feature transformation:

Feature selection differs from feature transformation, where new features are created by combining the original features. While feature transformation methods (such as[31](PCA), Linear Discriminant Analysis (LDA) [32], and Locally Linear Embedding (LLE)) [33] generate new, transformed features, feature selection methods preserve the original features. This preservation is often desired in many domains because it allows the model to maintain interpretability and relevance to the problem at hand [34].

- **Feature selection as a dimensionality reduction method:**
  Feature selection is considered one of the dimensionality reduction techniques. It is closely related to other methods aimed at reducing the number of input variables or features, such as PCA (Principal Component Analysis) or compression-based methods [35]. However, feature selection is distinct from these methods in that it does not modify the original features but instead selects a subset of the most relevant features. This is particularly important in fields where understanding and preserving the original features is critical for interpreting the model.
- **Position of feature selection in dimensionality reduction:**
  Figure 1 illustrates how feature selection fits within the broader family of dimensionality reduction techniques. Unlike transformation-based methods that generate new feature spaces, feature selection retains the original variables, which makes it especially valuable in fields that demand clear interpretability, including medical data analysis, financial forecasting, and scientific discovery. Rather

than merely reducing dimensionality, it strategically identifies the most relevant variables to balance computational efficiency with domain-specific insight[34].

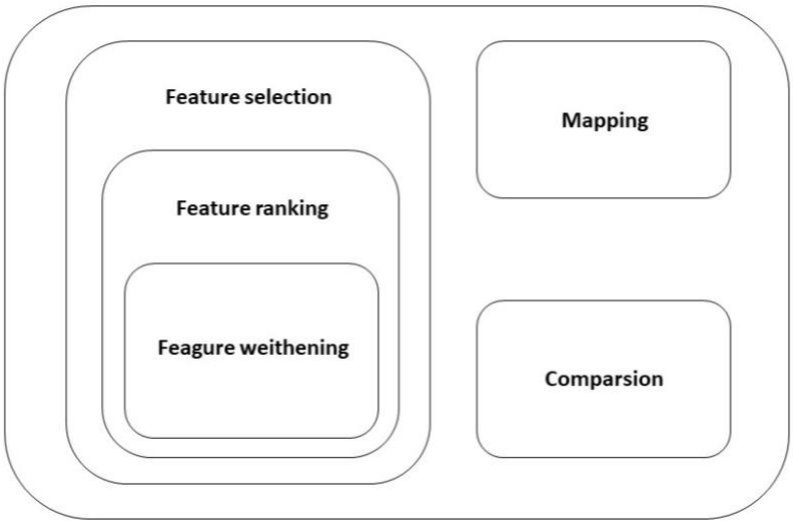

Figure 1: Relationship between feature selection and other dimensionality reduction techniques

2.2. Expression of feature selection problem

The problem of selecting a subset of features in order to improve the desired classification performance is explained as follows: V shows the main set of features with dimensions m. d shows the desired number of features in the selected subset of X. Conventionally, the selection problem is to find a subset of $X \subseteq V$ such that $|X| = d$ is The symbol $J(X)$ is used to display the feature selection criterion function [36]. A value above J refers to the optimal subset of the feature. The feature selection problem is expressed as relation (1) [28]:

$$J(X) = max_{Z \subseteq V, |Z|=d} J\,(Z) \quad (1)$$

- **Purpose of feature selection**

  Different feature selection algorithms may pursue varying objectives depending on the context and the specific application. Below are some of the common goals that researchers typically aim for when applying feature selection:

  1. **Minimizing the feature subset while retaining sufficient information**: One key goal is to identify the smallest subset of features that is both necessary and sufficient to accurately represent the target concept or task. In other words, the objective is to reduce the dimensionality of the data by eliminating irrelevant or redundant features while ensuring that the remaining features are capable of capturing all the important information required for classification or prediction. This is critical in scenarios where model efficiency is important, as fewer features mean faster processing and less computational cost [37].

  2. **Selecting an optimal subset of features for optimization**: Another common goal is to choose a subset of NNN features from a larger set of MMM features, where N<MN < MN<M, in such a way that the selected subset optimizes a certain criterion function. This means that the algorithm aims to find the best possible combination of NNN features that provides the highest performance based on a specific evaluation metric (such as classification accuracy, information gain, etc.). This

approach often involves evaluating all possible feature combinations to ensure the best-performing subset is selected. However, this can be computationally expensive, especially when dealing with large datasets and a large number of features [37].

3. **Improving prediction accuracy or reducing model complexity**: Feature selection can also aim to improve the prediction accuracy or reduce the complexity of the model without significantly compromising performance. The goal here is to select a subset of features that either enhances the predictive power of the model (i.e., improves accuracy) or reduces the size and complexity of the model (e.g., smaller models may be easier to interpret and faster to run), while ensuring that there is no substantial drop in accuracy when compared to the full feature set. This approach strikes a balance between optimizing performance and minimizing model complexity, which is especially valuable in real-time or resource-constrained applications [38].

- **Explanation:**
  1. **smallest subset for target concept**: This goal focuses on the necessity and sufficiency of the selected features. It's about finding the minimum set of features that can still capture all the important patterns in the data. In practice, this reduces the model's complexity and makes it easier to understand and interpret, while avoiding overfitting [39].

  2. **Optimizing feature subsets**: Here, researchers aim to find the best possible feature subset based on a specific criterion function. By exploring all subsets of NNN features from a larger set of MMM features, the goal is to identify the combination that optimizes performance. This process often involves algorithms like exhaustive search, genetic algorithms, or sequential feature selection, but it can become computationally expensive for high-dimensional data.

  3. **Balancing accuracy and model complexity**: This goal involves a trade-off between performance and efficiency. Selecting fewer features may make a model more computationally efficient and easier to interpret, but the challenge is ensuring that this reduction doesn't come at the cost of significant accuracy loss. For instance, selecting only the most relevant features may help speed up predictions without drastically affecting the model's ability to classify or predict outcomes[30].

By understanding these goals, researchers and practitioners can choose feature selection algorithms that align with their specific objectives, whether it's reducing computational time, improving model accuracy, or creating more interpretable models. Table 1 highlights three different goals in feature selection algorithms. Each goal addresses a specific challenge, such as reducing dimensionality, optimizing model performance, or balancing accuracy and complexity. The selected goal will depend on the specific requirements of the task at hand, such as computational resources, model interpretability, and accuracy.

Table 1: Highlights three different goals in feature selection algorithms

Goal	Description	Key objective	Challenges	Common use cases
Minimizing the feature subset while retaining	Identifying the smallest subset of features that is both necessary and sufficient to	Reduce the dimensionality of the data by eliminating irrelevant features,	Ensuring that the reduced set of features still captures all important	Real-time applications, high-dimensional datasets,

Sufficient information	represent the target concept.	keeping only those that contain the most critical information.	information; risk of oversimplification.	interpretability-focused models.
selecting an optimal subset of features for optimization	Choosing NNN features from a larger set of MMM features (N<MN < MN<M) that optimize a specific evaluation criterion.	Optimize performance by selecting the best possible combination of NNN features that improves model accuracy or any other performance metric (e.g., information gain).	Computational complexity of evaluating all possible feature subsets; may be time-consuming for large datasets.	Classification tasks, feature subset exploration, performance optimization in model-building.
Improving prediction accuracy or reducing model complexity	Selecting a subset of features that improves accuracy or reduces the size/complexity of the model without losing significant accuracy.	Achieve a balance between accuracy and model simplicity by selecting only the most relevant features, resulting in a more efficient and faster model.	Maintaining accuracy while reducing the number of features; avoiding performance drops after simplification.	Real-time systems, embedded devices, medical data, financial forecasting, and interpretability-focused models.

- **The basic steps of feature selection**

  There are four basic steps in different feature selection methods:

  1. A Generation procedure to produce the candidate subset.
  2. An evaluation function to evaluate the subset under test.
  3. A stop criterion to determine when to stop.
  4. A Validation procedure to check the correctness of the selected subset [30].

  The following steps are reviewed:

- **Production process:**

  The production process is a search process aimed at generating feature subsets for evaluation. It can proceed in different ways, depending on how the initial set of features is chosen:

  1. **Case 1: Starting without any features:** In this case, no features are initially selected. The process progressively adds features in a stepwise manner based on their relevance.
  2. **Case 2: Starting with all features** Here, all features are initially considered, and features are then iteratively removed to find the most relevant subset.
  3. **Case 3: Starting with a random subset** A random subset of features is selected initially, and the algorithm either deletes features in a stepwise fashion or generates new subsets

randomly after each step [31].

The goal of the production process is to generate feature subsets and evaluate their effectiveness in the context of the task at hand.

- **Evaluation function:**
  The evaluation function plays a key role in assessing the quality of the feature subsets produced by the search process. It compares the current subset with the previous best subset according to a predefined criterion. If the new subset performs better than the previous one, it replaces the previous subset as the best solution. The evaluation function could measure aspects such as accuracy, information gain, or any other relevant metric for assessing the subset's performance [40].
- **Stop criterion:**
  Without a clear stopping criterion, the feature selection process might end up searching the entire space of possible feature subsets, which can be computationally expensive and time-consuming. The stopping criterion ensures that the process halts when certain conditions are met. These conditions can be based on:
- **Search-based stopping criteria:**
  1. The number of selected features reaches a predefined threshold.
  2. A predefined number of iterations or steps is reached.
- **Evaluation-based stopping criteria:**
  1. Further adding or removing features does not improve the quality of the current subset.
  2. An optimal subset is identified according to the evaluation function [22].
  3. These stopping conditions help balance computational efficiency with the quality of the selected feature set.

2.3. Validation

Once the feature selection process is completed and a subset of features is selected, it is validated through a separate process. The validation step is crucial in ensuring that the chosen feature subset is truly optimal and performs well in real-world conditions. This process involves testing the selected subset through various validation techniques to assess its reliability and accuracy. The evaluation criteria used during the feature selection process play a decisive role in this validation phase. Proper selection of evaluation criteria is essential because if the wrong criterion assigns inappropriate values to the feature subsets, the correct subset will never be identified as optimal. In essence, the choice of the evaluation function directly influences the feature selection outcome. For classification problems, the Bayesian error rate E(S) is often considered an optimal criterion. The Bayesian error rate is the lowest possible error rate that any classifier can achieve, serving as a theoretical benchmark in classification tasks [41]. The goal is to minimize this error rate, which is computed using Eq. (1). In discrete spaces, Eq. (2) is typically used to calculate the error rate. In the concept of classification and its related problems, an optimal criterion must have an optimal Bayesian error rate E(S). E(S) is calculated from Eq. (1) [42]. Eq. (2) is also used in discrete space.

$$E(S) = \int_S p(S)(1 - \max_i (p(c_i|S)))dS \tag{1}$$

$$or \sum_S p(S)(1 - max(p(c_i|S))) \tag{2}$$

As it can be seen from the relations (1) and (2), E(S) is desired as sum or integral and also $p(S)(1-max(\,p(c|s)))$ is non-linear and non-negative [35]. In Eq. (3), the upper limit of E(S) is calculated, $H(C|S)$ conditional entropy C is given in each S.

$$E(S) \leq \frac{H(C|S)}{2} \tag{3}$$

Calculating E(S) directly is very difficult because S is a combination of features. As a result, most researchers prefer to use criteria based on correlation and distance. Eq. (4) shows the evaluation criterion of the correlation coefficient. The covariance of the variables a and b is the variance.

$$r(a,b) = \frac{cov(\,a,b)}{\sqrt{var(\,a)}\sqrt{var(\,b)}} \tag{4}$$

Pearson correlation coefficient is calculated in relation (6).

$$r(a,b) = \frac{N\sum a_i b_i - \sum a_i \sum b_i}{\sqrt{N\sum a_i^2 - (\sum a_i)^2}\sqrt{N\sum b_i^2 - (\sum b_i)^2}} \tag{5}$$

The Mutual Information (MI) is obtained from Eq. (6). $p(0)$ is the Probability Density Function (PDF).

$$I(a,b) = \sum_a \sum_b p(ab)\, log\, \frac{p(ab)}{p(a)p(b)} \tag{6}$$

The relation (8) shows the Symmetric Uncertainty (SU). Entropy is any property [43].

$$SU(a;b) = \frac{2I(a;b)}{H(a) + H(b)} \tag{7}$$

Distance information is calculated in relation (8). $H(a|b)$is the conditional entropy of a on condition b.

$$d(a,b) = \frac{H(a|b) + H(b|a)}{2} \tag{8}$$

Finally, the last criterion which is very common for evaluation is calculated from Eq. (9) and it is the Euclidean distance criterion.

$$d(a,b) = \sqrt{\sum (a_i - b_i)^2} \tag{9}$$

These common criteria are used to evaluate feature selection methods. Although there are other criteria such as Laplacian score1, Fisher score and other criteria, information criteria require features in discrete state and discretization is required if they are used [44]. Having reviewed the definitions, objectives, and fundamental steps of feature selection, the next logical step is to examine the various approaches used to implement this process. Since each approach comes with its own advantages and challenges, Section 3 provides a comprehensive categorization of feature selection methods, enabling clearer distinctions and facilitating the comparison and selection of appropriate techniques for medical applications.

3. Categorizing of feature selection methods

The search strategies used in the first step of feature selection is presented in following. If the main feature set has N number of features, each feature selection method searches among 2^N candidate feature subsets. These methods try to select the best subset according to the evaluation criteria. Although this complete process tries to find only one better subset, it may be computationally very complicated and heavy even with a moderate size N feature set. Other strategies are based on exploratory methods or random search,

which try to have the best performance by reducing the computational cost. These strategies need a stopping criterion to prevent the complete search of candidate subsets [45].

In the presented methods for feature selection, a classification can be provided. In this article, this classification is done as follows. Methods based on tags of data in the database, methods based on feature space search, and finally methods based on learning. Data tag-based methods are classified into three categories: supervised, semi-supervised and unsupervised. Learning-based methods are divided into three categories: filter, wrapper, permutation and combination methods. Search-based methods also consist of complete stochastic and exploratory search strategies. Figure 2 shows this classification. In the following, this category will be examined.

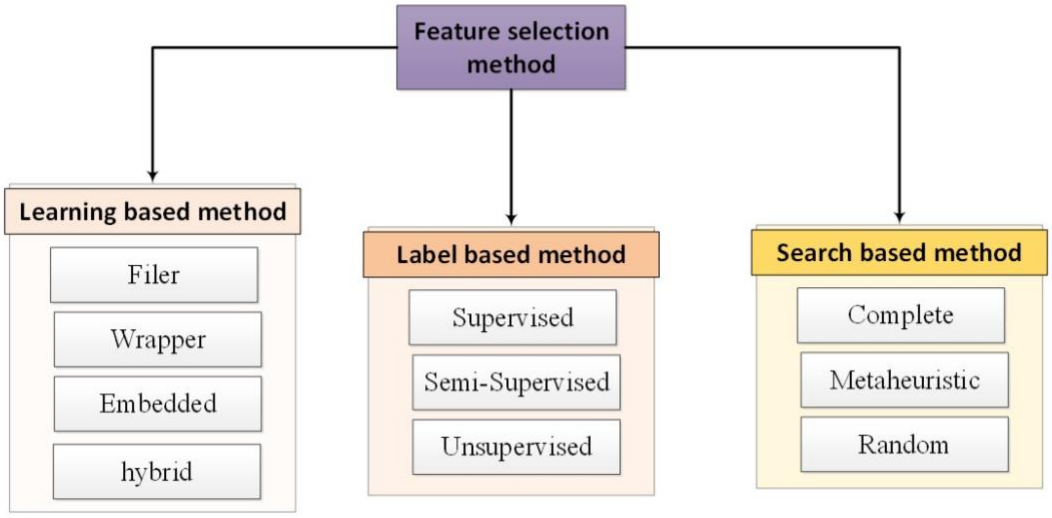

Figure 2: Classification of feature selection methods.

3.1. Search based algorithms

3.1.1. Heuristic search

Heuristic search is an efficient strategy used in feature selection to explore the space of possible feature subsets. In this approach, at each iteration of the search process, all remaining features are evaluated to decide whether they should be included or excluded from the current feature subset. This method aims to find good (though not necessarily optimal) subsets of features that improve model performance, especially when the search space is large and exhaustive search is computationally impractical. Heuristic search methods typically use a greedy approach, where decisions are made based on a local search criterion, which may lead to good solutions without evaluating every possible feature combination[46]. For instance, in [47] proposed two heuristic filter-based methods for gene selection in medical datasets, demonstrating that variance performs well on standard datasets, while Mutual Congestion considerably improves accuracy in high-dimensional data.

3.1.2. Complete search

This production process performs a complete search to find the optimal subset according to the evaluation function. A comprehensive search is a complete search, but if the search is complete, it does not mean that the search is comprehensive. There are various functions to reduce the search space without reducing the probability of finding the optimal subset. Here, although it $O(2^N)$ is the order of the search

space, but a smaller subset is evaluated. The optimality of the subset is guaranteed according to the evaluation function, because the production process can be reversed. Going back can be done using different methods. An example of this type of search is the Beam Search (BS) search algorithm. This algorithm is one of the best and first search engines that uses a limited queue to limit the search space. The queue is sorted from the best to the worst feature subset, where the best subset is at the top of the queue. This production process proceeds by removing the subset in the first queue and generating all possible subsets by adding features to it. Each subset is placed in a suitable place of the queue. If there is no limit on the queue size, the BS is an exhaustive searcher [39].

3.1.3. Random search

In this type of strategies, although the search space is of order, usually the number of subsets is less than by adjusting the possible number of repetitions. The optimality of the selected subset depends on the available resources [31]. An example of this type of search process is evolutionary algorithms. Evolutionary algorithms for the application of feature selection are approaches that have the task of generating subsets of features and optimal search to find the desired set of features. Evolutionary algorithms such as Genetic Algorithm (GA) [48], Particle Swarm Optimization (PSO) [49], Ant Colony Optimization (ACO) [50], Charged System Search (CSS) have been developed to solve large-scale combinatorial problems and have been widely used in the feature selection problem [20].

Table 2 is the comparison for the different search strategies used in feature selection: heuristic search, complete search, and random search. These strategies vary in terms of computational complexity, optimization guarantees, and search efficiency.

Table 2: Comparison for the different search strategies

Search Strategy	Search Process	Computational Complexity	Advantages	Disadvantages	Examples
Heuristic search	Iteratively selects or excludes features by evaluating their relevance. Often uses greedy or local search methods.	**$O(N^2)$**	- Simple and fast to implement. - Efficient for small to medium-sized datasets. - Can handle noisy features.	- Can get stuck in local optima. - May not find the optimal subset due to greedy nature. - Limited flexibility.	-Relief -Sequential Forward Selection (SFS) -Sequential Backward Selection (SBS)
Complete search	Explores the entire search space for the optimal subset using evaluation functions. Can be exhaustive or constrained.	**$O(N^2)$**	- Guarantees finding the optimal subset based on the evaluation function. - Can explore all possible subsets.	- Computationally expensive for large datasets. - Search space grows exponentially, making it impractical for	- Beam Search (BS) - Exhaustive Search

				high-dimensional data.	
Random search	Selects random feature subsets and evaluates them. The search space is reduced by limiting iterations or repetitions.	Depends on number of iterations. Typically O(N) or lower for specific algorithms.	- Can explore a wide range of feature subsets. - Requires fewer resources for computation compared to exhaustive methods. - Useful for large-scale problems.	- Optimality is not guaranteed. - The solution depends on the number of iterations and available computational resources. - May miss the optimal subset.	- Genetic Algorithm (GA) - Particle Swarm Optimization (PSO) - Ant Colony Optimization (ACO) - Charged System Search (CSS)

3.2. Learning-based methods

The feature selection methods are divided based on the interaction with the learning algorithm and label information in the data [51]. Feature selection methods are divided into three categories based on their interaction with learning algorithms: wrapper, filter and Embedded. If the feature selection algorithm uses learning algorithms for feature selection, this method will be a wrapper. In rapper-based methods, according to the predicted accuracy of the classifier (classifier) as a black box, it selects the feature and ranks the subsets of the feature according to their predictive power It binds. Filter-based methods select features using a pre-processing step. Learning algorithms are not included in this category of methods. The main drawback of this method is that it does not consider the effect of the selected feature subset in the learning algorithm. Under these two methods, comprehensive search can be done if the number of variables is not too large. As the number of features increases, the computational cost increases, in this case the search becomes impossible [52]. Combined or integrated methods benefit from the advantages of both previous methods by using different evaluation criteria in different stages of the search. In contrast to Rapper's methods, they select features while also considering the design of the classifier. In this control method, it is difficult to select the appropriate number of features and usually there is an excess of features. In the following, these methods are described in feature selection [53].

3.2.1. Wrapper method

In wrapper-based methods, the performance (for example, predicted classification accuracy) of a classification algorithm is used to evaluate a subset of features. Figure 3 shows the general idea of rapper methods. For each generated feature subset X, the wrapper evaluates the goodness by applying the classification algorithm on the data set using the features in the subset X. Rapper can find feature subsets with high accuracy, and the reason is that the features match well with the learning algorithm [54].

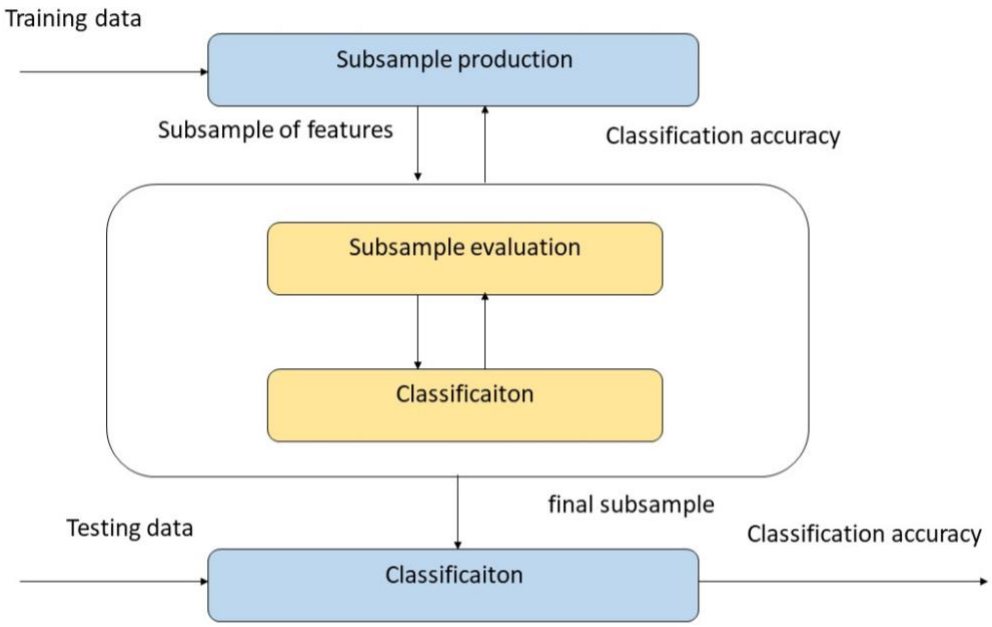

Figure 3: The general idea of wrapper-based methods.

A wrapper-based feature selection model consists of two parts:

- **The first part**: feature subset selection, which selects the best subset using the predicted accuracy of the classifier (training data) as a criterion.
- **The second part**: learning and testing, when a classifier was taught from the training data with the best feature subset, it is tested on the test data [55].

Rapper methods are often criticized for their computational time. Due to the fact that a comprehensive search requires different evaluation, forward selection or backward elimination methods have been proposed and used. These methods are referred to sequential methods. The sequential forward selection method of SFS starts with a zero set and the variables increase in a developmental way and make larger subsets, while in SBS with a set composed of all It starts variables and deletes them in a developmental way. Both of these methods suffer from a problem called nesting effect. This means that the backward deletion method cannot select the deleted features again, and in the forward selection method, once the features are selected, the feature cannot be deleted. The result is that the mentioned methods are not optimal. In addition, the forward selection and backward elimination methods require high computing time when the number of variables is large [14].

3.2.2. Filter-based methods

Filter-based feature selection applies a selected metric to find irrelevant features and filters out redundant data. The selection process is independent of the educational process. Filter-based methods rank the features as a pre-processing step before the learning algorithm and select those features with high-ranking scores. The score is calculated by measuring the variance between the expected value of information and the observed value. The filter evaluation criterion is usually used to analyze the internal features of the feature subset, including correlation, distance, information acquisition, etc. In practice, filter-based feature selection can be first checked by expert knowledge, then filtered by filter-based methods. Traits that have been shown to be associated with a specific disease or physiological response are often selected directly in trait selection [56]. Columns with poor feature selection scores are ignored. Filter-based feature selection

provides various performance measures to evaluate information value, such as correlation coefficient, mutual information, Kendall correlation, Spearman correlation, Chi Squared, Laplacian score, Trace Ratio criterion and Fisher score. Commonly used criteria for selecting supervised features. Fahim Sufi and colleagues. ranked feature subsets based on a correlation-based heuristic evaluation function. The algorithm selected the features by calculating the average feature class correlation and the average feature-feature correlation. These two criteria ensure that irrelevant features and redundant features are removed from the feature set because they are not correlated with the class or other features. Some researchers evaluated each feature from the set of primary features with Fisher's score. Fisher's score selected each database feature independently according to their scores under Fisher's criterion, resulting in a subset of features not being optimal. It is often used to select feature sets with lower dimensions. The filter method uses statistical indicators to score and filter each feature, focusing on the features of the data itself [47].

The advantage of the filter method is that the calculation is fast and does not depend on a specific model. However, the final classification accuracy may not be high because the selected features are not customized for the specific model. As shown in Figure 4, the filter-based method, unlike the wrapper-based method, instead of using a specific classifier for learning, will choose based on the inherent characteristics of the data. Then it uses them to evaluate subsets of features.

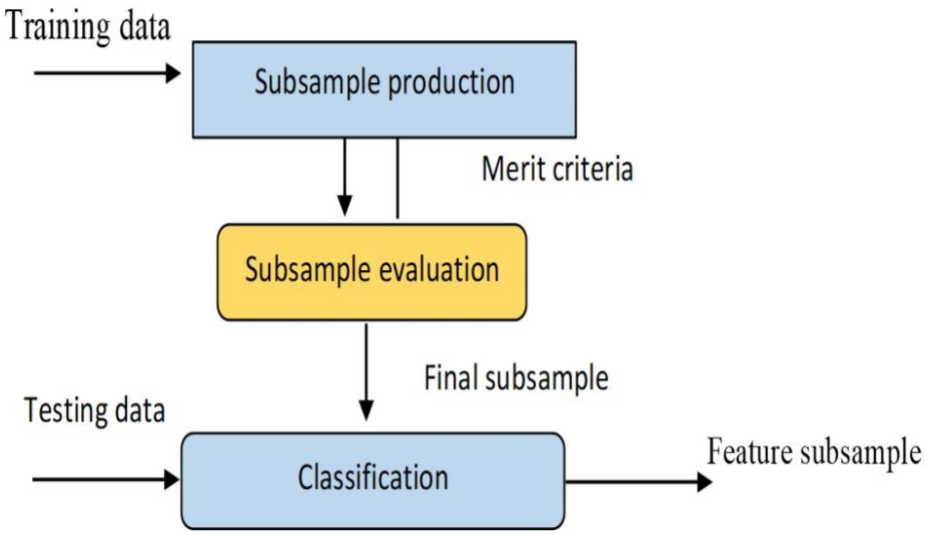

Figure 4: Feature selection approach using filter-based method [45].

Filter-based methods are evaluated based on four evaluation criteria: distance, information, dependence, and consistency. Independent of any classification algorithm, filter-based methods remove irrelevant features, extraneous and noise in pre-processing steps, before the classification takes place. The basis of the filter-based method is to search for related features and remove unrelated features [57].

A filter feature selection model consists of two parts:

- **The first part**: feature selection using criteria of distance, information, dependence and consistency. The classifier is not used in this section.
- **The second part**: This part is similar to the wrapper model, where a classifier is trained on the training data with selected features and tested on the test data.

In addition to the fact that the filter method is based on the inherent characteristics of the data, it also has the following characteristics:

- Normally, the measurement of distance, information, dependence and compatibility criteria has a lower computational cost compared to the measurement of the predicted accuracy of the classifier, so a filter method can be a subset if the rest of the issues are equal. produce more quickly.
- Due to the simplicity of criteria and low time complexity, a filter method can use a larger amount of data than a classifier, so when a classifier cannot be trained directly from a large amount of data. See, this method can be used to reduce data dimensions.
- The problem of the filter approach is that the internal relationship between the features is not considered, and the selected set may have a large number of interrelated features, as a result, it does not achieve our goal, which is to reduce features as much as possible and eliminate unnecessary features [58].

In this category, feature ranking methods and feature weighting methods are located. In feature ranking methods, instead of returning a set of the best features by the feature selection approach, the features They sort from the most relationship to the least relationship. Then this ranking can be used to discard some features. Weighting algorithms belong to the category of filter-based methods, they assign a communication weight to each specific feature. These weights can then be stored for ranking and then used for feature removal, for example by choosing a threshold weight and removing all features with a weight below the threshold. An important question about the definition of "subset of related features" is the definition of "related". Two degrees can be defined for the definition of "related": weak and strong. Feature X is strongly correlated if the removal of this feature causes the performance of an ideal classifier to degrade greatly. A feature X is weakly related if this feature is not strongly related and there is a subset of features, S, that the classifier's performance on this subset of features is worse than on set $S \cup \{X\}$. A feature is unrelated if it is not weakly or strongly related [59].

Related features do not necessarily mean that they should be in the subset of optimal features. Similarly, unrelated features do not always imply that these types of features should not be optimal in the subset. This case can happen when the classifier is limited in the space of assumptions and this case makes it unable to use all the features. As an integrated example, in [60] a novel framework that combines filter, wrapper, and embedded methods for medical data, achieving 94.45% accuracy on a microarray dataset and 91% accuracy on the Cleveland Heart Disease dataset using an SVM classifier.

3.2.3. Embedded methods

Alternative methods use learning-based methods for feature selection. Alternative methods work based on combined and group learning. The use of group learning methods reduces the computational cost. It also improves the classification accuracy. It should be noted that even by adding a feature to the desired feature set, the computational cost does not increase in this method. In replacement methods, feature selection and learning evaluation are done simultaneously, while in rapper methods, the feature selection process is done after learning. The simultaneous performance of learning and feature selection has reduced the computational complexity [61].

The selection of embedded features is embedded in the construction of the machine learning algorithm. It provides a trade-off solution between the filter method and the wrapper method, which can solve the high redundancy of the filter algorithm and the computational complexity of the wrapper algorithm. Permutation-based feature selection is performed automatically during the learner training process. Compared to the other two methods, the process of searching and selecting the subset of features is included in the construction of the classifier. Regularization and tree methods are widely used in embedded methods.

Regularization models in the form of normative regularized regression models, such as lasso, sparse linear discriminant analysis, and regularized support vector machine, are widely used in embedded methods. Regularization is the application of additional constraints or penalties on the loss function when training a neural network, which can reduce the complexity and instability of the model in the learning process, thus preventing overfitting and improving the generalization ability. The decision tree is a classic embedded feature selection method [62].

Features with good classification ability are selected at the nodes of the tree, and then the selected feature subsets are used to perform learning tasks. Feature subsets are selected during the decision tree generation process. Random forest has the advantages of high accuracy, good robustness and ease of use, which makes it one of the most popular machine learning algorithms. Random forest provides two methods for feature selection, including average impurity reduction and mean decrease impurity and mean decrease accuracy. Tree-based predictive models can be used to calculate the importance of features and thus remove irrelevant features. Embedded feature selection can be applied to high-dimensional datasets, but the design of the embedded method is strongly associated with a particular learning algorithm, which in turn limits its applicability to other learning algorithms [63]. As an example of embedded approaches, in study [64] introduced a Time Selection Layer in deep learning models for ozone pollution prediction, which reduced complexity, enhanced interpretability, and improved model performance by 9% on average across multiple monitoring sites.

3.2.4. Hybrid methods

In feature selection, combined or hybrid methods integrate the strengths of multiple approaches to improve the performance and efficiency of the selection process. Typically, these methods combine the filter and wrapper techniques, each of which has its own advantages and limitations. The main goal of hybrid methods is to leverage the benefits of both approaches while mitigating their individual weaknesses. The filter method is generally fast and computationally efficient, as it evaluates features independently of any machine learning algorithm, using statistical measures like correlation or mutual information [65]. On the other hand, wrapper methods are more computationally expensive but tend to provide more accurate feature subsets, as they evaluate subsets based on their impact on a specific model's performance. By combining these two methods, hybrid approaches aim to balance computational efficiency with classification accuracy, addressing the shortcomings of each method when used alone [66].

- **Combining filter and wrapper methods**: The integration of filter and wrapper methods in hybrid feature selection provides a powerful strategy for enhancing feature relevance and reducing redundancy. The filter component can quickly eliminate irrelevant or redundant features, reducing the search space and making the wrapper phase more focused and efficient. By filtering out features that are highly correlated or statistically insignificant, the wrapper method can then focus on evaluating more promising subsets that are likely to improve model performance. Conversely, the wrapper component enhances the filter approach by using a predictive model to assess the usefulness of feature subsets, thereby ensuring that only the most relevant features are selected. In this way, the hybrid method combines the computational efficiency of the filter approach with the high accuracy of the wrapper approach, resulting in a feature subset that not only reduces dimensionality but also improves model performance in real-world applications [67]. In [68] proposed a hybrid serial filter–wrapper framework with an elite-guided mutation strategy (SFEMEO) for cancer gene expression data, which outperformed nine benchmark algorithms across ten medical datasets, achieving accuracy improvements ranging from 3.73% to 18.13% and significantly better optimal fitness.

- **Advantages and challenges of hybrid methods**: The main advantage of hybrid feature selection methods lies in their ability to overcome the individual weaknesses of filter and wrapper methods. Filter methods can quickly remove irrelevant features, but they often ignore feature interactions that could improve classification performance. Wrapper methods, while more accurate, tend to be computationally expensive and may suffer from overfitting if the feature subset is too large. By combining both approaches, hybrid methods can exploit the strengths of each, reducing computational costs while maintaining or even improving predictive accuracy. However, the integration of these methods comes with its own challenges. Hybrid approaches often involve more complex search strategies, which can increase the computational burden compared to using filter or wrapper methods alone. Additionally, careful tuning of the filter and wrapper components is essential to ensure that the combination does not result in suboptimal feature subsets. Despite these challenges, the use of hybrid methods in feature selection has gained popularity due to their ability to provide high-quality and efficient solutions, especially in large-scale machine learning tasks and real-time applications[69].

3.2.5. Comparison of learning-based methods

Feature selection methods Filter-based feature selection, to select a feature or to select a subset of features, starts with all the features in the database. Then, to select each feature, statistical criteria are used to evaluate the desired features for the purpose of selection or purpose. Among the statistical criteria is the Pearson coefficient [33]. LAD linear discriminant analysis, AWOVA, Chi square [34], MI mutual information or other methods are used. All these statistical methods depend on the change of feature in the database. The methods of Pearson's coefficient and mutual information are among the most common statistical methods in the field of filters. Filter methods use characteristics such as Information gain, stability, correlation dependence and distance criterion to select a feature.

Wrapper-based methods are known as black box methods. These methods do not have information about the features and the type of selected features are selected based on learning methods. In other words, the features are selected based on the estimation of accuracy in learning methods. Based on the accuracy obtained in the learning method used in selected, removed or selected, the feature is determined. Rapper-based methods have higher complexity, but have much higher accuracy than filter-based methods. It is possible to refer to the methods of forward hierarchical feature selection, backward hierarchical feature selection, methods based on evolutionary algorithms [70].

Wrapper-based feature selection uses a predefined classifier to evaluate a set of features. This method scores the features using the learning algorithm, these features are finally used in the classification. The feature selection process is integrated with the training process, and the predictive ability of the model is used as a selection criterion to evaluate the feature subset, such as classification, accuracy, complexity penalty factor. The forward and backward selection algorithm in multiple linear regression is a simple implementation of the wrapper. Sequential floating forward search (SFFS) algorithm uses sequential forward selection (SFS) and sequential backward selection (SBS) respectively to obtain the best set of database features [71].

Compared with the filter-based method, the wrapper-based method performs better in generating high-quality subsets, but the data processing is computationally complex because the learner needs to be trained several times during the feature selection process. Unlike filter selection, which does not consider subsequent classification algorithms, complex selection directly takes the performance of the final classification algorithms as the evaluation standard for the subset of features. In other words, complex

feature selection is to select the most optimal feature subset for a given learning algorithm. However, the performance of the subset of features is influenced by the specific learning algorithm. Feature subset stability and consistency is poor because each additional feature must be built into the feature subset for evaluation. Rapper-based feature selection has high time complexity and is not suitable for high-dimensional datasets [4]. Table 3 shows the comparison of these methods [72].

Table 3: Comparison of feature selection methods based on the type of learning

Feature selection method	Interaction with learning algorithm	Key characteristics	Advantages	Disadvantages	Common usecases
Wrapper-based methods	Uses learning algorithm for feature selection	- Evaluates feature subsets by training a model using the selected features. - High accuracy based on model performance.	- High accuracy due to model-specific evaluation. - Can find highly relevant subsets for specific tasks.	- Computationally expensive (requires multiple model evaluations). - High time complexity.	- Tasks requiring high accuracy, such as classification in complex datasets. - Real-time predictions.
Filter-based methods	Independent of learning algorithm	- Features are ranked or filtered based on statistical measures (e.g., correlation, mutual information).	- Computationally efficient. - Does not depend on any classifier. - Can handle large datasets.	- May ignore feature interactions that could improve model performance. - May not result in optimal subsets for classification.	- Preprocessing for large datasets. - Applications where computational efficiency is critical.
Embedded methods	Integrated with learning algorithm	-Feature selection occurs during the model training process. - Simultaneous learning and selection.	-Lower computational cost compared to wrapper methods. - Features selected based on model's requirements.	-Model-dependent, which limits its applicability to other learning algorithms. - Can be biased toward specific models.	- Regularized models (e.g., Lasso, Ridge), decision trees, random forests. - High-dimensional datasets.
Hybrid methods	Combination of filter and	- Combines strengths of both filter and	- Combines computational efficiency	- More complex and computationally	- Complex tasks requiring high accuracy

	wrapper methods	wrapper methods. - Uses different evaluation criteria in different stages.	(filter) with high accuracy (wrapper). - Reduces redundancy and improves relevance.	intensive. - Requires careful tuning to avoid suboptimal subsets.	and computational efficiency. - Real-time machine learning tasks.

3.3. Label based feature selection

Feature selection methods based on tag information can be broadly categorized into three types: supervised, unsupervised, and semi-supervised methods, each with its own approach to evaluating the relevance of features based on the nature of the available data [23].

3.3.1.Supervised feature selection:

In supervised feature selection, the method relies on labeled data to identify the most relevant features. The key idea is to evaluate the importance of each feature by measuring its correlation with the labels (i.e., the class or target variable) [73]. These methods use statistical measures such as mutual information, correlation coefficients, or information gain to assess how well each feature distinguishes between different classes or outcomes. Since supervised feature selection utilizes label information, it tends to offer higher accuracy and performance compared to unsupervised methods because the model is guided by the target variable during the selection process. However, the main drawback of supervised methods is their dependence on sufficient labeled data, which is often costly and time-consuming to acquire. In many practical applications, labeled data is limited, while unlabeled data is abundant. This makes supervised feature selection impractical or infeasible in certain contexts, especially when labeling is expensive or time-consuming [74].

Supervised feature selection methods are mostly used in classification problems and use the correlation and relationship between features and labels as the basis for feature selection. The importance of features is measured by relevant criteria. Suppose there are $D(X, C)$ databases in which $X = \{x_1, x_2, \ldots, x_n\}$ are feature sets and C is the class label. The goal of a supervisory feature selection method is to find an optimal subset of X features in the form of $\hat{S}(\left|\hat{S} = K\right|)$ in such a way that the classification accuracy and accuracy are maximized. CFS correlation feature selection [52] Relief [53] methods of selecting supervisory features are among the significant methods in this category of methods. CFS is a feature subset selection method that works from hierarchical selection methods to obtain the optimal subset. Relief, which is an improved result of Relief, has worked well in multi-class classification. The main idea of this method is to use the Euclidean criterion and weight the features based on the differential sample in different classes.

Eq. (10) shows the Hilbert-Schmidt dependency criterion for feature selection. In this relation, the dependencies of a subset of the relation (10) show the Hilbert-Schmidt dependency criterion $J(0)$ for feature selection. In this regard, it shows the dependence of a subset of features on class C. A good subset of the feature maximizes $J(0)$ and as a result the feature selection problem is optimally solved. It maximizes a good subset of features, and as a result, the feature selection problem is optimally solved.

$$D_{FS} = \underset{\hat{S} \subseteq F}{argmax}\left[J(\hat{S})\right] \tag{10}$$

If value $\hat{f}(S)$ is put instead of $I(\hat{S};C)$ in relation (10), relation (11) is obtained. If value is placed in relation (10) instead of value.

$$D_{FS} = arg\,max\left[I\left(\hat{S};C\right)\right] \tag{11}$$

which $I(\hat{S};C)$ is calculated from Eq. (13).

$$I\left(\hat{S};C\right) = H(C) - H(C|\hat{S}) \tag{12}$$

In Eq. (12), $H(C)$ is a limited value. The highest value of $I\left(\hat{S};C\right)$ indicates the lowest value of $H\left(C;\hat{S}\right)$. Therefore, based on the Eq. (4), the maximization of $I\left(\hat{S};C\right)$ causes the decrease of $\rightleftarrows E\left(\hat{S}\right)$.

The central idea is the feature selection method in information $arg\,max\left[I\left(\hat{S};C\right)\right]$ theory. In the past, feature selection methods based on information theory have been very successful. Filter-based feature selection methods usually use evaluation criteria to improve the correlation between features and labels and reduce the correlation between features. Also, correlation between features is replaced by repetition or variation (distance). Correlation, repetition and variation (distance) criteria can be similar or different.

In Minimum redundancy and maximum relevance methods (mRMR) based on association and repetition, association (attribute-category) and repetition (attribute-attribute) are the basis of this class of methods. mRMR selects features that are highly relevant to the prediction target while avoiding features that carry overlapping or redundant information. For example, in a medical dataset, it may choose one biomarker from a group of highly correlated biomarkers instead of keeping them all [75]. These models use Euclidean distance criterion, Pearson's correlation criterion, as well as information theory for communication and repeatability. Features in the main database are divided into four groups.

- Fully connected and noiseless features
- Correlated weakness and repetitive features
- Weakly related and non-repetitive features
- Strong related features

Strong related properties are also called basic properties of hard theory, which form the core of conditional properties [54].

Relevance and redundancy analyze become two main optimization problems. A classical criterion for feature selection based on repetition or relationship is mRMR (least repetition and maximum relevance), which uses mutual information as an evaluation criterion. Eq. (13) shows the first-order analysis of repetition, and its extended form is given in Eq. (14) as mutual information of conditions.

$$mRMR = max\left[\frac{1}{|S|}\sum_{x_i\in S} I(x_i;S) - \frac{1}{|S|^2}\sum_{x_i\in S}\sum_{x_j\in S} I(x_i;x_j)\right] \tag{13}$$

$$CmRMR = max\left[\frac{1}{|S|}\sum_{x_i\in S} I(x_i;S) - \frac{1}{|S|^2}\sum_{x_i\in S}\sum_{x_j\in S}\left(I(x_i;x_j) - I\left(x_i;x_j|c\right)\right.\right] \tag{14}$$

Although mRMR based on mutual information only minimizes feature-feature mutual information and eliminates the classification efficiency of the desired features and can affect the selected features, feature selection based on mutual information Conditional has also been given a lot of attention. Feature selection

methods based on information theory increase information regardless of labels. Incremental hierarchical methods and approximate expression directly measure mutual information between subsets of features as well as labels. mRMR also has an incremental hierarchical structure [76].

3.3.2. Unsupervised feature selection:

In contrast, unsupervised feature selection methods do not rely on any label information. Instead, these methods evaluate features based on their ability to preserve the intrinsic structure of the data, such as maintaining variance or clustering patterns. For example, methods like Principal Component Analysis (PCA) or Independent Component Analysis (ICA) focus on selecting features that capture the most variance or the underlying statistical structure of the data without any regard to class labels. These methods are particularly useful in scenarios where labeled data is unavailable or scarce, as they do not require any supervision. However, unsupervised methods may not always result in the most relevant features for classification tasks because they do not directly consider how well features correlate with the target variable, potentially leading to suboptimal feature sets when the goal is predictive performance [77].

The main goal of unsupervised feature selection methods is to cover data classification and increase clustering accuracy by finding subsets of features based on clustering or evaluation criteria. Unsupervised feature selection methods are divided into unsupervised wrapper and unsupervised filter based on clustering algorithm.

Unsupervised wrapper feature selection uses clustering algorithm to set the validity of feature selection. The subset of features with the best clustering efficiency is considered as the most optimal subset of features. Since each subset of features must be evaluated by the clustering algorithm, it has a high computational cost. Because it will be involved with a large amount of data. Unsupervised rapper feature selection methods can be divided into two categories: 1- method based on local rapper, method based on global rapper. If the algorithm is applied to all features and clusters, it will be global and otherwise it will be local.

In unsupervised filter feature selection methods, a subset of features are selected based on the essence and nature of the data features, and learning and clustering algorithms are not implemented in the feature selection process. Therefore, the time of clustering and the complexity of the algorithm are reduced. Unsupervised filter feature selection methods directly use the statistical efficiency of all training data as an evaluation criterion, which seems to be very efficient and useful for large databases. The efficiency of clustering in selected features of unsupervised filter is usually lower than that of unsupervised wrapper. The main reason is that in the unsupervised filter, the evaluation criterion is independent from the specific clustering algorithm [78]

3.3.3. Semi-supervised feature selection:

Given the limitations of supervised and unsupervised methods, semi-supervised feature selection has emerged as a solution to handle situations where only a small portion of the data is labeled, but a large amount of unlabeled data is available. Semi-supervised methods aim to leverage the local structure of both labeled and unlabeled data to assess feature relevance. These methods typically use techniques that exploit the relationship between labeled and unlabeled instances, such as graph-based methods, which utilize the manifold structure of the data. Semi-supervised feature selection methods often aim to preserve the structure of the data while also taking into account the class information provided by the labeled data [31]. The advantage of semi-supervised methods is that they make use of both labeled and unlabeled data, allowing them to perform better in scenarios where labeled data is scarce but abundant unlabeled data is available.

This makes them particularly useful in real-world applications where acquiring labeled data is expensive or impractical, such as in medical diagnostics or social media data analysis [79].

Suppose the database $D = \{D_L, D_U\}$ is given. D_L is a subset of data that all have labels and D_u is data that does not have labels. Semi-supervised feature selection methods use D_u to improve learning efficiency with the help of D_L. In other words, in this category of methods, labeled data is used to improve the efficiency of feature selection in unlabeled data [34].

Semi-supervised feature selection methods are divided into three categories: semi-supervised wrapper, semi-supervised filter and semi-supervised integrated. Score functions used in semi-supervised filter methods are divided into four groups:

1. Constraint Score
2. Fisher's score
3. Variance score
4. Laplacian score

Limit and Fisher scores are supervised and variance and Laplacian scores are unsupervised [80]. Semi-supervised feature selection methods based on constraint score use binary constraint to evaluate the features and the features that have the best ability to maintain the constraint are selected. Semi-supervised feature selection methods based on Fisher's score take advantage of Fisher's features and use the local structure and information distribution of labeled and unlabeled data in feature selection in such a way that the discriminability and keeping the local information completely preserved.

In semi-supervised feature selection methods based on variance score, features that have the most variance are selected. Semi-supervised feature selection methods based on the Laplacian score combine the Laplacian criteria and the output information to select the feature. These methods are based on the Laplacian graph that constructs the neighborhood graph and evaluates the features based on the ability to preserve the local structure of the data [81].

3.3.4. Comparison of labeled based feature selection:

Table 4 provides a comparative overview of three major paradigms in feature selection: supervised, unsupervised, and semi-supervised approaches. Supervised methods take advantage of label information to select the most discriminative features, which makes them highly effective in classification tasks such as disease diagnosis or image recognition. However, their dependence on sufficient labeled data can limit their applicability in many medical and real-world settings. Unsupervised approaches, in contrast, operate without labels by exploiting patterns such as variance or intrinsic structure in the data. While this independence from labels is beneficial in exploratory analysis and clustering, it may reduce their predictive power in tasks that rely on accurate class separation. Semi-supervised methods represent a middle ground by incorporating both labeled and unlabeled samples, thereby improving performance when labeled data are limited but unlabeled data are abundant. Although computationally more demanding, these methods are particularly relevant to modern healthcare and web data applications where annotation is expensive, yet large-scale raw data are readily available.

Table 4: Summarizing the three types of feature selection methods

Feature selection method	Data type	Key idea	Advantages	Disadvantages	Common Usecases
Supervised feature selection	Labeled data	Selects features based on their correlation with the target labels	- High accuracy and performance due to label information - Suitable for classification tasks	- Requires sufficient labeled data, which can be costly and time-consuming - Not practical in the absence of labeled data	- Classification problems with sufficient labeled data - Medical diagnosis, image classification, etc.
Unsupervised feature selection	Unlabeled data	Selects features based on data variance or intrinsic structure	- Does not require labeled data - Useful when labeled data is scarce or unavailable	- May not select the most relevant features for predictive tasks - Cannot directly consider class labels	- Clustering, dimensionality reduction - Exploratory data analysis, feature extraction in unsupervised contexts
Semi-supervised feature selection	Both labeled and unlabeled data	Leverages both labeled and unlabeled data to assess feature relevance	- Makes use of both labeled and unlabeled data - Useful when labeled data is scarce but unlabeled data is abundant	- More complex and computationally intensive than supervised or unsupervised methods - May still suffer from limited label quality	- Applications with limited labeled data but abundant unlabeled data - Text mining, medical applications, web data analysis

4. Discussion

The review presented in this paper shows that feature selection remains a crucial step in medical data analysis, where balancing accuracy, computational efficiency, and interpretability is essential. While filter methods are valued for their scalability, their inability to capture feature interactions limits their use in complex healthcare datasets. Wrapper methods typically achieve higher predictive accuracy but are computationally expensive, whereas embedded approaches attempt to combine the strengths of both but may still face challenges with generalization.

Another important point is that the effectiveness of feature selection strongly depends on the characteristics of medical data. For instance, high-dimensional genetic data, imaging modalities, and electronic health records each bring unique problems such as redundancy, noise, and heterogeneity. This

indicates that no single approach is universally optimal, and hybrid or adaptive methods may provide more robust outcomes.

Additionally, the choice of evaluation measures significantly influences the assessment of feature selection techniques. Measures such as the Pearson correlation coefficient are computationally efficient but limited to detecting linear relationships. In contrast, mutual information can capture both linear and non-linear dependencies, offering richer insights at the expense of higher computational costs. Similarly, distance-based measures are intuitive but may lose discriminative power in high-dimensional settings, whereas entropy-based measures provide deeper analysis but require more complex computations. Such comparisons highlight that the evaluation process must be tailored to both the dataset and the application domain.

Finally, open challenges remain, including scalability for real-time medical applications, integration with deep learning models, and ensuring interpretability in clinical practice. Addressing these issues is critical for transforming feature selection research into practical, trustworthy tools that can support clinicians, enhance diagnostic accuracy, and advance personalized healthcare.

5- Conclusion

In this article, we provided a comprehensive review of the concepts and methods of feature selection, which play a fundamental role in improving the accuracy, efficiency, and interpretability of machine learning models, particularly in the medical domain. Methods were categorized based on their interaction with learning algorithms as well as the type of data (labeled, unlabeled, or semi-labeled). These categories highlight that the choice of feature selection technique strongly depends on dataset characteristics and available computational resources. In the context of medical data, the nature of the dataset—whether labeled or unlabeled, structured or unstructured—greatly influences the performance of learning models, making the choice of an appropriate feature selection method even more critical. Despite the progress made, several open challenges remain that define directions for future research. One key challenge is scalability, that is, developing feature selection methods that are effective for real-time medical applications such as continuous monitoring of vital signs. Another challenge is interpretability: the selected features must be understandable to clinicians to ensure transparency and reliability in medical decision-making. Furthermore, integration with deep learning architectures is an important avenue, as feature selection could reduce complexity and computational costs in these models. Finally, heterogeneity of medical data ranging from images to genomic information and electronic health records calls for more flexible and adaptive approaches. Additionally, one promising direction is the design of adaptive frameworks capable of adjusting to changing data conditions and clinical requirements. Such approaches could not only enhance computational efficiency but also make feature selection more practical in real clinical environments, such as decision-support systems and personalized medicine platforms. Moreover, integrating feature selection with emerging technologies like medical Internet of Things (IoT) and federated learning may open up new possibilities for scalable and secure medical applications in the future .In summary, feature selection remains an essential tool for medical data analysis, but achieving real-world clinical applicability requires further work on scalability, interpretability, and adaptability to complex and heterogeneous datasets. Addressing these challenges can pave the way for more accurate, efficient, and clinically meaningful applications of feature selection in healthcare informatic

Funding: No specific funding received for this research.

Data Availability: The data that support the findings of this study are available in the article at Ref. [4,37,47].

Conflicts of Interest: The authors declare no conflicts of interest regarding this study.

Authors contributions: Conceptualization: MK, ZK, MAA, NT; methodology: ZD, MHD, SAN, MAA, NT, validation: MK, ZK, SAN, MAA, NT; writing—original draft preparation, MK, ZK, MK, ZD, MHD, , NT; writing—review and editing: MK, ZK, MAA, NT; visualization: MK, SAN, MAA, NT; supervision: MK, ZK, MK, ZD,; project administration: HA; funding: MK, ZK, MK, ZD, , NT; The author had approved the final version.

AI Usage Disclosure: Artificial intelligence tool Grammarly was used only for language polishing and grammatical correction. All scientific and technical contents were created and validated by the authors

References

[1] Larabi-Marie-Sainte, S., Aburahmah, L., Almohaini R., et al., (2019). "Current techniques for diabetes prediction: review and case study". *Applied Sciences*, 9(21), 4604.

[2] Wang, A., Liu, H., Yang, J., et al., (2022). "Ensemble feature selection for stable biomarker identification and cancer classification from microarray expression data". *Computers in Biology and Medicine*, *142*, 105208.

[3] Bridgelall, R. (2024). "Alignment of unsupervised machine learning with human understanding: A case study of connected vehicle patents". *Applied Sciences*, *14*(2), 474.

[4] Doe, J., and Smith, J. (2024). "Application of machine learning in medical diagnostics". *Journal of Medical Informatics*, *12*(4), 123-134.

[5] Alaidany, A. A., and Mahdi, M. M. (2025). "A review of IoT-based wearable sensor systems for healthcare monitoring". American Journal of Pediatric Medicine and Health Sciences, 2(5), 132-159.

[6] Taylor, R. (2024). "Building reliable industrial systems: Structural and operational approaches". *Journal of Industrial Reliability and Maintenance*, 18(4), 250-270.

[7] Ali A, A., Ali K, M., Marwah M, M., et al., (2025). "A review of machine learning in banking risk management and possible research topics". *Journal of Engineering, Mechanics and Modern Architecture*, *4*(1), 50-57.

[8] Emadi, M., Karimi, M., and Davoudi, F. (2023). "A review on examination methods of types of working memory and cerebral cortex in EEG signals". *Majlesi Journal of Telecommunication Devices*, *12*(3), 161-169.

[9] Asgarnezhad, R., Alhameedawi, K. A. M., and Akram Mahfoud, H. (2023). "An effective model of autism spectrum disorder using machine learning". *Indonesian Journal of Electrical Engineering and Informatics*, 11(11), 43-54.

[10] Abbood, S. H., Hamed, H. N. A., Rahim, M. S. M., et al., (2022). "Hybrid retinal image enhancement algorithm for diabetic retinopathy diagnostic using deep learning model". *IEEE Access*, 10, 73079-73086

[11] S. Sonawane, S. (2024). "Enhanced feature optimization for multiclass intrusion detection in IOT fog computing environments". *Scalable Computing: Practice and Experience*, *25*(2), 1246-1263.

[12] Aliabadi, M., Karimi, M., Karimi, Z., et al., (2025). "The effect of photoplethysmography signal denoising on compression quality". *Iranian Journal of Electrical & Electronic Engineering*, 21(1), 1-14.

[13] Ali, H. (2025). A Meta-Review of Computational Intelligence Techniques for Early Autism Disorder Diagnosis. *International Journal of Theoretical & Applied Computational Intelligence*, 2025, 1–21

[14] Sofi, F. (2024). "Feature selection methods: A comparative study of filter, wrapper, and embedded approaches". *International Journal of Machine Learning and Data Analysis*, *12*(4), 123-145.

[15] Rajakani, M., Kavitha, R. J., and Rajesh, S. (2024). "Feature selection using modified chaotic satin bowerbird algorithm with deep transfer learning for multispectral image classification". *International Journal of Information Technology*.

[16] Malik, W., Javed, R., Tahir, F. et al., (2025). "COVID-19 Detection by Chest X-Ray Images through Efficient Neural Network Techniques". *International Journal of Theoretical & Applied Computational Intelligence*, vol. 2025, 35–56

[17] Lee, J., Song, M. S., Yoo, S. Y., et al., (2024). "Multimodal-based machine learning approach to classify features of internet gaming disorder and alcohol use disorder: A sensor-level and source-level resting-state electroencephalography activity and neuropsychological study". *Comprehensive Psychiatry*, *130*, 152460.

[18] Sun, L., Si, S., Ding, W., et al., (2023). "TFSFB: Two-stage feature selection via fusing fuzzy multi-neighborhood rough set with binary whale optimization for imbalanced data". *Information Fusion*, *95*, 91-108.

[19] Karimi, Z., Najafabadi, S. A., Nezhad, A. R., et al., "A big survey on biometrics for human identification", *Studies in Big Data*, Singapore: Springer Nature Singapore, 371-402, (2022).

[20] Kaur, S., Kumar, Y., Koul, A., et al., (2023). "A systematic review on metaheuristic optimization techniques for feature selections in disease diagnosis: Open issues and challenges". *Archives of Computational Methods in Engineering*, *30*(3), 1863-1895.

[21] P., D., A., B., and Ponniah, T. (2023). "Machine learning model for breast cancer data analysis using triplet feature selection algorithm". *IETE Journal of Research*, *69*(4), 1789-1799.

[22] Grünwald, P., "Introducing the minimum description length principle", *Advances in Minimum Description Length*, The MIT Press, 3-22, (2005).

[23] Gupta, K. K., Ritu Vijay, Pahadiya, P., et al., (2023). "Novel feature selection using machine learning algorithm for breast cancer screening of thermography images". *Wireless Personal Communications*, *131*(3), 1929-1956.

[24] Srividya, T. D., and Arulmozhi, V. "Feature selection classification of skin cancer using genetic algorithm". In 2018 3rd International Conference on Communication and Electronics Systems, IEEE 412-417, (2018).

[25] Deng, X., Li, Y., Weng, J., et al., (2019). "Feature selection for text classification: A review". *Multimedia Tools and Applications*, *78*(3), 3797-3816.

[26] Abed, M. H., Alsaeedi, A. H., Alfoudi, A. D., et al., (2020). "Palm vein identification based on hybrid features selection model".

[27] Khan, A. R., Doosti, F., Karimi, M., et al., (2021). "Authentication through gender classification from iris images using support vector machine". *Microscopy Research and Technique*, *84*(11), 2666-2676.

[28] Raftarai, A., Mahounaki, R. R., Harouni, M., et al., "Predictive models of hospital readmission rate using the improved adaboost in covid-19", *Intelligent Computing Applications for Covid-19*, Boca Raton: CRC Press, 67-86, (2021).

[29] Akram, S., Javed, M. Y., Qamar, U., et al., (2015). "Artificial neural network based classification of lungs nodule using hybrid features from computerized tomographic images". *Applied Mathematics &Information Sciences*, *9*(1), 183-195.

[30] Karimi, M., Harouni, M., Nasr, A., et al., "Automatic lung infection segmentation of covid-19 in CT scan images", *Intelligent Computing Applications for Covid-19*, Boca Raton: CRC Press, 235-253, (2021).

[31] Ibrahim, S., Nazir, S., and Velastin, S. A. (2021). "Feature selection using correlation analysis and principal component analysis for accurate breast cancer diagnosis". *Journal of Imaging*, *7*(11), 225.

[32] Song, F., Mei, D., and Li, H. "Feature selection based on linear discriminant analysis". In 2010 International Conference on Intelligent System Design and Engineering Application: IEEE 746-749, (2010).

[33] Jafari, P., Espandar, E., Baharifard, F., et al., "Linear local embedding", *Dimensionality Reduction in Machine Learning*, 129-156, (2025).

[34] Harouni, M., Karimi, M., Nasr, A., et al., "Health monitoring methods in heart diseases based on data mining approach: A directional review", *Studies in Big Data*, Singapore: Springer Nature Singapore, 115-159, (2022).

[35] Rehman, A., Harouni, M., Karimi, M., et al., (2022). "Microscopic retinal blood vessels detection and segmentation using support vector machine and k-nearest neighbors". *Microscopy Research and Technique*, *85*(5), 1899-1914.

[36] Karimi, E., Ebrahimi, A., and Tavakoli, M. R. (2019). "How optimal PMU placement can mitigate cascading outages blackouts?". *International Transactions on Electrical Energy Systems*, *29*(6), 1-13.

[37] Prasetiyo, B., Alamsyah, Muslim, M. A., et al., (2021). "Evaluation of feature selection using information gain and gain ratio on bank marketing classification using naïve bayes". *Journal of Physics: Conference Series*, *1918*(4), 042153.

[38] Sánchez-Hernández, S. E., Salido-Ruiz, R. A., Torres-Ramos, S., et al., (2022). "Evaluation of feature selection methods for classification of epileptic seizure EEG signals". *Sensors*, *22*(8), 3066.

[39] Haider, F., Pollak, S., Albert, P., et al., (2021). "Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods". *Computer Speech & Language*, *65*, 101119.

[40] Mahindru, A., and Sangal, A. (2020). "Dldroid: Feature selection based malware detection framework for android apps developed during covid-19". *Int J Emerg Technol*, *11*(3), 516-525.

[41] Chen, Q., Cao, F., Xing, Y., et al., (2023). "Evaluating classification model against bayes error rate". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(8), 9639-9653.

[42] Wu, J., Lu, L., and Tang, T. (2011). "Bayesian analysis for uncertainty and risk in a groundwater numerical model's prediction". *Human and Ecological Risk Assessment: an International Journal*, *17*(6), 1310-1331.

[43] Rehman, A., Harouni, M., Zogh, F., et al., (2024). "Detection of lungs tumors in CT scan images using convolutional neural networks". *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *21*(4), 769-777.

[44] Sharif, U., Mehmood, Z., Mahmood, T., et al., (2019). "Scene analysis and search using local features and support vector machine for effective content-based image retrieval". *Artificial Intelligence Review*, 52, 901-925.

[45] Hu, Y., Zhao, L., Li, Z., et al., (2022). "Classifying the multi-omics data of gastric cancer using a deep feature selection method". *Expert Systems with Applications*, *200*, 116813.

[46] Sadeghian, Z., Akbari, E., Nematzadeh, H., et al., (2025). "A review of feature selection methods based on meta-heuristic algorithms". *Journal of Experimental & Theoretical Artificial Intelligence*, *37*(1), 1-51.

[47] Alirezanejad, M., Enayatifar, R., Motameni, H., et al., (2020). "Heuristic filter feature selection methods for medical datasets". *Genomics*, *112*(2), 1173-1181.

[48] Oreski, S., and Oreski, G. (2014). "Genetic algorithm-based heuristic for feature selection in credit risk assessment". *Expert Systems with Applications*, *41*(4), 2052-2064.

[49] Ibrahim, R. A., Ewees, A. A., Oliva, D., et al., (2019). "Improved salp swarm algorithm based on particle swarm optimization for feature selection". *Journal of Ambient Intelligence and Humanized Computing*, *10*(8), 3155-3169.

[50] Nayar, N., Gautam, S., Singh, P., et al., "Ant colony optimization: A review of literature and application in feature selection", *Lecture Notes in Networks and Systems*, Singapore: Springer Nature Singapore, 285-297, (2021).

[51] Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., et al., (2022). "A review of feature selection methods for machine learning-based disease risk prediction". *Frontiers in Bioinformatics*, 2*, 1-17*.

[52] Uzma, Al-Obeidat, F., Tubaishat, A., et al., (2022). "Gene encoder: A feature selection technique through unsupervised deep learning-based clustering for large gene expression data". *Neural Computing and Applications*, *34*(11), 8309-8331.

[53] Ssekidde, P., Steven Eyobu, O., Han, D. S., et al., (2021). "Augmented cwt features for deep learning-based indoor localization using WIFI RSSI data". *Applied Sciences*, *11*(4), 1806.

[54] Wadhera, T., Bedi, J., and Sharma, S. (2023). "Autism spectrum disorder prediction using bidirectional stacked gated recurrent unit with time-distributor wrapper: An EEG study". *Neural Computing and Applications*, *35*(13), 9803-9818.

[55] Wang, A., An, N., Yang, J., et al., (2017). "Wrapper-based gene selection with Markov blanket". *Computers in Biology and Medicine*, 81, 11-23.

[56] Ismi, D. P., Panchoo, S., and Murinto, M. (2016). "K-means clustering based filter feature selection on high dimensional data". *International Journal of Advances in Intelligent Informatics*, *2*(1), 38.

[57] Cherrington, M., Thabtah, F., Lu, J., et al., "Feature selection: Filter methods performance challenges". In 2019 International Conference on Computer and Information Sciences, IEEE 1-4, (2019).

[58] Lyu, Y., Feng, Y., and Sakurai, K. (2023). "A survey on feature selection techniques based on filtering methods for cyber attack detection". *Information*, 14(3), 191.

[59] Peng, J., Wang, N., El-Latif, A. A. A., et al., "Finger-vein verification using Gabor filter and sift feature matching". In 2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing: IEEE 45-48, (2012).

[60] Bashir, S., Khattak, I. U., Khan, A., et al., (2022). "A novel feature selection method for classification of medical data using filters, wrappers, and embedded approaches". *Complexity*, *2022*(1).

[61] Jiang, K., Tang, J., Wang, Y., et al., (2020). "EEG feature selection via stacked deep embedded regression with joint sparsity". *Frontiers in Neuroscience*, *14*.

[62] Cancian, P., Di Donato, G. W., Rana, V., et al., "An embedded Gabor-based palm vein recognition system". In 2017 IEEE International Conference on Biomedical & Amp; Health Informatics, IEEE 405-408, (2017).

[63] Wang, Y., Grant, A. V., and Li, Y. (2023). "Implementation of a graph-embedded topic model for analysis of population-level electronic health records". *Star Protocols*, *4*(1), 101966.

[64] Jiménez-Navarro, M., Martínez-Ballesteros, M., Martínez-Álvarez, F., et al., (2024). "Explaining deep learning models for ozone pollution prediction via embedded feature selection". *Applied Soft Computing*, *157*, 111504.

[65] Zavrak, S., and Yilmaz, S. (2023). "Email spam detection using hierarchical attention hybrid deep learning method". *Expert Systems with Applications*, *233*, 120977.

[66] Hu, Y., Zhang, Y., Gao, X., et al., (2023). "A federated feature selection algorithm based on particle swarm optimization under privacy protection". *Knowledge-based Systems*, *260*, 110122.

[67] Lamba, R., Gulati, T., and Jain, A. (2022). "A hybrid feature selection approach for Parkinson's detection based on mutual information gain and recursive feature elimination". *Arabian Journal for Science and Engineering*, *47*(8), 10263-10276.

[68] Song, Y., Wang, J., Qi, Y., et al., (2025). "Serial filter-wrapper feature selection method with elite guided mutation strategy on cancer gene expression data". *Artificial Intelligence Review*, *58*(4).

[69] Khoirunnisa, A., Adytia, D., Deris, M. M., et al., (2025). "Comprehensive review of hybrid feature selection methods for microarray-based cancer detection". *IEEE Access*, *13*, 78739-78768.

[70] Zhang, R., Nie, F., Li, X., et al., (2019). "Feature selection with multi-view data: A survey". *Information Fusion*, *50*, 158-167.

[71] Abdolzadegan, D., Moattar, M. H., and Ghoshuni, M. (2020). "A robust method for early diagnosis of autism spectrum disorder from EEG signals based on feature selection and DBSCAN method". *Biocybernetics and Biomedical Engineering*, *40*(1), 482-493.

[72] Too, J., and Mirjalili, S. (2021). "A hyper learning binary dragonfly algorithm for feature selection: A covid-19 case study". *Knowledge-based Systems*, *212*, 106553.

[73] Gautheron, L., Redko, I., and Lartizien, C., "Feature selection for unsupervised domain adaptation using optimal transport", *Lecture Notes in Computer Science*, Cham: Springer International Publishing, 759-776, (2019).

[74] Feng, F., Li, K., Yang, E., et al., (2023). "A novel oversampling and feature selection hybrid algorithm for imbalanced data classification". *Multimedia Tools and Applications*, *82*(3), 3231-3267.

[75] Peng, H., and Ding, C. (2205). "Minimum redundancy and maximum relevance feature selection and recent advances in cancer classification". *Feature Selection for Data Mining*, *52*.

[76] Harouni, M., Karimi, M., and Rafieipour, S., "Precise segmentation techniques in various medical images", *Artificial Intelligence and Internet of Things*, Boca Raton: CRC Press, 117-166, (2021).

[77] Zhou, H., Xin, Y., and Li, S. (2023). "A diabetes prediction model based on Boruta feature selection and ensemble learning". *BMC Bioinformatics*, *24*(1).

[78] Emadi, M., Jafarian Dehkordi, Z., and Iranpour Mobarakeh, M. (2023). "Improving the accuracy of brain tumor identification in magnetic resonanceaging using super-pixel and fast primal dual algorithm". *International Journal of Engineering*, *36*(3), 505-512.

[79] Srijayanthi, S., and Sethukarasi, T. (2023). "Design of privacy preserving model based on clustering involved anonymization along with feature selection". *Computers & Security*, *126*, 103027.

[80] Agrawal, P., Abutarboush, H. F., Ganesh, T., et al., (2021). "Metaheuristic algorithms on feature selection: A survey of one decade of research (2009-2019)". *IEEE Access*, *9*, 26766-26791.

[81] Abualigah, L. M., Khader, A. T., and Hanandeh, E. S. (2018). "A new feature selection method to improve the document clustering using particle swarm optimization algorithm". *Journal of Computational Science*, *25*, 456-466.