Feature Selection in Medical Data as Coping Review from 2017 to 2022

Sara S. Emam, Mona M. Arafa, Noha E. El-Attar, and Tarek Elshishtawy

Abstract The number of medical applications with large datasets that require great speed and accuracy is continually growing. A large number of features in medical datasets is one of the most critical issues in data classification and prediction models. Furthermore, irrelevant and redundant features have also harmed the complexity and functioning of data classification systems. Feature selection is a reliable dimensionality reduction strategy for identifying a subset of valuable and non-redundant features from massive datasets. This paper reviews the state-of-the-art feature selection techniques on medical data in the last five years.

Keywords High-dimensional dataset · Feature selection · Classification

1 Introduction

Massive data expansion in medical domains has made medical data mining methods challenging. To make medical diagnoses, treatments, predictions, and prognostic schedules on time, doctors and professionals in the field of medicine must examine a vast amount of medical data. As a result, it is critical to provide an intelligent model that can accurately handle an enormous amount of medical data. Therefore, intelligent and machine learning-based techniques have become increasingly important

S. S. Emam (✉) · M. M. Arafa · N. E. El-Attar · T. Elshishtawy
Information Systems Department, Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt
e-mail: sara.samir@fci.bu.edu.eg

M. M. Arafa
e-mail: mona.abdelmonem@fci.bu.edu.eg

N. E. El-Attar
e-mail: noha.ezzat@fci.bu.edu.eg

T. Elshishtawy
e-mail: T.shishtawy@fci.bu.edu.eg

D. Magdi et al. (eds.), *Green Sustainability: Towards Innovative Digital Transformation*, Lecture Notes in Networks and Systems 753,
https://doi.org/10.1007/978-981-99-4764-5_10

in medical health care. In various areas of health care, such as diagnosis, screening, prognosis, monitoring, therapy, survival analysis, and hospital management, machine learning classification algorithms are used in the decision-making process. However, machine learning faces a considerable challenge when dealing with medical datasets with a high-dimensional feature space and a limited number of samples [1]. Many features are used to represent data, but only a handful of them are relevant to the desired concept. Thus, the original datasets may have redundancy, which is not required to be included in the modeling process. Dimensionality reduction is one popular strategy for removing irrelevant, redundant, and insignificant features. It is a practical way to increase accuracy, reduce computational complexity, create more generalized models, and reduce storage requirements [1]. Two key strategies for reducing dimensionality have recently been developed: feature extraction and feature selection. Individual features or feature subsets are not searched for during feature extraction. Instead, feature extraction converts the original feature set from a higher to a lower dimensional space. The features are not chosen; instead, they are projected onto a new feature area [2]. Principal component analysis (PCA) is an example of feature extraction. Feature selection is choosing a subset of relevant features to create enhanced prediction models.

1.1 Feature Selection

Feature selection is a preprocessing technique that selects the most critical and relevant features, which may enhance machine learning performance by removing redundant or unnecessary features. As a result of the application of feature selection, modeling accuracy is improved, while the overall computing cost is reduced [2]. Furthermore, feature selection provides various advantages, including [3]:

- Improving the machine learning algorithm's performance.
- Data comprehension, including learning about the process and possibly assisting with visualizations.
- Data compression, limiting storage requirements and lowering processing costs.
- Simplicity and the ability to use simpler models and gain speed.

Feature Selection Approaches

In general, there are three feature selection approaches: filter, wrapper, and embedded, as shown in Fig. 1 [4].

In the *filter approach* algorithms, the classifier is independent. Thus, the feature selection and learning models are also separate. Information Gain (IG), correlation coefficients, Relief method, Relief-F (RF), Fisher score method, Chi-squared (CS), and Gain Ratio are examples of filter approaches. Generally, the filter selection algorithms do not use interrelationships between features to evaluate features [5]. Instead, they employ a scoring method that determines the statistical score of each feature and ranks the most likely highest. The higher a feature's score, the more

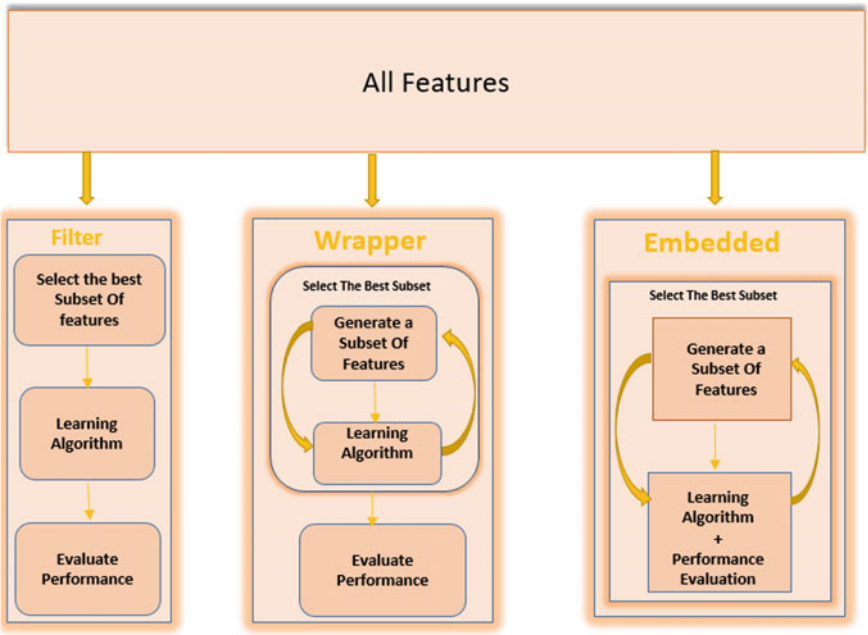

Fig. 1 Feature selection approaches [4]

likely it is to be chosen [2]. The main problem of this approach is that it ignores feature dependencies and relationships across classifiers, resulting in an incorrectly classified model [4].

In contrast, *wrapper-based feature selection* is classifier-dependent. It searches for a combination of features, each of which is referred to as a feature subset. Subsets can be used for prediction, and performance is determined using some metric. The selected feature subset has the best performance metric [2]. Several algorithms are used as wrapper feature selections, such as forward selection, backward elimination, and recursive feature elimination.

Finally, *the embedded model* is primarily concerned with identifying features that rate highly in terms of accuracy. The learning and feature selection processes are inextricably linked, and the feature search process is included in the classification algorithm. Examples of embedded methods are Lasso and Ridge regression algorithms [5]. Wrapper and embedded methods are frequently more accurate at classification than filter methods, although they take longer. As a result, several researchers have proposed hybrid strategies for identifying the essential features [4].

Optimization Algorithms in Feature Selection

Optimization algorithms aim to find the optimal solution for a well-defined problem. It is an iterative process that compares various solutions until an optimum or satisfying

one is discovered. Optimization methods are used in multiple fields to identify solutions that maximize or minimize specific study criteria, such as reducing expenses in manufacturing a good or service, maximizing profits, minimizing raw materials in developing a good, or maximizing efficiency. Metaheuristics is an example of an optimization strategy that involves simulating the behaviors of physical phenomena and live creatures to create a general-purpose optimization search framework independent of the task at hand [6]. Genetic algorithms (GA) and evolution strategies such as particle swarm optimization, ant colony optimization, and bee colony optimization are examples of metaheuristics. In recent years, optimization algorithms have been a powerful approach to feature selection. This approach gets the essential feature subset that achieves high classification accuracy according to determined objective function criteria. This approach improves results in most cases than traditional approaches of feature selection. Optimization techniques may be used alone, such as using the brain storm optimization (BSO) in [7], the improved teacher–learner-based optimization (ITLBO) algorithm in [8], and marine predators algorithm (MPA) in [9]. Otherwise, the optimization technique can be hybridized with other traditional feature selection techniques such as in [10] which combined rough set, chaos theory, and binary gray wolf optimization to produce the RS-CBGWO-FS model. In [11], an ensemble of multifilters algorithms such as IG, GR, CS, and RF has been used to utilize a harmonized classification technique based on PSO and SVM.

1.2 Medical Data

Clinical (medical) data include narrative, textual data (HPI, social/family Hx), and numerical measurements (laboratory results, vital signs, and measurement).

- Laboratory tests are a well-known type of medical data. Several types of data can be driven from laboratory tests, like:
  - Hematological or blood tests: This type of test is considered one of the most popular laboratory procedures performed to examine and analyze the hemic system. Microscopes and hematologic analyzers are used in these tests to look at the concentration of HBC in the blood flow (i.e., the oxygen levels in the blood flow), the white blood cell (WBC) count, the red blood cell (RBC) count, the number of platelets (PLTs), the iron concentration, and the number of erythrocytes and leukocytes. Hematological tests are also used to assess and monitor several diseases and disorders, including the prothrombin time and thrombin time, hematocrit (HCT), blood sedimentation, blood coagulation time, fibrin clot lysis time, and bone marrow, among others [12].
  - Urine tests: This test checks the urine using a urinalysis method. Urinalysis examines and analyzes the flow of urine, the gravity of urine, the levels of urine color, and the presence of germs and cellular debris using chemical screening tests and microscopes. Urine tests are commonly used to diagnose kidney and

liver disease, as well as diabetes and prostate cancer, urinary tract infections, and prostatic hypertrophy [12].
  - Histopathological (histological) testing: These tests examine the various types of tissues that indicate the nature of the disease (e.g., muscular, neural, epithelial). The most common way to diagnose cancer is via histopathological tests. A tissue sample is first collected in the least invasive method feasible for a biopsy test, with the amount of the recovered tissue sample varied according to the tissue area under review. A histological report can design a more tailored pharmacological treatment for a malignancy (or metastases) by providing information on the tumor type, hormone responsiveness, and other tumor markers [12].
  - Skin tests: These examinations are used to check the changes in the skin. An allergy, a skin ailment, or even skin cancer could cause these changes. For example, skin tests are commonly used to detect skin redness caused by enlarged blood vessels, non-blanching hemorrhages (such as purpura and palpable purpura), skin carcinoma, and skin lesions that could progress to skin cancer as allergies via a skin prick test [12].

- The vital signs of a live organism are an objective measure of its essential physiological functioning. They are "vital" because assessing and measuring them is the first and most crucial stage in any clinical evaluation. An assessment of the patient's vital signs constitutes the initial set of clinical examinations. Vital signs are the foundation for patient triage in an emergency department or urgent care setting because they can show a doctor how far a patient has deviated from the norm. Temperature, pulse, blood pressure, and respiratory rate are the traditional vital signs. Even though many other indicators besides the standard four vital sign parameters may also be helpful, studies have only demonstrated that pulse oximetry and smoking status are significantly related to patient outcomes [13].

2 Literature Review

This section presents different feature selection approaches on clinical (biomedical) datasets.

In [14], Li et al. have developed a novel prediction framework for medical diagnostics, known as IGWO-KELM, by combining an enhanced gray wolf optimization (IGWO) and kernel extreme learning machine (KELM). It consists of two primary sections. By adaptively looking for the best feature combination in the medical data, IGWO is used to weed out the redundant and irrelevant information in the first stage. In the suggested IGWO, GA is initially used to produce the population's initial positions, and GWO is then used to update the population's current positions in the discrete searching space. Based on the ideal feature subset obtained in the first step, the second stage conducts the effective and efficient KELM classifier. Two typical medical diagnosis issues, including the diagnosis of Parkinson's illness and breast cancer, were looked into in order to assess the proposed IGWO-KELM methodology.

The datasets of them are available in UCI. On the two common disease diagnosis issues, the new solution is contrasted with the original GA and GWO using a set of performance indicators. The results have demonstrated that the proposed method is superior to its two rival competitors.

In [7], Tuba et al. have suggested the brain storm optimization approach for feature selection and optimizing SVM parameters to classify medical datasets. The BSO method was modified for a binary solution to accomplish feature selection. The fitness function combines the number of features selected and classification accuracy. Three well-known medical datasets related to Hepatitis, liver disorders, and diabetes were used to test the suggested methodology. Results were compared to existing state-of-the-art approaches. The proposed method improved classification accuracy for the three datasets under consideration while using fewer features or preserving the same amount of features. The datasets used are Hepatitis (which has two classes, 19 features, and 155 instances), liver (which has two classes, seven features, and 345 instances), and diabetes (which has seven classes, eight features, and 768 cases).

Identifying the ideal feature subset using a feature selection method that is independent of the governing parameters of an algorithm tailored to the particular situation at hand is a difficult task. As a result, in [8], Manonmani et al. have introduced an algorithm based on the original TLBO algorithm's operating principle, which does not call for any algorithm-specific parameters. The improved teacher–learner-based optimization (ITLBO) algorithm, which is the name of the proposed research project, aimed to choose the best feature subset based on the Chebyshev distance formula in the evaluation of the fitness function and standard control parameters (i.e., population size and several generations) to find the ideal feature subset for early diagnosis of chronic diseases. The chronic kidney disease (CKD) dataset was used to test the proposed feature selection technique, resulting in a significant feature reduction of 36% compared to the 25% obtained using the original TLBO algorithm. Furthermore, by assessing the accuracy of classification algorithms (support vector machine (SVM), convolutional neural networks (CNNs), and gradient boosting), the generated optimal feature subset produced from the TLBO algorithm and the feature subset obtained from the ITLBO algorithm are validated. According to experimental results, the suggested feature selection algorithm improves overall classification accuracy for the resulting feature subset compared to the original TLBO approach.

In [10], Azar et al. have introduced a robust hybrid dynamic model for feature selection called RS-CBGWO-FS. Rough set (RS), chaos theory, and binary gray wolf optimization are combined in RS-CBGWO-FS (BGWO) to select the ideal number of features and accomplish an effective classification procedure in the medical area. Ten different chaotic maps are used to estimate and fine-tune GWO parameters. Before moving on to the classification and feature selection process, the process of handling missing values and the max–min normalization process are applied to medical datasets. This proposed strategy is tested on five complex datasets retrieved from the UCI repository. The overall result shows that RS-CBGWO-FS with the Singer and piecewise chaos maps offers greater efficacy, less error, faster convergence, and shorter computation times. The dataset used is cervical cancer (which has 36 attributes and 858 instances), dermatology (which has 33 features and 366

cases), diabetic retinopathy (which has 20 attributes and 1151 instances), arrhythmia (which has 279 attributes and 452 cases), and sonar (which has 60 features and 208 instances).

In [15], Spencer et al. use relevant features selected using various feature selection approaches to evaluate the performance of models produced using machine learning techniques. First, principal component analysis, Chi-squared testing, Relief-F, and symmetrical uncertainty have all been used to examine four widely used heart disease datasets to produce different feature sets. Then, to increase the accuracy of heart condition predictions, various classification algorithms have been employed to develop models that are compared to find the best feature combinations. The advantages of applying feature selection differ based on the machine learning technique utilized to analyze the heart datasets. However, the best model we produced used the BayesNet method and Chi-squared feature selection to reach an accuracy of 85.00% on the datasets under consideration. The dataset combines four heart disease datasets from the UCI ML repository (Cleveland Dataset, Long-Beach-VA Dataset, Hungarian Dataset, and Switzerland Dataset). The new combined dataset has fourteen features and 720 cases.

In [2], Shah et al. have introduced an automatic methodology for diagnosing clinical heart disease. By utilizing feature selection and extraction techniques, the suggested method calculates the essential feature subset. First, mean Fisher-based feature selection algorithm (MFFSA) and accuracy-based feature selection algorithm (AFSA) are introduced to carry out the feature selection. The feature extraction method, principal component analysis, is then used to refine the chosen feature subset further. The suggested approach has been tested with Cleveland, Hungarian, and Switzerland data and a combination of the three. Radial basis function kernel-based support vector machines classify humans as either heart disease patients (HDP) or standard control subjects (NCS). Accuracy, specificity, and sensitivity measures assess the suggested methodology. This paper used three datasets from UCI (Cleveland, Hungarian, and Switzerland). The original dataset comprised 76 features, 14 of which were chosen, including the class label. The following are the details and descriptions of the datasets: Cleveland (which consists of 13 features and 303 instances), Hungarian (which consists of 13 features and 283 cases), Switzerland (which consists of 13 features and 123 instances), and combined (which comprised of 13 features and 709 cases).

In [1], Rostami et al. have combined the multi-objective PSO algorithm and the node centrality methodology to create a feature selection method called MPSONC. This approach is classified as filter-based model feature selection, and its optimization process considers relevance and redundancy ideas. The MPSONC procedure is broken down into three stages: (1) graph presentation, (2) computation of node centrality, and (3) final feature selection utilizing a multi-objective PSO search algorithm. Converting the feature space into an undirected, weighted graph is the goal of the first step. A node in this representation represents each feature, and the weight of each edge reflects the similarity of their associated features. To determine feature popularity, the second phase of the suggested method applies the node centrality criterion to every feature. The initial population in the PSO method will be created

using the node centrality criterion in this step. Finally, the most essential and non-redundant features are selected in the last phase using a novel PSO-based search algorithm. Instead of using a single-objective fitness function to evaluate the generated particles as in many earlier PSO-based feature selection approaches, the innovative approach used in this study considers a feature subset utilizing a mix of feature separability index, similarity, and feature subset size. The proposed strategy is creative and performs better than the previous ones for three reasons: Three criteria of relevance, redundancy, and subset size of the chosen feature are considered in the fitness function of the proposed PSO-based technique. To illustrate the performances of the proposed strategy, five medical datasets with various properties are used. The results demonstrated that the introduced method is more efficient and effective than similar prior methods.

In most cases, the data dimensionality and classifier parameters significantly impact the accuracy of a diagnosis system. Because these two procedures are dependent, performing them separately could reduce accuracy. Based on ranking, the filter algorithm is employed to remove unimportant features. On the other hand, independent filters can still not account for feature dependency, resulting in an imbalanced selection of significant features and, as a result, a reduction in classification performance. To address this issue, in [11], Hamid et al. used an ensemble of multi-filters algorithms such as Information Gain (IG), Gain Ratio (GR), Chi-squared (CS), and Relief-F (RF), which takes into account feature intercorrelation. However, kernel parameter values may also influence classification performance. As a result, a harmonized classification technique based on PSO and SVM is used to optimize the simultaneous search for the best relevant features and kernel parameters while maintaining accuracy. As a result, this research proposed an ensemble filter feature selection with PSO and SVM harmonized classification (ensemble-PSO-SVM). The efficiency of the suggested strategy is evaluated using common lymphography and breast cancer datasets compared to other approaches already in use, such as PSO-SVM and standard SVM. Experimental findings show that the suggested method successfully indicates the classifier accuracy performance with the best essential features. As a result, the recommended method can be used as a substitute for selecting the best solution for dealing with high-dimensional data. Two datasets from UCI are used to verify the efficiency of the suggested model. The first dataset is breast cancer, which has 286 cases, nine features, and two predicting classes: class recurrence event and class no-recurrence event. The second dataset is lymphography, which has 148 instances, represented by 18 features and four predictive classes: standard, metastases, malignant, and fibrosis.

In [16], Bania et al. have used the feature-class, feature-feature rough dependence, and feature-significance measures to present a new rough set theory (RST)-based heterogeneous EFS approach (R-HEFS) to select the less repeated and more essential features during the aggregation of varied feature subsets. As a base feature selector, R-HEFS employs five state-of-the-art RST-based filter techniques. The experiments use ten standard medical datasets from the UCI repository. In addition, the *k*-nearest neighbor (KNN) imputation approach and RST-based discretization techniques are used for missing value imputation and continuous feature discretization. They use

four classifiers, namely random forest (RF), Naive Bayes (NB), AdaBoost, and support vector machine (SVM). The effectiveness of the proposed R-HEFS technique is assessed and studied. By eliminating irrelevant and redundant features during the aggregation of base feature selectors, the suggested R-HEFS technique proves to be efficient and helps to improve classification accuracy. R-HEFS has obtained improved average classification accuracy on 7 out of 10 diverse medical datasets. As a result, the overall findings strongly show that the suggested R-HEFS method can minimize the dimension of substantial medical datasets, potentially assisting physicians or medical specialists in diagnosing (classifying) various diseases with fewer computational difficulties. The datasets for cancer, heart, skin, liver, thyroid, and cardiac illnesses were gathered from the UCI machine learning repository. And they have a medium to a high level of complexity.

In [5], Omuya et al. presented a hybrid filter approach based on principal component analysis (PCA) and Information Gain for feature selection. By building the main components of the dataset, PCA allows datasets with many linked features to be reduced in size so that the present data can be stated with fewer variables. It is performed by determining the most significant primary components by assessing the association between features. Information Gain Evaluation: This stage uses Information Gain (IG) to examine the feature set selected above to find the most relevant attributes. The final feature set is chosen based on a predetermined threshold, and the IG for features is calculated (t). After that, using machine learning techniques such as the Naive Bayes methodology, the hybrid model is used to support classification (classify breast cancer data). According to experimental results, the hybrid filter model picks relevant feature sets, decreases training time, and minimizes data dimensionality, resulting in higher classification performance as assessed by accuracy, recall, and precision. The dataset used in this paper is the breast cancer dataset. It was created by Zwitter and Soklic of the Institute of Oncology University Medical Center and Ljubljana, Yugoslavia. It has nine features that can be used to detect the existence or onset of cancer.

In [17], Pavithra and Jayalakshmi have proposed a hybrid feature selection technique HRFLC, which combines random forest (RF), AdaBoost (AD), and Pearson coefficient (PC). A subset of features is chosen based on the previous three techniques, and accuracy will be tested for several models. The model's results demonstrate that it effectively predicts diseases and enhances prediction accuracy—the dataset used in this paper was taken from UCI repository. Dataset (heart disease dataset) contains 13 features and includes 280 patient records, 10 of which have missing values that are eliminated during data preprocessing. The dataset is a binary classification problem, with 1 indicating heart illness and 0 indicating no heart disease. The dataset is balanced with 120 heart disease patients and 150 records of those without heart disease patients.

In [9], Elminaam et al. presented a new method for reducing dimension in feature selection. In a seminal attempt, this paper selects the appropriate feature subset to increase classification accuracy using binary variations of the recent marine predators algorithm (MPA). MPA is a brand-new metaheuristic inspired by nature. This study offers the MPA-KNN method, a mix of MPA and k-nearest neighbors (KNN). On

medical datasets with feature sizes varying from small to huge, KNN is utilized to evaluate the selected features. The suggested methods are compared to eight well-respected metaheuristic wrapper-based algorithms and tested on 18 well-known UCI medical dataset benchmarks. In MPA, the fundamental exploratory and exploitative processes are modified to choose the best and most significant features for the most accurate classification. The findings show that the suggested MPA-KNN strategy can select the most relevant and optimal features. Furthermore, it outperformed the well-known metaheuristic algorithms that were put to the test. On average, MPA-KNN outperforms all other datasets in terms of accuracy, sensitivity, and specificity.

Based on symptoms and data from patients' electronic medical records, in [4] El-Attar et al. have presented a new multilayer perceptron (MLP) with feature selection (MLPFS) to predict positive COVID-19 instances (EMR). The MLPFS model comprises a layer that determines the most valuable symptoms to reduce the number of symptoms based on their relative value. Using only the most informative symptoms when training the model can hasten to learn and improve accuracy. Three separate COVID-19 datasets and eight different models, including the suggested MLPFS, were used in the experiments. According to the results, MLPFS outperforms all other experimental models in feature reduction across all datasets. It also performs better than the other models regarding classification outcomes and processing speed. In this research, three types of clinical reports served as datasets. The SARS-CoV-2 RT-PCR and further laboratory testing carried out on about 6000 COVID-19 cases during their visits to the emergency room were used to create this dataset. It has one class label and 109 features. Clinical features for symptomatic and asymptomatic individuals are included in the second COVID-19 dataset. There are 34,475 records in this dataset, each with one class label and 41 features. The third dataset uses clinical information to forecast the intensive care unit (ICU) admission for COVID-19 positive cases. There are 1926 cases total, 228 features, and 1 class label.

In [18], Piri and Mohapatra have proposed a multi-objective quadratic binary Harris Hawk optimizer for dealing with the feature selection issue in medical data. The continuous MOHHO is changed to a binary version using four quadratic transfer functions to make the approach practical for the FS problem. As a measure of each Hawk's fitness, two objective functions—the number of features in the candidate feature subset and the KNN classifier's classification accuracy—are considered. The four versions of the proposed MOQBHHO are implemented to extract the best feature subsets. Finally, the crowding distance (CD) value is used as a third criterion for selecting the best non-dominated option. Twelve standard medical datasets are used in this study to measure the performance of the suggested technique. MOBHHO-S (with a sigmoid function), MOGA, MOALO, and NSGA-II are all compared to the proposed MOQBHHO. Compared to deep-based FS approaches, the experimental results reveal that the suggested MOQBHHO effectively discovers a set of non-dominated feature subsets. The used datasets are BreastCancerW, Arrhythmia, Diabetic, Hepatitis, ILPD, Cardiotocography, Lymphography, LungCancer, Primary tumor, Parkinsons, Colon tumor, and SRBCT. The first ten datasets are from the UCI library, and the last two high-dimensional datasets are from Ref [18].

In [19], Gutowski et al. have provided a novel MOFS technique for binary medical classification. It is based on a genetic algorithm and a three-dimensional compass, intended to direct the search to the desired trade-off between the number of features, accuracy, area under the ROC curve (AUC), and accuracy. On several real-world medical datasets, our approach—the genetic algorithm with multi-objective compass (GAwC)—performs better than any other genetic algorithm-based MOFS technique. Furthermore, GAwC guarantees the classification quality of its solution by including AUC as one of the objectives, making it a particularly intriguing method for medical situations where both healthy and ill patients need to be reliably diagnosed. Finally, GAwC is used to solve a real-world medical classification problem, and the results are analyzed and supported from both a medical and a classification quality perspective. The datasets used in this paper are Breast cancer (which consists of 569 instances and 30 feature) from the UCI ML repository, Cardiotocography (which comprises 2126 and 21 features) from the UCI ML repository, Diabetes (which consists of 768 instances and eight features) from UCI ML repository, Kaggle Heart (which comprised of 270 cases and 13 features) from UCI ML repository, Musk1(which consists of 476 instances and 166 features) from UCI ML repository, and ASA-DI (which comprised of 822 cases and 48 features) from University Hospital of Angers (Table 1).

From the above literature review, we see that the feature selection method can be done by using traditional feature selection approaches alone, such as in [4, 15, 16], or by using the optimization technique to optimize the feature selection process, such as in [7–9, 18, 19]. Feature selection method can be done also by using a hybrid model between traditional methods, such as in [2, 5, 17], or making a hybrid model between optimization techniques themselves or between optimization techniques and traditional FS approaches such as in [1, 10, 11, 14].

3 Conclusions

Medical dataset suffers from the curse of dimensionality due to including redundant and irrelevant feature, so feature selection plays a vital role in solving this problem, and it chose most important feature subset. However, the traditional feature selection approaches increase the classification accuracy, but the hybrid model and optimization technique achieve the best classification accuracy.

Table 1 Presents a summary of the above-mentioned related work

References	Year	Method	Classifier	Dataset	Result
[14]	2017	IGWO-KELM	KELM classifier	Parkinson and Wisconsin diagnostic breast cancer	The best accuracy is 97.45 for Parkinson and 95.43 for WDBC
[7]	2019	Brainstorm optimization algorithm (BSO)	(SVM) where the BSO algorithm also tunes parameters	Hepatitis Liver Diabetes	Accuracy is Hepatitis: 97.16% Liver disorder: 84.31% Diabetes: 91.46%
[8]	2020	Improved teacher–learner-based optimization (ITLBO) algorithm	Support vector machine (SVM), convolution neural networks (CNNs), and gradient boosting	Chronic kidney disease (CKD) dataset	Accuracy SVM: 91.75% CNN:95.25% Gradient boosting:94.5%
[10]	2020	RS-CBGWO-FS	KNN	Cervical cancer, dermatology, diabetic retinopathy, arrhythmia, sonar	The highest accuracy achieved for each dataset is, cervical cancer: 97%, dermatology: 96%, diabetic retinopathy: 65%, arrhythmia:71%, sonar: 85%
[15]	2020	Using principal component analysis, Chi-squared testing, Relief-F, and symmetrical uncertainty. Then, several classification methods were used to develop models, which were then compared to find the best feature combinations	BayesNet, logistic, stochastic gradient descent (SGD), KNN (or in WEKA: IBK with K¼21), AdaBoost M1 with decision stump, AdaBoost M1 with logistic, repeated incremental pruning to produce error reduction (RIPPER or in WEKA: JRip), and random forest.29–36	Combination of four heart disease dataset (Cleveland Dataset, Long-Beach-VA Dataset, Hungarian Dataset, and Switzerland Dataset)	The best combination is Chi-squared feature selection with the BayesNet algorithm and achieved an accuracy of 85.00% Keywords

(continued)

Table 1 (continued)

References	Year	Method	Classifier	Dataset	Result
[2]	2020	$T = E(S(F))$, where $E \in$ FET and $S \in$ FST $S = \text{Filter}(F) \cup W(F)$ To accomplish the feature selection, two algorithms are used ((MFFSA) and (AFSA)) To accomplish the feature extraction (PCA)	RBF-based SVM	Heart disease Cleveland, Hungarian, Switzerland, and combined dataset	Cleveland:82.90%, Hungarian: 83.70%, and Switzerland: 91.30% combined dataset: 83.30%
[1]	2020	SIMPSONS	Support vector machine (SVM), Naive Bayes (NB), and AdaBoost (AB)	Colon, SRBCT, Leukemia, Prostate tumor, Lung cancer	SVM, Naive Bayes, AdaBoost Colon: 85.19 (3.21) 80.44 (2.56) 81.87 (1.12) SRBCT: 82.10 (0.25) 78.78 (1.04) 79.52 (1.65) Leukemia: 88.89 (2.08) 83.29 (1.88) 83.89 (2.42), Prostate tumor: 81.67 (0.25) 78.18 (2.83) 77.14 (0.55), Lung cancer: 88.19 (3.21) 88.44 (2.56) 88.87 (1.12)

(continued)

Table 1 (continued)

References	Year	Method	Classifier	Dataset	Result
[11]	2021	Ensemble-PSO-SVM	Harmonize classification of PSO-SVM	UCI breast cancer and lymphography datasets	UCI breast cancer: 96.15 and UCI lymphograph: 96.62
[16]	2021	R-HEFS	Naïve Bayes (NB), random forest (RF), support vector machine (SVM), and AdaBoost	Lung cancer, thyroid, Wisconsin diagnostic breast cancer (WDBC), Indian liver patient (ILP), Dermatology, Arrhythmia, SPECT heart, Hepatitis, SCADI, Hepatocellular carcinoma cancer (HCC)	The SVM accuracy Lung cancer: 90.66, Thyroid: 96.33, Wisconsin Diagnostic Breast Cancer (WDBC): 90.66, Indian Liver Patient (ILP): 58.06, Dermatology: 85.80, Arrhythmia: 74.60, SPECT heart: 83.55, Hepatitis: 85.80, SCADI: 86.18, Hepatocellular carcinoma cancer (HCC): 74.50
[5]	2021	PCA-IG model	Naïve Bayes	Breast cancer dataset	97.81%
[17]	2021	HRFLC, a combination of random forest (RF), AdaBoost (AD), and Pearson coefficient (PC)	Applied to different machine learning technique	Heart disease	79%

(continued)

Table 1 (continued)

References	Year	Method	Classifier	Dataset	Result
[9]	2021	MPA-KNN	KNN	18 dataset from UCI; an example is a lymphography	The best accuracy rate in 77.7% of the dataset
[4]	2022	MLPFS	MLPFS	SARS-CoV-2 RT-PCR dataset, ICU dataset	SARS-CoV-2 RT-PCR: 0.914 ICU dataset: 0.884
[18]	2022	Multi-objective quadratic binary HHO (MOQBHHO)	KNN	Arrhythmia	The heist accuracy using Q4 is 0.95
[19]	2022	GAwC: genetic algorithm with multi-objective compass	Extreme learning machine (ELM)	Breast cancer, cardiotocography, diabetes, heart, Musk1, ASA-DI datasets	The average accuracy for each dataset is: 97.48, 90.7, 79.75, 86.52, 82.2, 77.45

References

1. Rostami M, Forouzandeh S, Berahmand K, Soltani M (2020) Integration of multi-objective PSO based feature selection and node centrality for medical datasets. Genomics 112(6):4370–4384. https://doi.org/10.1016/j.ygeno.2020.07.027
2. Shah SMS, Shah FA, Hussain SA, Batool S (2020) Support vector machines-based heart disease diagnosis using feature subset, wrapping selection and extraction methods. Comput Electr Eng 84:106628. https://doi.org/10.1016/j.compeleceng.2020.106628
3. Sánchez-Maroño N, Alonso-Betanzos A, Tombilla-Sanromán M (2007) Filter methods for feature selection—a comparative study. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol 4881 LNCS, pp 178–187. https://doi.org/10.1007/978-3-540-77226-2_19
4. El-Attar NE, Sabbeh SF, Fasihuddin H, Awad WA (2022) An improved DeepNN with feature ranking for Covid-19 detection. Comput Mater Continua 71(2):2249–2269. https://doi.org/10.32604/cmc.2022.022673
5. Odhiambo Omuya E, Onyango Okeyo G, Waema Kimwele M (2021) Feature selection for classification using principal component analysis and information gain. Expert Syst Appl 174(November 2020). https://doi.org/10.1016/j.eswa.2021.114765.
6. Kitagawa S, Takenaka M, Fukuyama Y (2004) Recent optimization techniques. Rev Lit Arts Am 89–93
7. Tuba E, Strumberger I, Bezdan T, Bacanin N, Tuba M (2019) Classification and feature selection method for medical datasets by brain storm optimization algorithm and support vector machine. Procedia Comput Sci 162:307–315. https://doi.org/10.1016/j.procs.2019.11.289
8. Manonmani M, Balakrishnan S (2020) Feature selection using improved teaching learning based algorithm on chronic kidney disease dataset. Procedia Comput Sci 171(2019):1660–1669. https://doi.org/10.1016/j.procs.2020.04.178
9. Elminaam DSA, Nabil A, Ibraheem SA, Houssein EH (2021) An efficient marine predators algorithm for feature selection. IEEE Access 9:60136–60153. https://doi.org/10.1109/ACCESS.2021.3073261
10. Azar AT, Anter AM, Fouad KM (2020) Intelligent system for feature selection based on rough set and chaotic binary grey wolf optimisation. Int J Comput Appl Technol 63(1–2):4–24. https://doi.org/10.1504/IJCAT.2020.107901
11. Hamid TMTA, Sallehuddin R, Yunos ZM, Ali A (2021) Ensemble based filter feature selection with harmonize particle swarm optimization and support vector machine for optimal cancer classification. Mach Learn Appl 5(May):100054. https://doi.org/10.1016/j.mlwa.2021.100054
12. Pezoulas VC, Exarchos TP, Fotiadis DI (2020) Types and sources of medical and other related data
13. Sapra A, Bhandari P (2020) Vital sign assessment, no January, 2020, PMID : 31985994
14. Li Q et al (2017) An enhanced grey wolf optimization based machine for medical diagnosis. Comput Math Methods Med 2017:1–16
15. Spencer R, Thabtah F, Abdelhamid N, Thompson M (2020) Exploring feature selection and classification methods for predicting heart disease. Digit Health 6:1–10. https://doi.org/10.1177/2055207620914777
16. Bania RK, Halder A (2021) R-HEFS: rough set based heterogeneous ensemble feature selection method for medical data classification. Artif Intell Med 114(March).https://doi.org/10.1016/j.artmed.2021.102049
17. Pavithra V, Jayalakshmi V (2021) Hybrid feature selection technique for prediction of cardiovascular diseases. Mater Today Proc 81:336–340. https://doi.org/10.1016/j.matpr.2021.03.225

18. Piri J, Mohapatra P (2021) An analytical study of modified multi-objective Harris Hawk Optimizer towards medical data feature selection. Comput Biol Med 135(June):104558. https://doi.org/10.1016/j.compbiomed.2021.104558
19. Gutowski N, Schang D, Camp O, Abraham P (2022) A novel multi-objective medical feature selection compass method for binary classification. Artif Intell Med 127(March):102277. https://doi.org/10.1016/j.artmed.2022.102277