Hindawi
Complexity
Volume 2022, Article ID 8190814, 12 pages
https://doi.org/10.1155/2022/8190814

Research Article

A Novel Feature Selection Method for Classification of Medical Data Using Filters, Wrappers, and Embedded Approaches

Saba Bashir,[1] **Irfan Ullah Khattak**,[2] **Aihab Khan**,[2] **Farhan Hassan Khan**,[3] **Abdullah Gani**,[4] **and Muhammad Shiraz**[1]

[1]*Department of Computer Science, Federal Urdu University of Arts Sciences and Technology, Islamabad, Pakistan*
[2]*Department of Computing and Technology, Iqra University, Islamabad, Pakistan*
[3]*Knowledge & Data Science Research Center (KDRC), Department of Computer Engineering, College of E & ME, National University of Sciences & Technology (NUST), Islamabad, Pakistan*
[4]*Faculty of Computing and Informatics, University Malaysia Sabah, Jalan UMS, Kota Kinabalu 88400, Malaysia*

Correspondence should be addressed to Saba Bashir; saba.bashir3000@gmail.com and Abdullah Gani; abdullahgani@ums.edu.my

Received 13 June 2022; Revised 7 July 2022; Accepted 21 July 2022; Published 8 August 2022

Academic Editor: Xiaoan Yan

Feature selection is the process of identifying the most relevant features from the given data having a large feature space. Microarray datasets are comprised of high-quality features and very few samples of data. Feature selection is performed on such datasets to identify the optimal feature subset. The major goal of feature selection is to improve the accuracy by identifying a minimal feature subset. For this purpose, the proposed research focused on analyzing and identifying effective feature selection algorithms. A novel framework is proposed which utilizes different feature selection methods from filters, wrappers, and embedded algorithms. Furthermore, classification is then performed on selected features to classify the data using a support vector machine (SVM) classifier. Two publically available benchmark datasets are used, i.e., the Microarray dataset and the Cleveland Heart Disease dataset, for experimentation and analysis, and they are archived from the UCI data repository. The performance of SVM is analyzed using accuracy, sensitivity, specificity, and f-measure. The accuracy of 94.45% and 91% is achieved on each dataset, respectively.

1. Introduction

A microarray dataset is a dataset that contains a large number of feature spaces where rows represent the number of records and columns represent the dimensions to be analyzed. As the dimensionality of data rises, the effort to attain the most relevant information from the data also increases. The main problem in large feature space datasets (i.e., microarray datasets) is the projection of a large feature space into a smaller one by preserving the information as much as possible. In a microarray dataset, each record can contain up to 450,000 features, and processing of a large amount of data can result in high computation costs [1].

Moreover, when the dimensionality of a dataset grows significantly, the sparsity of meaningful data also increases, which results in a lack of meaningful information. Datasets with large feature space and small records or observations tend to be prone to overfitting. The model built on overfitted data has a high fluctuation rate where a small change in data can result in a high classification error. Noisy features also play an important role in increasing difficulty. Noisy data is error data, which has a deviation pattern from the original value. Noisy data also affect the performance of machine learning algorithms. In order to reduce complexity and attain an efficient machine learning model, noisy data should be eliminated [1, 2].

The issues such as the complex nature of high-dimensional data, irrelevancy, overfitting, and high computational cost raise the need for feature selection methods. The feature selection method identifies the most relevant features from thousands of feature space by preserving the information. As a result, efficient classification results and reduced

computation time is achieved from the most relevant and small feature space [3–5].

The success of a mining algorithm depends on many factors, and task-relevant data is the most important factor among them. The quality of input data generates quality results. If a data mining algorithm is applied to irrelevant data, which contains redundant information and noise, then the results of the mining process would not be acceptable, and also the learning process would be complex on a large feature space [6].

Feature selection has become the focus of interest for many researchers, and its importance is increasing rapidly with the advancements in data mining. Feature selection is the process of selecting a subset of relevant features from a dataset having a large feature space, for example, the selection of variables or predictors for use in model building. Feature selection helps the data mining process in reducing the hypothesis space by reducing the irrelevant features from the given data [7, 8].

Three different types of techniques are used for feature selection, namely, filters, wrappers, and embedded approaches. Filter approaches do not take into consideration any classification algorithm. These approaches select the subset of features with the most relevant information [7]. Some evaluation techniques for filter approaches are relief, information gain, and chi-square [8]. A wrapper approach is a form of feature selection method that takes into consideration a classification algorithm that is used for the selection of a feature subset. An optimized set of feature space for a particular classification algorithm is the output of the wrapper approach. Examples of wrapper approaches are principal component analysis, genetic algorithms, decision trees, and particle swarm optimization. In embedded approaches, feature selection is a part of model fitting, and a particular technique is selected based on the model. Random forest, singular value decomposition, and probability approximation correct Bayes are some examples of embedded approaches. The solutions of the filter and wrapper approaches are said to be suboptimal as the external approaches do not take into consideration all possible combinations of features [9].

A feature selection method must carry the search on entire feature space by selecting different candidates of feature subsets. This search is completely independent of the evaluation measures. The search can be of three different types, namely complete, sequential or random. The complete search is also termed an exhaustive search, which guarantees the optimum results depending upon the particular evaluation measure. However, its drawback is that it is computationally expensive $\Theta(2n)$ when executed over a large feature space (n). The sequential search does not produce optimal solution/subset as it is based on the previous ranking generated by other techniques. However, its computational complexity is much better than exhaustive search which is $\Theta(n^2)$. The random search also does not guarantee optimal results as it starts with a random subset of feature space and executes in sequential search manner. The next subset is generated completely random and the process continuous until it reaches the final decision [10, 11].

Feature selection techniques are often used in domains where there are many features and comparatively few samples (or data points). Archetypal cases for the application of feature selection include the analysis of written texts and DNA microarray data, where there are many thousands of features and a few tens to hundreds of samples [12].

A small feature space improves classification efficiency and accuracy by reducing redundant and irrelevant information. Furthermore, the generalization capability of classifiers also increases as model over-fitting is reduced over small feature space. The model/classifier is trained on only the important features of the data without getting hung up on irrelevant features. Such models are mostly workable for the clinical environment. In disease profiling, a small feature set based gene microarray dataset can result in better classification and disease prediction results. There are a number of research methodologies introduced for reducing the feature space in gene microarray datasets. Their main goal is to achieve high disease classification and prediction accuracy. Another importance of feature selection is that it provides reduced feature space, which results in a reduced response time of the model.

There are a lot of existing databases for health care organizations in structured and semi-structured forms, such as images, radiology reports, medication profiles, signals, patient history, pathology reports, and treatment records. This type of data can be very heterogeneous, complex in nature, uncertain, contain noise, and have a large feature space. Therefore, data mining offers some methodologies that can extract useful hidden patterns and information from such databases efficiently by employing relevant feature selection methods and classification algorithms. These valuable patterns and information can be very helpful for clinical researchers and practitioners in making intelligent decisions and early prediction of diseases with high accuracy.

This study proposed a data mining classification model along with a novel feature selection method for the identification and progression of disease. The main contribution of this research is a classification model which incorporates different feature selection algorithms to introduce an optimized and accurate model which predicts the disease with high accuracy. The proposed model employed 27 different feature selection methods from the categories of filter, wrappers, and embedded approaches. These feature selection methods are executed on different medical datasets consisting of large feature space obtained from publicly available online data repositories. The top 3 feature selection methods are then identified from each feature selection category, i.e., filters, wrappers, and embedded, on the basis of high performance.

Finally, an efficient feature selection method has been proposed which combines the results of all the selected feature selection methods. Later on, a support vector machine classifier is employed on selected features to perform the classification. Different metrics like accuracy, sensitivity, specificity, and F-measure are used to evaluate the performance. The proposed feature selection method can be applied to any dataset for selection of efficient and most

relevant features from a large feature space. The classification method applied to the selected feature set produced high disease diagnosis accuracy and reduced execution time when compared with state-of-the-art frameworks.

The rest of the paper is organized as follows: Section 2 discusses the literature review performed on feature selection methods for large datasets; Section 3 elaborates proposed methodology in detail. Experiments, results, evaluation, and discussion are given in Section 4. Finally, Section 5 discusses the conclusion and future work.

2. Literature Review

The feature selection method is considered the most important step in machine learning-based models. A lot of research has been performed on feature selection algorithms using classification methods. These feature selection methods are categorized into three categories: filter approaches, wrapper approaches, and embedded approaches. Some of the state-of-the-art research in each category is given as follows.

Khan et al. [13] focused on the disease diagnosis issue and mentioned that radiologists can use the Gabor filter bank for feature extraction. The proposed model is based on the Gabor filter bank for the selection of appropriate features. The Cuckoo search method is employed as a metaheuristic algorithm. Most descriptive features are attained from the image samples. In the segmented region, there are high-dimensional features, and feature selection is used to obtain a subset of features. The DDSM database was used to obtain 2000 mammograms. The proposed model has achieved high performance when compared with other algorithms. However, the proposed framework can be improved by reducing its computational cost.

Another filter approach is used in [14], where microarray data classification is performed in combination with a clustering approach. The K-mean clustering method is employed to generate the clusters of features where features are categorized on the basis of the same characteristics. The best features are selected from each cluster, and then classification is performed on the selected features. A random forest is used for classification. The evaluation of the proposed model is performed on multiple datasets, such as colon, lung cancer, and prostate. The analysis of the results indicates the highest accuracy of 98.9%. The proposed approach, along with the clustering approach, has generated high performance results. The proposed framework can be improved by using an optimization algorithm to fine tune the parameters.

Feature selection is considered as a preprocessing step in order to improve the classification performance. Ke et al. [15] discussed that microarray datasets contain large feature space, small sample size, and mostly irrelevant features, which makes it difficult to contribute towards a model with high classification accuracy. The proposed approach focused on score-based criteria for feature selection (SCF). The Cancer dataset is used for prediction, and prediction performance is analyzed. The proposed model is evaluated on five gene microarray datasets and three low-dimensional datasets. Support vector machines and K nearest neighbors are used for classification. The analysis shows that the proposed SCF method can be used as a preprocessing step to perform the disease diagnosis with high accuracy. The proposed approach is complex in terms of weight assignment parameters. A score normalization method can be incorporated to overcome this limitation. The ensemble of filter methods is then introduced by Ghosh et al. [16] for the identification of important features from microarray datasets. The proposed methodology is based on two stages. In the first stage, the ensemble of filter methods was introduced based on the union and intersection of top n features, and then in the next stage, the genetic algorithm was applied to generate the results. The analysis shows that the union of features produced better results. The proposed method is independent of classifiers; therefore, three different classifiers like multilayer perceptron, support vector machine, and k nearest neighbor are used for classification. The evaluation was performed on five different cancer datasets and attained the highest accuracy as compared to state-of-the-art models. The proposed methodology can be further improved by using ensembles for classification instead of single classifiers.

Mostly, the feature extraction and selection processes are quite complicated and hazardous. To improve this process and make it easy to extract features and make learning time short, improvement is needed. Saw et al. [17] proposed a model based on wrapper-based feature selection, i.e., particle swarm optimization (PSO) to reduce the feature set. As a result, the classification accuracy of healthcare data is improved. Five medical datasets are used for experimentation and analysis and are downloaded from the UCI data repository. Multiple classification algorithms are used to perform the classification of data along with the PSO method. It is clear from the experiments that computational complexity is reduced and accuracy is improved by reducing the feature set. However, the proposed approach can be evaluated on gene microarray data to justify the results. Xue et al. [18] then proposed an ensemble-based wrapper feature selection for medical data. The proposed methodology is based on extreme learning machines (ELM) and genetic algorithms (GA) for classification. Simultaneous optimization is performed for both the feature subset and classifier parameters. ELM is modified using error mimization (EM) and a novel machine. EM-ELM is introduced to perform feature selection in a reasonable time. Ranking and selecting strategies are used to further refine the selection of features. Benchmark datasets are used for evaluation. Experimental results show that the proposed model has achieved better performance and execution time as compared to other models. Ensemble classification algorithms can be incorporated to further improve the classification performance.

In order to further improve the feature selection performance, González et al. [19] introduced a novel method based on the multiobjective wrapper method for feature selection. The proposed model performs analysis and classification of BCI applications. Over-fitting is avoided by selecting a small subset of features. The stability of the proposed wrapper method is analyzed using the feature ranking method. Four different classification methods are

used to evaluate the effectiveness of the proposed framework. It is analyzed that KNN has achieved the highest performance in terms of kappa values and proved more stable as compared to other classifiers. However, it has the limitation of execution time, which rises directly proportional to the increase in value of ok. K. Mustaqeem A. et al. [20] also proposed a wrapper-based feature selection method for the cardiac arrhythmia dataset. A random forest is used to select the best features from the arrhythmia dataset. Various machine learning classifiers such as Naïve Bayes (NB), support vector machine (SVM), K nearest neighbor (K-NN), and multilayer perceptron (MLP) are then used for classification based on selected features. The best accuracy is achieved by MLP as compared to other classifiers. Arrhythmia dataset is archived from UCI machine learning repository. However, the proposed technique can be evaluated on other diseases and microarray datasets to justify the performance.

Alzaqebah et al. [21] further introduced an improvement in the wrapper feature selection method. The proposed model is based on the neighborhood search method along with moth optimization for performing optimized feature selection. The moth-flame optimization (MFO) algorithm is used along with neighborhood to find the best features from a given large feature space. Premature convergence and local optima problems are also avoided by the proposed method. The modified MFO has achieved better results when compared with other population-based algorithms such as particle swarm optimization (PSO), firefly algorithm (FFA), and genetic algorithms (GA), etc. The evaluation is performed on 8 different medical datasets and performance is analyzed. The analysis of results shows that the proposed method outperforms when compared with other state-of-the-art methods.

The most recent advancement of the wrapper feature selection method was introduced by Sun et al. [22]. The proposed model is based on a combination of filter and wrapper methods. A common spatial pattern (CSP) based feature selection framework is proposed which performs evaluation of BCI datasets. An improved binary gravitational search algorithm (IBGSA) is proposed to find the optimal subset of features. The experimentation is performed on three publicly available BCI datasets and then MI classification is performed. The accuracy is compared with existing feature selection models, and it is analyzed that the proposed model outperforms on the same datasets. The proposed model can be improved by incorporating embedded methods to further improve the selection of optimized features. The summary of data mining algorithms using feature selection is shown in Table 1.

Although a number of feature selection methods have been used in decision support systems for medical datasets, there is still room for improvement. There is a need to optimize the combination of feature selection methods along with classifiers to produce high performance for medical datasets, specifically for datasets having a large number of feature space. The proposed framework focuses on an optimized combination of feature selection methods which produce high quality results for different types of datasets.

3. Materials and Methods

The prime focus of the proposed methodology is on feature selection. The proposed research identified optimized feature selection techniques that result in high performance. Multiple feature selection techniques are incorporated into clinical decision support systems, and a framework is proposed. The proposed framework can be used for feature selection and classification/prediction of any medical dataset. Figure 1 shows the detailed overview of the proposed framework.

The proposed framework is divided into following modules i.e., Data Extraction, Feature Selection and Classification and Experimental Evaluation. Each of these modules is discussed as follows.

3.1. Data Extraction. The first and foremost step of the proposed framework is data extraction. Two medical datasets are downloaded from UCI's online data repository. These datasets are named the Cleveland heart disease dataset and the Gene expression cancer microarray datasets. These two datasets are different in nature as they have different levels of attributes and instances. The reason behind using these datasets is to show the unbiased behavior of the proposed model such that its performance on large feature space and comparatively small feature space is comparable. Each dataset contains a large number of features as compared to the number of samples. The summary of these datasets is shown in Table 2.

3.2. Feature Selection. After data extraction, feature selection is performed to obtain the most relevant features for the classification task. The archived datasets contain a large number of feature space and all of these features do not contribute towards classification. Therefore, the most relevant features are selected to make the process efficient and to improve the accuracy. Multiple feature selection methods are evaluated from the filters, wrappers, and embedded category. Then the top 3 methods are selected from each category (filters, wrappers, and embedded) based on accuracy. The accuracy threshold is set to 80%, which reflects that, after selecting the top 3 feature selection methods from each category, features selected by each method are evaluated using an SVM classifier. Finally, those features will be selected as a final subset from each category where the SVM classifier will give accuracy greater than 80%. The following methods are used from the filter, wrapper, and embedded feature selection methods category based on literature.

3.3. Filter Approaches. Following filter approaches are used in the proposed methodology for feature set selection.

3.3.1. Correlation Based Feature Selection (CFS). Correlation based feature selection [39] is a multivariate FS method that has been successfully applied to many classification tasks. CFS evaluates the features based on the individual's performance of prediction along with the certainty

Table 1: Summary of data mining algorithms using feature selection on medical datasets.

Reference	Year	Techniques	Datasets	Accuracy	Limitations
Ghosh et al [23]	2021	Relief, LASSO	Cleveland, long beach, Hungarian, stat log	99.05%	Classification algorithm is dependent on feature selection method.
Alirezanejad et al. [24]	2020	Heuristic methods	Colon, leukemia	92%	Meta heuristics can be applied to remove unnecessary attribute prior to classification
Khan et al. [25]	2019	Gabor filter bank + SVM	DDSM database	92.48%	The proposed system computational cost is much higher and they proposed further research to optimize the computational cost
Lü et al. [26]	2019	RBM + SVM	MNIST handwritten database	81.87%	Due to the determinacy of configuration parameters, the RBM feature extraction is not feasible, which needs to be improved in further research.
Shrinivas D. Desai et al. [27]	2019	BPNN + LR	Cleveland dataset	78.88%	The proposed model cannot be used as a clinical expert, it only complements the decision of clinician for taking better diagnostic decisions
Vijayashree et al. [28]	2018	PSO + SVM	Cleveland heart disease	—	The proposed model can further be improved using ensemble classifiers.
Kalantaria et al. [29]	2018	GA-SVM	California at Irvine (UCI) machine learning repository	84.44%	The proposed system further optimization needs in terms of achieving high performance in detection on medical datasets.
Dwivedi et al. [30]	2018	SVM	Statlog heart disease dataset	90%	The proposed system cannot be used for the predication of disease levels.
Jianguo Chena et al. [31]	2018	Disease diagnosis and treatment recommendation system	PubMed dataset	90%	The security prospective is not been addressed. Feature selection is not considered.
Tayefi, et al. [32]	2017	Decision tree, hs-CRP	UCI dataset	94%	Due to some risk factors in diabetic patients, the proposed model does not consider some key factors to evaluate the system for high performance.
Hoque et al. [33]	2016	Decision tree Random forests Naïve bayes kNN SVM	UCI dataset	71.2% 83.12% 28.60% 94.50% 51.14%	To incorporate incremental fuzzy feature selection technique for classification of DDoS attack traffic.
Bennasar et al. [34]	2015	mRMR JMIM NJMIM	Sonar datasets	88% 87% 86%	It disregards the interaction between the features and the classifier, as well as the higher dimensional joint mutual information between more than two features, which sometimes can lead to a suboptimal choice of features.
EMary et al. [35]	2015	Gray wolf optimization	UCI dataset	—	Improvement can be made using advanced feature selection method.
Veronica et al. [36]	2015	ReliefF Information gain mRMR CFS FCBF Chi-Squared	Micro array datasets	90.24% 82.32% 65.23% 65.36% 55.21% 81.2%	The developed techniques should be tested on multiplatform, distributed learning, and real-time processing. It provides a new line of research for researchers to work on datasets with numerous increases in dataset use in feature selection.
Zhang et al. [37]	2014	Sequential forward floating selection (SFFS) SBS MBPSO	Spam based data set	94.07% 95.28% 91.97%	Slow processing, used only decision tree for classification
Verónica et al. [38]	2014	Correlation-based feature selection (CFS) Chi-square Minimum redundancy maximum relevance Support vector machine	Brain (UCI) CNS (UCI) GLI (UCI)	66.67% 65.00% 69.41% 96.99%	To distribute the microarray data vertically (i.e., by features) in order to reduce the heavy computational burden when applying wrapper methods.

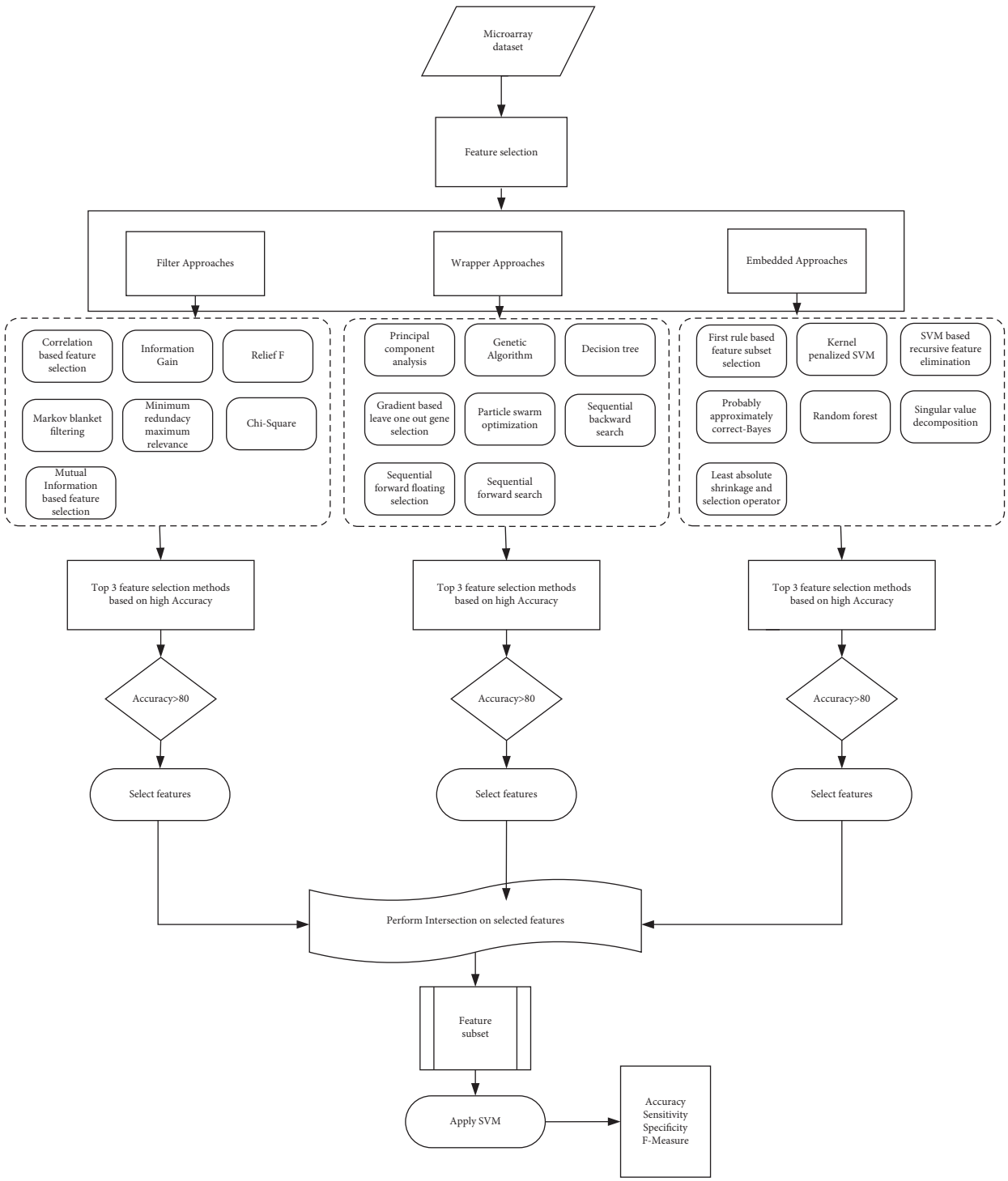

Figure 1: Proposed framework for feature selection and classification.

Table 2: Summary of medical datasets.

Datasets	Samples	Features	Access link
Cleveland heart disease	303	75	https://archive.ics.uci.edu/ml/datasets/heart + disease
Gene expression cancer dataset	801	20531	https://archive.ics.uci.edu/ml/datasets/gene + expression + cancer + RNA-seq

of redundancy between them. The correlation between subsets of attributes and class is determined by the correlation coefficient. Equation for CFS is given as

$$r_{ac} = \frac{k\overline{r_{zi}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}}, \quad (1)$$

where r_{ac} represents the correction between feature set and class attributes, k shows number of features, r_{zi} is average of correlation between class attribute and feature subset, and r_{ii} shows intercorrelation between feature subset.

3.3.2. Information Gain (IG). Information gain [40] is a univariate method which selects those branches of a decision tree that most count towards predication. Ultimately, those features are selected from the dataset that contain the most information. The following formulas are used to calculate the information gain of features, and those that contain the highest value are selected as a subset.

$$\begin{aligned} \text{Gain}(S_j) &= E(P_j) - E(S_j), \\ E(P) &= \sum_{I=1}^{n} p_i \log_2 p_i, \\ E(S_J) &= \sum_{I=1}^{nj} I_j * E(Y)_j. \end{aligned} \quad (2)$$

3.3.3. ReliefF. ReliefF [41] is a multivariate feature selection method which selects features based on their nearest neighbors. Feature weights are calculated using a convex optimization problem. Following formula is incorporated for measuring relief and feature subset selection.

$$W_i = W_i - (x_i - \text{nearHit}_i)^2 + (x_i - \text{nearMiss}_i)^2, \quad (3)$$

where nearHit is the closest same class instance where as nearMiss is the closest different class instance. X is a feature vector and w represents the weight.

3.4. Wrapper Approaches. Following wrapper approaches are used for features evaluation.

3.4.1. Genetic Algorithm (GA). A genetic algorithm [42] generates a number of solutions/populations from given features and evaluates every possible solution as a whole. Fitness function, crossover, and mutation are applied and a final population is selected, providing an optimized solution. GA can be used for a large number of feature sets and outperforms traditional FS methods.

3.4.2. Decision Tree. Decision trees [43] are widely used for feature ranking based on split node attributes. The importance of is analyzed by their ranking. There are a number of decision tree algorithms for FS and ranking, such as ID3, CART, and C4.5, etc. The most relevant features are selected as a subset and used for further classification.

3.4.3. Sequential Forward Search (SFS). SFS [44] is used for feature sunset selection where relevant features are selected iteratively until a desired set of features is obtained. The inclusion equation is given as follows:

$$x^+ = \text{argmax} J_{x \in Y_k}(Y_k + x). \quad (4)$$

3.5. Embedded Approaches. Proposed methodology comprised of following embedded approaches.

3.5.1. Probably Approximately Correct (PAC)-Bayes. PAC Bayes [45] is a probability-based method which identifies the relationship between features and class label. A feature that has a minimum probability for a particular class is eliminated and an iterative process is applied for all feature space.

3.5.2. Random Forest. Random forest [45] is a highly accurate method for FS and has better generalization. 4–12 hundred multiple tress are constructed based on random selection of feature subset. The importance of features depends on how pure a node is. The most important features are finally identified.

3.5.3. Singular Value Decomposition (SVD). SVD [45] is used for reducing the features of a given feature space. Approximated computations are widely used for unsupervised feature reduction using the following equation:

$$A \approx A_k = U_{m*k} \sum_{k*k} V^T_{k*n}, \quad (5)$$

where rank r is greater than k, different k are generated as compared to r. It is assumed that first k is larger than the discarded ones.

4. Classification

Once feature selection is performed on a given feature set, a subset of features is identified. This feature subset is now used for the classification of medical datasets. The classification method will have improved performance and high accuracy as it is learned on a refined feature subset. Linear classifiers generate results based on a linear combination of features/values generated by selected features. A few examples of this category include Fisher discriminant analysis, Naïve Bayes, etc., whereas SVM constructs maximum margin hyperplane. A decision tree uses the feature space to make a decision. Recursive partitioning is used where the classification process is applied to feature space recursively and it stops when there is no change in prediction value due to partitioning.

4.1. SVM for Classification. An extensive literature survey has been performed to select the appropriate algorithm for medical data classification. Initially, 150 research papers have been downloaded. Then, only those papers are identified that used classification algorithms for medical datasets.

```
Input F: Original feature set
   N: Size of population
   D: Dimensions of feature
   O: Optimal feature subset
For each F = 1-N Do:
While (each feature selection method not evaluated)
      Evaluate each feature using feature selection methods
         Calculate weight of each feature
         Select top features based on weight
         Evaluate each feature selection method on selected features
            Calculate Accuracy, Sensitivity, Specificity, F-Measure
            Select Top 3 methods based on Accuracy
               If feature selection method Accuracy >80
                  Then select feature selection method
                     Select features subset D of each feature selection method
                     Perform intersection on selected features subsets D
                     Return O: Optimal feature subset
               Else ignore
      End While
      End For
      Output O: Optimal feature subset
```

Algorithm 1: The algorithm of proposed feature selection method is given as follows:

Next, screening has been performed to check which classification algorithm performed better. A literature review shows that SVM outperforms other classification methods for gene expression and other datasets for classification. There are a number of reasons for this, such as SVM not only classifies the dataset into classes correctly but also identifies those samples/instances whose class label is missing, such as outliers, etc. Strong mathematical background of SVM makes them fit for gene expression dataset classification, for example, they are flexible in selecting the similarity function. In case of large datasets, it provides sparse solutions, and most importantly, it is capable of handling large feature space. Therefore, the proposed framework employs the SVM algorithm for classification.

The filter, wrapper, and embedded approaches are utilized to perform feature selection. The proposed research applied these methods to multiple gene expression microarray datasets and calculated the accuracy. The top 3 feature selection methods from each FS category are selected based on high accuracy. Now, there are top 9 feature selection methods i.e., 3 feature selection methods from each category having accuracy greater than threshold value. Common features are selected from them by performing intersection. Finally, Support Vector Machine is applied on common feature set. The proposed classifier, along with the selected features subset, has attained marginal performance and attained the highest accuracy. Multiple evaluation matrices are used to evaluate the performance of the proposed SVM with feature selection such as accuracy, sensitivity, specificity, and F-measure.

5. Experimental Evaluation

The proposed framework is evaluated on multiple medical datasets. These datasets are obtained from online data repositories. Multiple filter, wrapper, and embedded feature selection methods are used for evaluation.

5.1. Cleveland Heart Disease Dataset. The table below shows filter, wrapper, and embedded classifier implementation in the Cleveland heart disease dataset. The result of SVM classifier implementation is calculated. The performance of each feature selection method is evaluated using accuracy, sensitivity, specificity and F-measure. Tables 3–5 show the top 3 feature selection methods based on their performance.

The proposed framework is initially evaluated using 7 different feature selection methods, and these methods are selected based on literature. Then the top 3 methods are selected based on accuracy. The idea behind their selection is that if feature selection methods will be of high quality, then the results generated by these methods will also be of high quality. Then these methods are evaluated using an SVM classifier and the results are shown in Tables 3–5. It is clear from the results that most of the methods have high performance results. ReliefF, information gain, and correlation-based feature selection methods are shown in Table 3 because these methods have high performance as compared to other filter methods. Similarly, sequential forward selection, genetic algorithm, and decision tree have high performance as compared to other wrapper methods, as shown in Table 4. In Table 5, singular value decomposition, random forest, and probably approximately correct are shown as they have high performance as compared to other embedded approaches.

Moreover, accuracy, sensitivity, specificity, and f-measure are calculated for each method to show the performance.

Table 6 shows results of SVM classifier on intersected features selected by the top 9 feature selection methods from different categories, as shown in Tables 3–5. Each feature

Table 3: Top 3 feature selection methods (filters) based on high performance.

S.no	Filters	Accuracy (%)	Sensitivity (%)	Specificity (%)	*F*-measure (%)
1	ReliefF	91.09	75.93	99.49	87.71
2	Information gain	90.76	75.93	97.62	86.77
3	Correlation-based feature selection	90.76	75.85	97.64	86.74

Table 4: Top 3 feature selection methods (wrapper) based on high performance.

S.no	Wrapper	Accuracy (%)	Sensitivity (%)	Specificity (%)	*F*-measure (%)
4	Sequential forward selection	88.5	73.05	72.06	72.55
5	Genetic algorithm	86.5	73.8	71.6	72.7
6	Decision tree	64.02	96.36	85.48	90.92

Table 5: Top 3 feature selection methods (embedded) based on high performance.

S.no	Embedded	Accuracy (%)	Sensitivity (%)	Specificity	*F*-measure (%)
7	Singular value decomposition	64.37	100	64.36%	82.18
8	Random forest	89.44	84.24	86.95%%	85.95
9	Probably approximately correct	65.65	19.96	57.69%	38.82

Table 6: SVM performance on most relevant features using intersection.

S.no	Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	*F*-measure
1	SVM	91	79	82	80.5

Table 7: Top 3 Filter approaches performance for gene microarray dataset.

S.no	Filters	Accuracy (%)	Sensitivity (%)	Specificity (%)	*F*-measure (%)
1	ReliefF	91.09	75.93	99.49	87.70
2	Information gain	89.63	75.93	97.62	86.77
3	Correlation-based feature selection	90.55	75.85	97.64	86.74

Table 8: Top 3 Wrapper approaches performance for gene microarray dataset.

S.no	Wrapper	Accuracy (%)	Sensitivity (%)	Specificity (%)	*F*-measure (%)
4	Sequential forward selection	87.3	72.04	72.06	72.05
5	Genetic algorithm	86.4	70.8	71.6	70.57
6	Decision tree	64.02	93.33	85.48	89.04

selection method selects some set of features, and then these features are intersected and common features are identified as a subset. Most relevant features are identified as common features, and then an SVM classifier is applied on commonly selected features. The performance of SVM is shown by accuracy, sensitivity, specificity, and F-measure.

5.2. Gene Micro Array Dataset. Filters, wrappers, and embedded approaches are also applied on gene dataset attributes, and the most relevant features are selected by each method. Accuracy, sensitivity, specificity, and f-measure are calculated and the top 3 feature selection methods are identified from each feature selection category. Finally, SVM is applied on intersection of selected features. Tables 7–9 show the implementation results.

The analysis of the results indicates that most of the feature selection methods have high performance results. Table 7 shows ReliefF, information gain, and correlation-based feature selection methods. These methods have high performance as compared to other filters. Similarly, Table 8 shows sequential forward selection, genetic algorithm, and decision tree as these methods have high performance as compared to other wrapper methods. Table 9 shows singular value decomposition, random forest, and probably approximately correct as they have high performance when compared with other embedded approaches. Next, these methods are evaluated against a threshold value, and the subset of features is selected from those methods which have an accuracy greater than 80%.

Table 10 shows the SVM result on intersected/common features used in all top three techniques (Filters, Wrappers, and Embedded). Accuracy, sensitivity, specificity, and F-measure are then calculated. High accuracy of 94.45% is achieved when it is applied on features subset.

Table 11 shows the result of the proposed feature selection method using the SVM classifier for two medical datasets. Two different types and range of datasets have been

Table 9: Top 3 Embedded approaches performance for gene microarray dataset.

S.no	Embedded	Accuracy (%)	Sensitivity (%)	Specificity	*F*-measure (%)
7	Singular value decomposition	64.89	98.02	63.33%	80.67
8	Random forest	88.82	83.14	85.75%%	84.44
9	Probably approximately correct	64.91	18.56	56.59%	37.57

Table 10: Performance of SVM classifier after intersection on Gene Microarray dataset.

S.no	Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	*F*-measure (%)
1	SVM	94.45	78.25	83.56	80.90

Table 11: Performance of proposed framework using medical datasets.

S.no	Data sets	Accuracy (%)	Sensitivity (%)	Specificity (%)	*F*-measure (%)
1	Cleveland	91	79	82	80.5
2	Gene micro array	94.45	78.25	83.56	80.90

used to show the effectiveness of the proposed model. The results indicate that high performance is achieved while feature selection is used with the SVM classifier. The accuracy of 91% is achieved for the Cleveland dataset, which has small feature space and has 75 features, whereas 94.45% accuracy is achieved for the Gene microarray dataset, which has 20531 features. The proposed model shows an acceptable level of performance for each type of dataset, which indicates that the model is stable. It can be applied on any kind/size of dataset and its performance remains stable.

High performance is achieved by the SVM classifier because of the best and the most relevant features used in the dataset. Most relevant features are selected in different stages, such as when initially the most relevant feature selection methods are selected based on literature, then the top 3 feature selection methods are selected based on high accuracy, and next, those methods are selected which have an accuracy greater than 80%. After that, intersection of features is performed to further scrutinize the features, and these features are already selected by high performance methods. Finally, the SVM classifier is executed on the most relevant features, and this is the reason behind achieving high performance results for each dataset.

6. Conclusions and Future Work

In this research study, the proposed classification model successfully identified the most efficient feature selection methods for medical datasets. Four performance measure criterions (accuracy, sensitivity, specificity, and f-measure) are used along with SVM in this study.

Initially, 9 different feature selection methods are selected from three feature selection categories (i.e., filters, wrapper, and embedded) based on the research review, which was mostly used in medical sciences. Further, those nine methods were classified by implementing weight factor using rapid miner to get the weight of each attribute, and then an SVM classifier was used to get the performance factors. The Gain SVM classifier is selected based on a literature survey as most of the research was conducted using the SVM classifier and generated high results.

Finally, to get better performance, the intersection of common attributes selected by each feature selection method is applied, and then SVM is implemented on the intersected/common attributes to get the performance factor that further evaluates the proposed model.

There are many other feature selection techniques available in the literature to implement in different ways to select the best possible features in the medical domain. The proposed methodology covered three different techniques (filters, wrappers, and embedded) in the medical domain for feature selection. This research work can be further extended using evolutionary algorithm-based feature selection algorithms, such as particle swarm optimization, etc. The proposed framework can be implemented in other fields of medical sciences as well. [46–52].

Data Availability

The datasets used in experimentation are available at the following links: https://archive.ics.uci.edu/ml/datasets/heart+disease and https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

References

[1] J. Jonnagaddala, S. T. Liaw, P. Ray, M. Kumar, H. J. Dai, and C. Y. Hsu, "Identification and progression of heart disease risk factors in diabetic patients from longitudinal electronic health records," *Bio Medical Research International*, vol. 2015, Article ID 636371, 2015.

[2] D. Yach, C. Hawkes, C. L. Gould, and K. J. Hofman, "The global burden of chronic diseases," *JAMA*, vol. 291, no. 21, pp. 2616–2622, 2004.

[3] J. G. Cleland, K. Swedberg, and F. Follath, "The Euro Heart Failure survey program, a survey on the quality of care among patients with heart failure in Europe. Part 1: patient characteristics and diagnosis," *European Heart Journal*, vol. 63, 2003.

[4] V. L. Roger, A. S. Go, and D. M. Lloyd-Jones, "Heart disease and stroke statistics–2012," *A report from the American Heart Association*, vol. 220, 2012.

[5] U. M. Mogensen, M. Ersboll, and M. Andersen, "Clinical characteristics and major comorbidities in heart failure patients more than 85 years of age compared with younger age groups," *European Journal of Heart Failure*, vol. 23, 2011.

[6] J. Muntwyler, A. Cohen-Solal, N. Freemantle, J. Eastaugh, J. G. Cleland, and F. Follath, "Relation of sex, age and concomitant diseases to drug prescription for heart failure in primary care in Europe," *European Journal of Heart Failure*, vol. 8, 2004.

[7] P. A. McCullough, E. F. Philbin, and J. A. Spertus, "Confirmation of a Heart Failure Epidemic: Findings from the Resource Utilization Among Congestive Heart Failure," *(REACH) study*, vol. 39, 2002.

[8] J. J. McMurray, S. Adamopoulos, and S. D. Anker, "ESC Guidelines for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2012: The Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2012 of the European Society of Cardiology," *European Heart Journal*, vol. 33, 2012.

[9] S. M. Dunlay, N. L. Pereira, and S. S. Kushwaha, "Contemporary strategies in the diagnosis and management of heart failure," *Mayo Clinic Proceedings*, vol. 89, no. 5, pp. 662–676, 2014.

[10] W. Ouwerkerk, A. A. Voors, and A. H. Zwinderman, "Factors influencing the predictive power of models for predicting mortality and/or heart failure hospitalization in patients with heart failure," *Journal of the American College of Cardiology: Heart Failure*, vol. 2, no. 5, pp. 429–436, April 15, 2014.

[11] Y. Gerber, S. A. Weston, M. M. Redfiled et al., "A contemporary appraisal of the heart failure epidemic in Olmsted County, Minnesota, 2000 to 2010," *JAMA Internal Medicine*, vol. 175, 2015.

[12] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, *The Accuracy of Machine Learning (ML) Forecasting Methods versus Statistical Ones: Extending the Results of the M#-Competition*, Working Paper, University of Nicosia, Institute for the Future, Greece, 2017.

[13] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," *Expert Systems with Applications*, vol. 36, no. 4, pp. 7675–7680, 2009.

[14] S. Bashir, U. Qamar, and F. H. Khan, "IntelliHealth: a medical decision support application using a novel weighted multilayer classifier ensemble framework," *Journal of Biomedical Informatics*, vol. 59, pp. 185–200, 2016.

[15] J. C. Prather, D. F Lobach, L. K Goodwin, J. W Hales, M. L Hage, and W. E Hammond, "Medical data mining: knowledge discovery in a clinical data warehouse," *Proceedings of the AMIA Annual Fall Symposium*, p. 101, 1997.

[16] N Diaz-Diaz, J. S. Aguilar-Ruiz, J. A. Nepomuceno, and J. Garcia, "Feature selection based on bootstrapping," in *Proceedings of the 2005 ICSC Congress on Computational Intelligence Methods and Applications*, p. 6, IEEE, Istanbul, Turkey, December 2005.

[17] B. Jiang, C. Li, M. D. Rijke, X. Yao, and H. Chen, "Probabilistic feature selection and classification vector machine," *ACM Transactions on Knowledge Discovery from Data*, vol. 13, no. 2, pp. 1–27, 2019.

[18] J. Agor and O. Y. Özaltın, "Feature selection for classification models via bilevel optimization," *Computers & Operations Research*, vol. 106, pp. 156–168, 2019.

[19] C. Wan and A. A. Freitas, "An empirical evaluation of hierarchical feature selection methods for classification in bioinformatics datasets with gene ontology-based features," *Artificial Intelligence Review*, vol. 50, no. 2, pp. 201–240, 2018.

[20] R. B. Pereira, A. Plastino, B. Zadrozny, and L. H. C. MerschmannMerschmann, "Categorizing feature selection methods for multi-label classification," *Artificial Intelligence Review*, vol. 49, no. 1, pp. 57–78, 2018.

[21] R. Ruiz, J. Riquelme, and J. Aguilar-Ruiz, "Projection-based measure for efficient feature selection," *Journal of Intelligent and Fuzzy Systems*, vol. 12, pp. 175–183, 2003.

[22] J. Doak, "An Evaluation of Feature Selection Methods and Their Apllication to Computer Security," *Univ. of California at Davis, Dept. Computer Science*, vol. 105, 1992.

[23] M. Alirezanejad, R. Enayatifar, H. Motameni, and H. Nematzadeh, "Heuristic filter feature selection methods for medical datasets," *Genomics*, vol. 112, no. 2, pp. 1173–1181, 2020.

[24] P. Ghosh, S. Azam, M. Jonkman et al., "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021.

[25] S. Khan, A. Khan, M. Maqsood, F. Aadil, and M. A. Ghazanfar, "Optimized gabor feature extraction for mass classification using cuckoo search for big data e-healthcare," *Journal of Grid Computing*, vol. 17, no. 2, pp. 239–254, 2019.

[26] X. Lu, L. Meng, C. Chen, and P. Wang, "Fuzzy Removing Redundancy Restricted Boltzmann Machine: improving learning speed and classification accuracy," *IEEE Transactions on Fuzzy Systems*, vol. 28, p. 1, 2019.

[27] S. D. Desai, S. Giraddi, P. Narayankar, N. R. Pudakalakatti, and S. Sulegaon, "Back-propagation neural network versus logistic regression in heart disease classification," in *Advanced Computing and Communication Technologies*, pp. 133–144, Springer, Singapore, 2019.

[28] J. Vijayashree and H. P. Sultana, "A machine learning framework for feature selection in heart disease classification using improved particle swarm optimization with support vector machine classifier," *Programming and Computer Software*, vol. 44, no. 6, pp. 388–397, 2018.

[29] A. Kalantari, A Kamsin, S Shamshirband, A Gani, R Alinejad, and A. T. Chronopoulos, "Computational intelligence approaches for classification of medical data: state-of-the-art, future challenges and research directions," *Neurocomputing*, vol. 276, pp. 2–22, 2018.

[30] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Computing & Applications*, vol. 29, no. 10, pp. 685–693, 2018.

[31] J. Chen, B LiRong, and K. YangLi, "A disease diagnosis and treatment recommendation system based on big data mining and cloud computing," *Information Sciences*, vol. 435, pp. 124–149, 2018.

[32] M. Tayefi, T Mohammad, S Sara et al., "hs-CRP is strongly associated with coronary heart disease (CHD): a data mining approach using decision tree algorithm," *Computer Methods and Programs in Biomedicine*, vol. 141, pp. 105–109, 2017.

[33] N. Hoque, B Ahmed, and J. Kalita, "A fuzzy mutual information-based feature selection method for classification,"

Fuzzy Information and Engineering, vol. 8, no. 3, pp. 355–384, 2016.

[34] M. Bennasar, Y. Hicks, and R. Setchi, "Feature selection using joint mutual information maximisation," *Expert Systems with Applications*, vol. 42, no. 22, pp. 8520–8532, 2015.

[35] E. Emary, H. M. Zawbaa, C. Grosan, and A. E. Hassenian, "Feature subset selection approach by gray-wolf optimization," in *Proceedings of the Afro-European Conference for Industrial Advancement*, pp. 1–13, Cham, Switzerland, 2015.

[36] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "Recent advances and emerging challenges of feature selection in the context of big data," *Knowledge-Based Systems*, vol. 86, pp. 33–45, 2015.

[37] Y. Zhang and G. WangPhillipsJi, "Binary PSO with mutation operator for feature selection using decision tree applied to spam detection," *Knowledge-Based Systems*, vol. 64, pp. 22–31, 2014.

[38] V. Bolón-Canedo, N Sánchez-Maroño, A Alonso-Betanzos, J.M Benítez, and F Herrera, "A review of microarray datasets and applied feature selection methods," *Information Sciences*, vol. 282, pp. 111–135, 2014.

[39] R. J. Palma-Mendoza, L. de-Marcos, D. Rodriguez, and A. Alonso-Betanzos, "Distributed correlation-based feature selection in spark," *Information Sciences*, vol. 496, pp. 287–299, 2019.

[40] T. A. Alhaj, M. M. Siraj, A. Zainal, H. T. Elshoush, and F. Elhaj, "Feature selection using information gain for improved structural-based alert correlation," *PLoS One*, vol. 11, no. 11, Article ID e0166017, 2016.

[41] R. J. Palma-Mendoza, D. Rodriguez, and L. de-Marcos, "Distributed ReliefF-based feature selection in spark," *Knowledge and Information Systems*, vol. 57, no. 1, pp. 1–20, 2018.

[42] Y. Srinivasa MurthySrinivasa Murthy and S. G. Koolagudi, "Classification of vocal and non-vocal segments in audio clips using genetic algorithm based feature selection (GAFS)," *Expert Systems with Applications*, vol. 106, pp. 77–91, 2018.

[43] S. J. Lee, Z. Xu, T. Li, and Y. Yang, "A novel bagging C4.5 algorithm based on wrapper feature selection for supporting wise clinical decision making," *Journal of Biomedical Informatics*, vol. 78, pp. 144–155, 2018.

[44] J. Lee, D. Park, and C. Lee, "Feature selection algorithm for intrusions detection system using sequential forward search and random forest classifier," *KSII Transactions on Internet & Information Systems*, vol. 11, no. 10, 2017.

[45] I Guyon, S Gunn, and M Nikravesh, L. A Zadeh, *Feature Extraction: Foundations and Applications*, Springer, vol. 207, 2008.

[46] Y. Hu, Y. Zhang, and D. Gong, "Multiobjective particle swarm optimization for feature selection with fuzzy cost," *IEEE Transactions on Cybernetics*, vol. 51, no. 2, pp. 874–888, 2021.

[47] K. Chen, B. Xue, M. Zhang, and F. Zhou, "An evolutionary multitasking-based feature selection method for high-dimensional classification," *IEEE Transactions on Cybernetics*, vol. 52, no. 7, pp. 7172–7186, 2022.

[48] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction," *Journal of Applied Science and Technology Trends*, vol. 1, no. 2, pp. 56–70, 2020.

[49] N. Abd-Alsabour, "On the role of dimensionality reduction," *Journal of Computers*, vol. 13, no. 5, pp. 571–579, 2018.

[50] M. Verleysen and D. François, "The curse of dimensionality in data mining and time series prediction," *Computational Intelligence and Bioinspired Systems. In International Work-Conference on Artificial Neural Networks*, Springer, Berlin, Heidelberg, pp. 758–770, 2005, June.

[51] D. L. Padmaja and B. Vishnuvardhan, "Comparative study of feature subset selection methods for dimensionality reduction on scientific data," in *Proceedings of the 2016 IEEE 6th International Conference on Advanced Computing (IACC)*, pp. 31–34, IEEE, Bhimavaram, India, February 2016.

[52] J. Jin Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.