Received: 6 April 2022 | Revised: 12 May 2022 | Accepted: 16 May 2022 | Published online: 17 May 2022

Journal of Computational and Cognitive Engineering
2023, Vol. 2(1) 57–67
DOI: 10.47852/bonviewJCCE2202238
BON VIEW PUBLISHING

RESEARCH ARTICLE

Novel Approach to Evaluate Classification Algorithms and Feature Selection Filter Algorithms Using Medical Data

Fawad Masood[1,2], Junaid Masood[3,*], Hina Zahir[4], Kaouthar Driss[5], Nasir Mehmood[6] and Hassan Farooq[7]

[1]*School of Computing, Edinburgh Napier University, UK*

[2]*College of Information Engineering, Yangzhou University, China*

[3]*Department of Computer Science, IQRA National University, Pakistan*

[4]*Department of Electrical Engineering, University of Engineering and Technology, Pakistan*

[5]*National School of Computer Sciences, ENSI Manouba University Campus, Tunisia*

[6]*Department of Basic and Applied Sciences, Air University, Pakistan*

[7]*Department of Computer Science, Bahria University, Pakistan*

Abstract: In today's world, hepatitis is a widespread problem related to the medical field, which directly affects the lives of mankind. For patient survival, data mining is essential in predicting future trends using various techniques. This paper uses three feature selection filter algorithms (FSFAs): relief filter, step disc filter, and Fisher filter algorithm and 15 classifiers using a free data mining Tanagra software having UCI Machine Learning Repository. This process is done on a medical dataset with 20 attributes and 155 instances. As a result, the error rate is obtained in terms of accuracy, which shows the performance of algorithms regarding patient survival. This work also shows the independent comparison of FSFAs with classification algorithms using continuous values and the FSFA without using classification algorithms. This paper shows that the obtained result of the classification algorithm gives promising results in terms of error rate and accuracy.

Keywords: data mining, classification algorithm, Tanagra, feature selection algorithm, error rate, input parameters, target parameter

1. Introduction

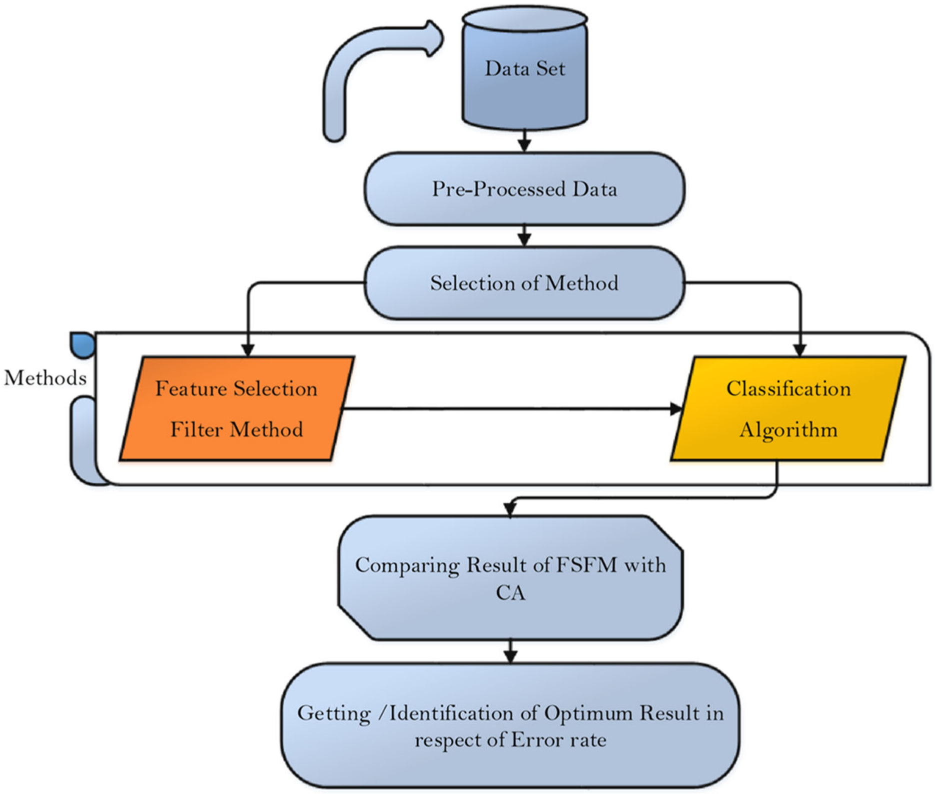
Different forms of algorithms exist in data mining in the field of computer science. Classification algorithms are also commonly used to deal with various methods and techniques in grouping the different subsets of entities. Using an algorithm that tests all subsets of potential functionality is the simplest way to reduce the error rate. Feature selection filter algorithms (FSFAs) is also known as variable selection algorithms. In this study, the medical dataset is selected from the UCI Machine Learning Repository, as shown in Figure 1. The dataset is then preprocessed to eliminate irrelevant and noisy data and then optimal data are chosen based on the study outcome. Finally, the dataset is analyzed in which two approaches have been suggested. The information is passed from the feature selection (FS) process to the classifier phase in one of two ways while another method is to transfer data from data preprocessing to a classifier algorithm and then compare both phases output FS and classifier algorithms in a subsequent step. The final stage is to determine the best classifier based in terms of error rate and accuracy. To get optimum results of classifier, error rate must be closer to zero. The growth of features has revealed a dimensionality issue for understanding algorithms. The impact and values of features vary depending on the methods and standard indication with criteria (Manikandan et al., 2017), which is a problem in classification.

In most cases, FS methods aim to reduce the number of input variables to improve results in the classification task. So, it is essential to filter out irrelevant attributes before applying any mining technique such as a classification algorithm, as shown in Figure 1.

The relevant attribute in the medical dataset has a lot of weight and significance for the classification task. Moreover, the redundant functions affect the algorithm's classification performance and add to the computational costs. The FS algorithm, which selects the best attribute, is essential for removing unnecessary features or attributes. Thus, FS is one of the most known techniques in the data mining area to optima. It can effectively reduce computation costs and eliminate overfitting and modeling errors, resulting in a better understanding of results (Liu et al., 2017). In this context, researchers have started to pay attention to the choice of the FSFA

*Corresponding author: Junaid Masood, Department of Computer Science, IQRA National University, Pakistan. Email: Junaidkttk@gmail.com

Figure 1
UCI diagram on feature selection


Data Set
Pre-Processed Data
Selection of Method
Methods
Feature Selection Filter Method
Classification Algorithm
Comparing Result of FSFM with CA
Getting /Identification of Optimum Result in respect of Error rate

method because it provides several benefits in terms of performance and accuracy.

(i) **Independent classifier**
- Fisher selection filter algorithm
- Relief selection filter algorithm
- Step disc filter algorithm.

(ii) **Dependent classifier**
- Wrapper method
- Embedded method (Liu et al., 2017).

In Figure 2, the first phase is known as the original feature set. Data will be added and processed. In the second step candidate subset is generated, where all irreverent and noisy data are eliminated and removed. During the third step known as the assessment phase, data will be evaluated (Visalakshi & Radha, 2017). After evaluation, the analyzed data transfer to the condition statement process which checks for the desired data result as requirements. If the data match the needs, it advances to the validation phase; otherwise, it returns to the state where the process began.

With the help of FS, the system can achieve:

- The efficiency of the classification algorithm has been improved.
- The scalability issue is addressed.
- Reduces classification complication due to a narrower set of interpreters (Visalakshi & Radha, 2017).

Many industries, such as health care, astronomy, social media, sensors, and others, now create massive amounts of data. It is necessary to remove useless, redundant, or unconnected data when working with such data. Dimensionality reduction (DR) is one of the several preprocessing procedures used to clean data. It aids classifiers by cutting down on training time and increasing classification accuracy. The most extensively used FS strategies in machine learning were studied in this work for enhancing classification and prediction accuracy (Karunakaran et al., 2021).

One possible explanation is that the datasets contain redundant, noisy, and uninformative features or attributes. As a result, FS methods are employed to identify the subset of relevant features capable of optimizing model performance. Furthermore, by reducing feature dimension, the model's training time and storage requirements can be reduced (Mandal et al., 2021).

FS is one of the most fundamental problems in machine learning, and it has received increased attention as a result of high-dimensional datasets emerging from fields such as bioinformatics. Filter methods are important for FS because they can be combined with any machine learning model and significantly reduce the run time of machine learning algorithms. The analyses' goal is to review how different filter methods work, compare their performance in terms of both run time and predictive accuracy, and provide application guidance (Bommert et al., 2020).

1.1. Introduction to liver

In human body, the liver is located on the upper right side of the stomach which is the largest internal organ. The liver is a vital organ that performs a variety of functions in the body, including protein synthesis, red blood cell decomposition, bile production, cholesterol, and glycogen synthesis, clotting factor (a protein found in the blood that helps to control bleeding) production of chemical liquid from the bile duct for digestion, etc.

The liver is located under the lungs and heart of the human body and above the large intestine, small intestine, gall bladder, and stomach.

1.2. Liver functions

- Removes potential toxic by-products of certain medications.
- Breaks down or metabolisms nutrients in food to generate energy when required.
- Helps the body in fighting infection by eliminating bacteria from the blood.
- Stores vitamins, minerals, and sugar to prevent food shortages.
- Produces most the body's protein requirements.
- Produces bile, a substance needed for fat digestion and absorption of vitamins A, D, and K.

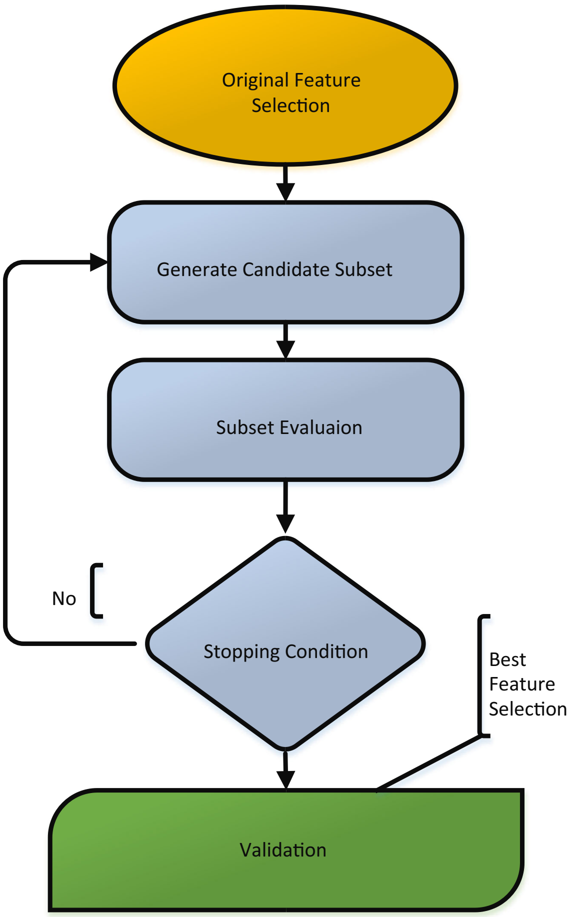
1.3. Liver diseases

- Many types of disease occur in the liver, that is:
- Hepatitis A, Hepatitis B, Hepatitis C
- Scarring
- Cancers
- Damage occurs by medications.

1.3.1. Hepatitis

Inflammation of the liver may cause hepatitis. Excessive alcohol intake and certain medications are two factors that cause hepatitis, one of which is infectious and the other not. A virus causes hepatitis. Hepatitis A, B, and C are the most common types of viral hepatitis, which cause a liver infection. Viral hepatitis is caused by a virus and can appear to be cold at first. But, unlike colds, chronic hepatitis C reveals how the hepatitis virus can endanger the patient's life due to liver failure and treatment difficulties. Symptoms of viral infection include fatigue, stomach upset, muscle aches, nausea, and loss of appetite in some of these patients. The signs of liver failure, on the other hand, only occur in advanced cases. Hepatitis B and C infection rates in children are high, ranging from 50 to 60%, compared to the adult population in Pakistan infected with hepatitis E, also known as epidemics.

Figure 2
Feature selection steps



1.3.2. Scarring

During an injury, the damaged skin is replaced by scar tissue, made up of fibrosis tissues. When the liver is damaged by self-medication or something else, fibrosis develops on the body's skin and the liver. Scarring tissues play an important role in determining whether the body will be cured or suffer more damage. Cirrhosis is when scarring tissue obstructs blood flow to the liver, restricting blood flow to the liver cells.

1.3.3. Cancer

Cancer is a group of virus infections characterized by abnormal cell growth that spreads to various human body parts. Different symptoms and signs occurred, such as blood clotting, long-term coughing, unexplained weight loss, hair loss, and a small number of cell divisions. These indications indicate cancer. There are over a hundred different forms of cancer that can affect the human body.

1.3.4. Damage occurs by medications

Everyone needs health care in today's world. When it comes to health issues, most people self-medicate rather than seeking advice from experts or physicians. As a result, self-medication can slowly harm one's life, such as certain medications that cause caused hepatitis, such as naproxen and ibuprofen. Medication-induced liver injury is often caused by other painkillers and the use of tablets to reduce fever.

Many other medications, such as birth control pills, statins, halothane, and others, can cause liver damage.

2. Literature Review

In this section, related work on FSFAs and classification by various researchers is discussed.

The author observed in Manikandan et al. (2017) that comprehensive research had been performed on FS, which has become increasingly profound, ranging from the most basic techniques to novel methods for obtaining unattended and semi-supervised designation from the supervised selection. This paper will show and help you how to choose the features you want to use.

Liu et al. (2017) suggested a new statistical measure called the (LW) label feature index as a possible replacement for the time-consuming cross-validation method when evaluating a subset of entities. As a result, a novel FS method has been developed by combining the suggested LW index with the sequence forward search algorithm. Furthermore, the author shows that the proposed technique can achieve classification accuracy comparable to encapsulation techniques using a centroid classifier or a support vector. Through various experiments performed on 9 UC Irvine Machine Learning Repository datasets, its measurement rate is approximate.

Visalakshi and Radha (2017) proposed to solve the problem of computational complexity; the authors proposed a filter-based algorithm. The article's authors claim that reducing the space of characteristics has a significant impact on water pollution. (i) The aim is to speed up the characteristics of the choosing production by applying pre-selection of the proposed feature based on the filter and ensuring that the valid data do not deviate from the initial step. (ii) The resulting functionalities were filtered once more using a genetic algorithm encoded in the support vector machine (SVM) technique, which aids in the synthesis of subsets with high correctness functions and lowers the price. Fact gathering is aided by using strategies that effectively eliminate irrelevant characteristics to improve classification accuracy.

Hui et al. (2017) proposed for detecting error or faultfinding method using rolling element bearing analysis, an improved (WFS) wrapper-based FS technique was previously combined with a SVM model classifier. The bearing vibration dataset provides a case study of Western Reserve University's bearing data center, which was maintained using the suggested (WFS) wrapper-based FS technique. The behavior was analyzed and discussed. The chosen (WFS) wrapper-based FS technique results considered the best feature subset with the least amount of statistical effort by removing reassessment data. As a result, the proposed approach is capable of performing FS tasks.

Hoseininejad et al. (2019) explained how to pick a subset of global resources for all applications in a space sample using standard FS methods. In localized feature selection (LFS), the sample space in various applications is linked to its optimum subset of the feature set. This method allows a function subset of resources to be passed to accommodate local distance in the sample space. Subsets of features are selected to keep those in a localized region and distances within the class to a minimum. In contrast, distances between types are kept to a maximum. LFS is used for solving using the fixed weight of regions using the randomized rounding technique.

Masoudi-Sobhanzadeh et al. (2019) highlight the FS approach in different science specialties such as software engineering, computer science field, medical field, and other engineering fields. For testing, various software and methods were implemented to find effective results, such as the WEKA program. These tools use a filter process, which yields lower results than wrapper methods. Researchers also developed function pick, which is an application program that deals with restrictions. Furthermore, the filter method function selects three types of learners for algorithm optimization. This tool has a user-friendly interface and can be used for any study. It can be used for sensitivity and performance. It can also be applied to any unbalanced or balanced dataset using various functions.

Akhiat et al. (2019) suggested ensemble selection based on FS, which is one of the selection methods. One function on the trained model was used from the selection process to generate models from the library. This shows how each model represents a function in the library. Furthermore, the ensemble construction provides a high-performing subset of features as well as a subset of models. Using eight benchmark datasets tested in the proposed approach, the results show that the ensemble selection approach is successful.

Qu et al. (2019) studied FS methods which are essential factors in the pattern recognition problems. The effect of FS quality on classifier output in terms of classification generalization and accuracy is clear. To reduce the size of the function, subset without sacrificing accuracy and increase algorithm performance. The study proposed an algorithm based on FS based on association rules (ARFS). Using this rule on a dataset for mine was able to find the most common two categories and attributes. It will combine the forward selection method with a sequential technique based on item set-2 confidence and a subset of feature criteria to evaluate the decision tree algorithm's efficiency. According to the results of the experiments, the researcher used five datasets from the UCI Machine Learning Repository to conduct three comparison experiments in this study, namely precision, learning algorithm, and run time. Therefore, ARFS is preferred in terms of size and precision. The FS algorithm based on association rules is more robust than the relief filter algorithm in run time measurement.

Driss et al. (2020) suggested a useful predictive model for the better diagnosis of diabetes. The proposed approach is based on an imputation process, and then the aim was to identify and replace missing values. Experimental results using a real dataset show that the proposed classification technique has higher accuracy after missing values are imputed.

Driss et al. (2020) presented a new approach, named MFAP, to detect frequent patterns in large networks using the inexact matching technique. The proposed method is based on three steps: JSON data processing, data visualization, and mining frequent patterns. The proposed algorithm was shown to be consistently more efficient than similar state-of-the-art algorithms in real-world networks.

Karunakaran et al. (2021) stated that many fields today generate massive amounts of data, including health care, astronomy, social media, sensors, and so on. When working with such data, it is necessary to remove irrelevant, redundant, or unrelated information. DR is one of the various preprocessing techniques used to clean data. It benefits classifiers by shortening training time and increasing classification accuracy. In this paper, the most commonly used FS techniques in machine learning were investigated in order to improve classification and prediction accuracy.

Mandal et al. (2021) stated that FS is regarded as an important step in data preprocessing in machine learning and data science. When we use raw data for classification or clustering, we sometimes find that the learning algorithms do not perform well. One theory is that the datasets have redundant, noisy, and uninformative characteristics or qualities. As a result, FS approaches are used to find the subset of relevant features that can improve model performance. The model's training time and storage requirements can also be decreased by lowering feature dimension. We present a three-stage wrapper-filter-based FS framework for medical report-based disease detection in this paper. In the first stage, four filter methods—Mutual Information, ReliefF, Chi Square, and Xvariance—were used to create an ensemble, and then each feature from the union set was evaluated by three classification algorithms—SVM, naive Bayes, and k-nearest neighbors (k-NNs)—and an average accuracy was calculated. To obtain a preliminary subset of optimal features, the features with the highest accuracy were chosen. Pearson correlation was used in the second stage to eliminate highly correlated features. XGBoost classification algorithm was used in these two stages to obtain the most contributing features, which in turn provided the best optimal subset. The obtained feature subset was then fed into a meta-heuristic algorithm called the whale optimization algorithm in the final stage to further reduce the feature set and achieve higher accuracy. The proposed FS framework was tested using four publicly available disease datasets from the UCI Machine Learning Repository: arrhythmia, leukemia, DLBCL, and prostate cancer. Our results show that the proposed method outperforms many state-of-the-art methods and can detect important features as well. Less features result in fewer medical tests for correct diagnosis, saving both time and money.

Bommert et al. (2020) stated that due to high-dimensional datasets arising from various domains such as bioinformatics, FS has become one of the most important difficulties in machine learning. Filter methods are crucial in FS because they may be coupled with any machine learning model and drastically lower the run time of machine learning algorithms. The analysis' goal is to examine how different filter algorithms work, assess their performance in terms of both run time and predicted accuracy, and offer application suggestions.

Zhang et al. (2019) stated that when working with a high-dimensional data collection, FS can considerably improve the effectiveness of a learning algorithm. The two most frequent ways are the filter method and the wrapper method. These methods, however, have drawbacks. The filter technique evaluates and selects characteristics via independent evaluation, which is more computationally efficient but less accurate than the wrapper method. The wrapper technique computes the assessment using a predefined classifier, which can provide high accuracy for specific classifiers but is computationally expensive. We offer a new FS approach called the big margin hybrid algorithm for FS in this paper (LMFS). To find a set of candidate feature subsets, we first use a new distance-based evaluation function, in which samples from the same class are ideally close together while samples from other classes are far apart, and a weighted bootstrapping search strategy. Then, from the candidate feature subsets, we use a specific classifier and cross-validation to select the final feature subset. Six vibrational spectroscopic datasets and three different classifiers were used to validate the LMFS method's performance: k-NNs, partial least squares discriminant analysis, and least squares SVM. According to the findings, LMFS can effectively overcome overfitting between the optimal feature subset and a given classifier. The LMFS method's features outperform the filter and wrapper methods in terms of classification performance and model interpretation. Furthermore, LMFS effectively overcomes the impact of classifier complexity on computational time, and distance-based classifiers were discovered to be more appropriate for selecting the final subset in LMFS.

Cherrington et al. (2019) stated that learning is the foundation of intelligence. The goal of machine learning is to automate methods for achieving goals, improving predictions, or encouraging informed behavior. FS is an important step in data analysis because it reduces dataset dimensionality by removing irrelevant and/or redundant attributes to simplify the learning process or improve the quality of the outcomes. This study examines various filter methods based on ranking procedures (Information Gain (IG), Chi-square (CHI), V-score, Fisher Score, mRMR, Va, and ReliefF) and identifies potential challenges. We focus on how threshold determination affects the results of various filter methods based on ranked scores. We demonstrate that this issue is critical, particularly in the age of big data, where users must deal with attributes in the tens of thousands with only a limited number of instances.

Venkatesh and Anuradha (2019) stated that being in the digital era, the data generated by various applications are increasing dramatically both row-wise and column-wise, creating a bottleneck for analytics and increasing the burden on machine learning algorithms that work for pattern recognition. This source of dimensionality can be addressed using reduction techniques. DR can be accomplished in two ways: FS and feature extraction (FE). This paper focuses on a survey of FS methods; we can conclude from this extensive survey that the majority of FS methods use static data. However, with the advent of IoT and web-based applications, data are generated dynamically and at a rapid rate, increasing the likelihood of noisy data and impeding algorithm performance. The scalability of the FS methods is jeopardized as the size of the dataset grows. As a result, existing DR algorithms do not address issues with dynamic data. Using FS methods not only reduces the data burden but also prevents the model from overfitting.

Remeseiro and Bolon-Canedo (2019) stated that a preprocessing technique that identifies the key features of a given problem is FS. It has traditionally been used to solve a variety of problems such as biological data processing, finance, and intrusion detection systems. FS, in particular, has been used successfully in medical applications, where it can not only reduce dimensionality but also help us understand the causes of a disease. We describe some fundamental concepts in medical applications and provide some background information on FS. We review the most recent FS methods developed for and applied in medical problems, including medical imaging, biomedical signal processing, and DNA microarray data analysis. To highlight the applicability of applying FS approaches to medical problems and to explain how these methods perform in real-world circumstances, a case study of two medical apps using actual patient data is used.

Deng et al. (2019) stated that in essence, big multimedia data are heterogeneous, meaning it may contain a mix of video, audio, text, and images. This is due to the recent prevalence of novel applications such as social media, video sharing, and location-based services, among others. Text classification techniques have been widely used to facilitate multimedia data processing in many multimedia applications, such as video/image tagging and multimedia recommendation. We provide a comprehensive review of FS techniques for text classification in this paper. We begin by discussing some popular document representation schemes and similarity measures used in the text classification. The most popular text classifiers are then examined, including the nearest neighbor (NN) method, SVM, neural networks, Nave Bayes (NB), and decision tree. Following that, we survey four FS models, namely the filter, wrapper, embedded, and hybrid, discussing the advantages and disadvantages of the most recent FS approaches. Finally, we conclude the paper by providing a brief overview of some interesting FS work that is not related to the four models.

Alirezanejad et al. (2020) stated that the technique of finding the best feature subset in any dataset is known as gene selection. Gene selection is important in high-dimensional datasets with a low number of samples and a high number of features, respectively. The main goals of gene selection are to improve accuracy, locate the smallest effective feature subset, and improve assessment performance. Xvariance versus Mutual Congestion and other heuristic approaches for gene selection were proposed in this work. Xvariance aims to categorize labels based on internal feature properties, whereas Mutual Congestion is based on frequency. Eight binary medical datasets were used to test the proposed approaches. The results show that Xvariance performs well with normal datasets, but Mutual Congestion significantly enhances the accuracy of high-dimensional datasets.

Solorio-Fernández et al. (2020) stated that because of its practical significance and use on a wide range of challenges in expert and intelligent systems where unlabeled data exist, unsupervised feature selection (UFS) has sparked a lot of interest in recent years. Because of their efficiency, scalability, and simplicity, UFS approaches based on the filter approach have gotten increased attention. However, there are no extensive studies in the literature for evaluating such UFS algorithms when they are applied to a wide variety of real-world data under the same conditions. To close this gap, we give a detailed empirical and systematic evaluation of the most popular and contemporary filter UFS algorithms, assessing their effectiveness in terms of clustering, classification, and run time in this study. Fifty datasets from the UCI Machine Learning Repository and 25 high-dimensional datasets from the ASU Feature Selection Repository were used in our research. The Friedman test and Holm post hoc procedure were used in the clustering and classification results to see if the outcomes achieved by the tested methods were statistically significant. We present some practical advice and insights for the application of the filter UFS approaches evaluated in our paper based on our tests.

3. Motivation

The manual and conventional methods of detecting hepatitis are slow and ineffective, resulting in losing precious lives of human being. Furthermore, hematology instruments used by specialists for manual hepatitis diagnosis and identification are costly and are not available in all hospitals. However, compared to automated detection and classification of hepatitis data, this task is prone to errors regarding time consuming, inefficient, and requires more resources.

For achieving the required goals and getting the results, the classification algorithms use medical dataset to find out the error rate along with accuracy. Considering both techniques we obtained results that are beneficial for the life of human beings in respect of error rates and accuracy. This study focuses on applying FE and reduction algorithms to find out unrelated data without disturbing or affecting the accuracy of the classifier that removes irrelevant and excessive entities from the raw data and classification using these features. Also, FS plays an important role, not because of dimension but also because of data complications with multiple disciplines. Also, do state the art of survey on Feature selection techniques and classified the data by using fifteen standard classifiers and three Feature selection filter methods by using medical dataset which comprised twenty attributes and one fifty-five instances.

Key Contributions:

- Identify irrelevant information and detect and classify results with a minimum error rate and correctness. If the values come closer to zero, it shows promising results in respect of error rate and accuracy.
- Get results, using combine classifier and FS algorithms and get results by the individual result of classification algorithm without FS algorithm.
- Comparing both situations and we can investigate which process is good to check the patient survival.

4. Effective Classification Algorithm and FS Using Tanagra Software

Tanagra software is an open-source project and research software in which users can access their source code and algorithm made by the user. Tanagra software is a free software for data mining purposes for research, project, and academic pursuit. Users can use this software in different ways like use of machine learning, data analyzing, and it is designed on a graphic user interface and synthetic data extraction. The version used in this research is Tanagra 1.4.50.

4.1. Methodology

Medical Data is utilized for fifteen classifiers to determine the error rate and accuracy and FSFAs along with classifiers to test the error rate again in the end, compared the results. Check the error rates to see which strategy is better for human life survivor from the two outcomes. This study aims to use FE and reduction algorithms to find unrelated data without interfering with or affecting the accuracy of the classifier. These features are used to eliminate irrelevant and excessive entities from the raw data and to classify them using these features. Often, due to data complexity in several disciplines. Besides, conduct a state-of-the-art survey on FS techniques and organize data using 15 general classifiers with various algorithms. Then, it will be applied to a medical dataset mentioned below in a Table 1 with 20 attributes and 155 observations.

4.2. Medical dataset obtained from UCI learning repository

4.3. Classification algorithm or classifiers

Classification in machine learning and statistics is a supervised learning approach in which the computer program learns from the data given to it and makes new observations via classes. It is used to categorize data into predefined class labels. There are many different types of classification algorithms that you may encounter in machine learning and specialized approaches to modeling that may be used for each.

4.4. Feature selection filter algorithm

To proceed further, some preprocessing of the data is needed after the missing values have been replaced. One of the preprocessing techniques is feature reduction. The important features needed to implement the classification algorithm are defined in this phase. The model's complexity is reduced, making it easier to understand, thanks to feature reduction. Furthermore, the attenuation of the variables to obtain is advantageous during

Table 1
Twenty attributes and 155 instances

S. No	Attributes	Values	Category
1.	Class	Die, Live	Discrete
2.	Age	10 20 30 40 50 60 70 80	Continues
3.	Sex	Female, Male	Discrete
4.	Steroid	Yes, No	Discrete
5.	Antivirals	Yes, No	Discrete
6.	Fatigue	Yes, No	Discrete
7.	Malaise	Yes, No	Discrete
8.	Anorexia	Yes, No	Discrete
9.	Liver Big	Yes, No	Discrete
10.	Liver Firm	Yes, No	Discrete
11.	Spleen Palpable	Yes, No	Discrete
12.	Spiders	Yes, No	Discrete
13.	Ascites	Yes, No	Discrete
14.	Varices	Yes, No	Discrete
15.	Bilirubin	0.39, 0.80, 1.20, 2.00, 3.00, 4.00	Continues
16.	Alka Phosphate	33, 80, 120, 160, 200, 250	Continues
17.	Sgot	13, 100, 200, 300, 400, 500	Continues
18.	Albumin	2.1, 3.0, 3.8, 4.5, 5.0, 6.0	Continues
19.	Protime	10, 20, 30, 40, 50, 60, 70, 80, 90	Continues
20.	Histology	Yes, No	Discrete

model implementation. In certain cases, variable selection makes it possible to increase model accuracy. Manual selection by a domain expert is unquestionably the best method.

4.4.1. Feature selection fisher filter algorithm

It is a supervised FS algorithm based on a filtering strategy, in which the selection is processed independently of the learning algorithm. This part prioritizes the input attributes based on their importance. The redundancy of the input attributes is not considered in this method.

4.4.2. Feature selection relief filter algorithm

This is a supervised algorithm which will not consider the redundancy of the input attributes. At least two attributes must be available, and the target attribute must be discrete. It was created with the intention of being used to solve binary classification issues with discrete or numerical features. Relief assigns each feature a feature score, which can subsequently be used to rank and select the highest-scoring features for FS. These scores could also be used as feature weights to aid downstream modeling. The discovery of feature value discrepancies between NN instance pairs is used to score relief features. The feature score drops if a feature value difference is discovered in a surrounding instance pair with the same class.

4.4.3. FS step disc filter algorithm

Discriminant is often synonymous with step disc. In TANAGRA, the FORWARD and BACKWARD methods are implemented. At each step of the FORWARD method, we identify the variable that truly leads to the discrimination between classes. If this variable makes a major contribution, we include it. In the BACKWARD method, we start with the entire model, including all descriptors. We try to figure out which variable is the least important. If eliminating this variable does not have a substantial impact on group prejudice, it is removed. When there are no more variables to delete, the process ends. To minimize the number of variables, we used the STEPDISC approach which allows us to select variables that are possibly interesting one by one. We use the procedure's FORWARD and BACKWARD techniques, with Fisher's statistic as a stop rule at the 3.1 level, to give our results greater credence.

5. Experiment Results and Discussions

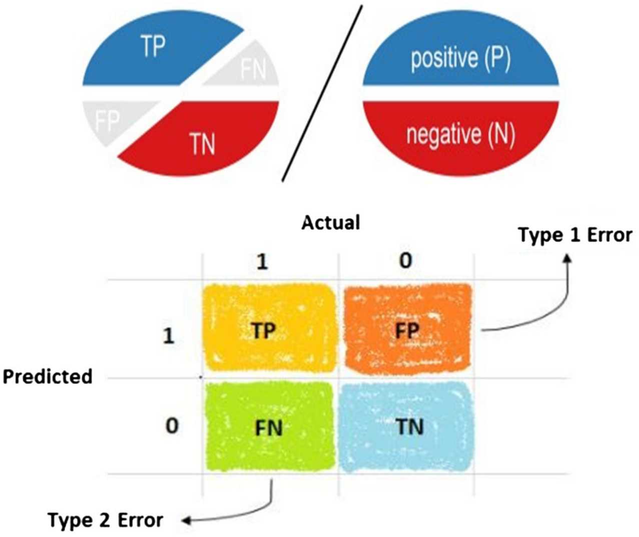
5.1. Accuracy in terms of evaluation matrices

The number of correctly predicted data points out of all data points is referred to as accuracy. It is defined as the number of true positives added with true negatives divided by true positives added true negatives added false positives added false negatives. The accuracy of a machine learning classification algorithm is one way to determine how regularly the algorithm correctly categorizes a data point. This is the formula of accuracy used in the machine learning field mentioned in Figure 3.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}).$$

With:
TP: true positive
TN: true negative
FP: false positive
FN: false negative

Figure 3
Accuracy in terms of evaluation matrices

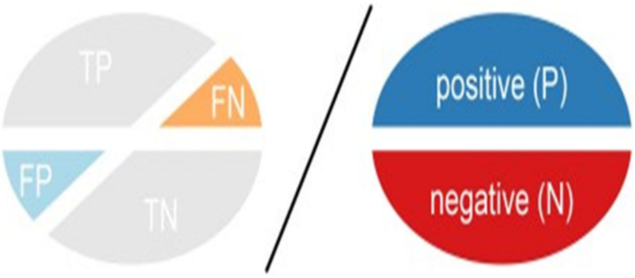


5.2. Error rate in terms of evaluation metrics

In this study, you can check the error rate in term of evaluation metrics showing in Figure 4 given below:

$$\text{Error rate (err)} \\ = \text{Number of incorrect predictions/total number of dataset}$$

Figure 4
Error rate in terms of evaluation metrics



5.3. Study of 15 classification algorithms

In this study, checked and examine the performance of 15 data mining classification algorithm that are given below in Table 2.

The first step in our approach we used to apply three common FSFAs that are:

- step disc filtering,
- relief filtering,
- -Fisher filtering

After using above-mentioned three feature selection filters algorithms along with classifiers, we apply 15 data mining classification algorithms or classifiers. Each algorithm will give

Table 2
Using 15 classification algorithms already built-in Tanagra software

S. No	Classification algorithm	S. no	Classification algorithm
1	BVM	9	K-NN
2	C4.5	10	LDA
3	C-RT	11	MLP
4	CS-CRT	12	NBC
5	CS-MC4	13	PLS-DA
6	C-SVC	14	PLS-LDA
7	Core Vector Machine	15	Random Tree
8	Iterative Dichotomiser 3		

error rate which shows the performance to analyze and measure classes. Besides, it shows which combined technique is best for prediction analysis. Table 3 gives detail information about using only classification algorithm or classifiers and using FSFA with classification algorithm.

5.4. Graphical view (without using FSFA)

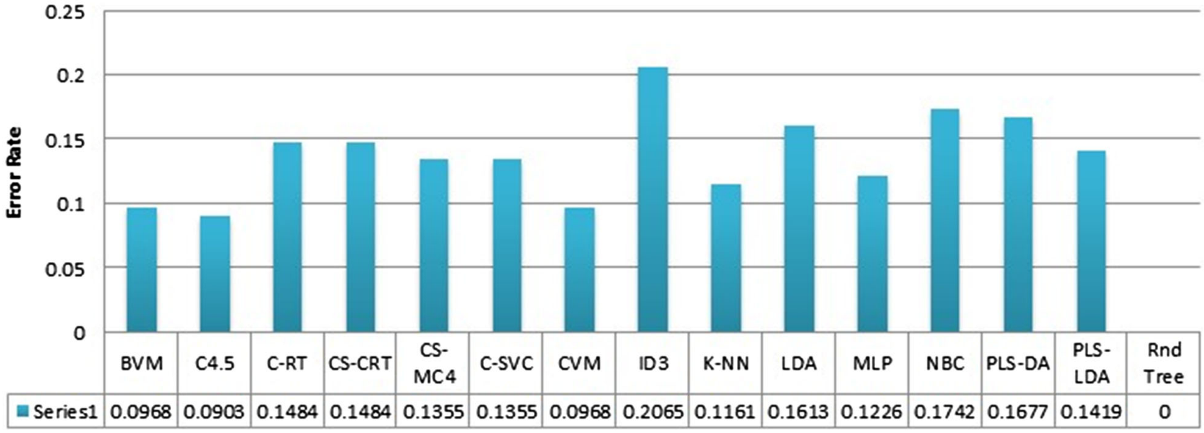
Figure 5 depicts the use of a classification algorithm without applying a FSFA. This figure presents the error rate of 15 classification algorithms, and zero indicates the lowest error rate, which is the best of all the outcomes. Horizontal lines depict the 15 classification algorithms, along with their names, while vertical lines depict the error rate. The error rate of 15 classification algorithms is

Table 3
Comparative study of classification algorithms used with and without feature selection filter method

		CA without using FSFA		Using FSFA along with CA					
				Err	Acc	Err	Acc	Err	Acc
S. no	Classifiers CA	Err	Acc	Step disc filter		Relief filter		Fisher filter	
1	BVM	0.0968	93 %	0.1419	87 %	0.1419	87 %	0.1419	87 %
2	C4.5	0.0903	94 %	0.1226	88.5 %	0.1226	88.5 %	0.1226	88.5 %
3	C-RT	0.1484	86 %	0.1484	86 %	0.1484	86 %	0.1484	86 %
4	CS-CRT	0.1484	86 %	0.1484	86 %	0.1484	86 %	0.1484	86 %
5	CS-MC4	0.1355	88 %	0.1355	88 %	0.1355	88 %	0.1355	88 %
6	C-SVC	0.1355	88 %	0.1484	79 %	0.1742	86 %	0.1484	86 %
7	CVM	0.0968	93 %	0.1419	87 %	0.1419	87 %	0.1419	87 %
8	ID3	0.2065	79 %	0.2065	79 %	0.2065	79 %	0.2065	79 %
9	K-NN	0.1161	89 %	0.1484	79 %	0.1290	88.4 %	0.1484	79 %
10	LDA	0.1613	84 %	0.1548	84.5	0.1871	82 %	0.1548	84.5
11	MLP	0.1226	88.5 %	0.1484	79 %	0.1742	83 %	0.1419	87 %
12	NBC	0.1742	83 %	0.1548	84.5	0.1871	82 %	0.1548	84.5
13	PLS-DA	0.1677	84 %	0.1484	86 %	0.1742	83 %	0.1484	86 %
14	PLS-LDA	0.1419	87 %	0.1484	86 %	0.1742	83 %	0.1484	86 %
15	Rnd Tree	0.0000	100 %	0.0194	95 %	0.0065	98 %	0.0194	95 %

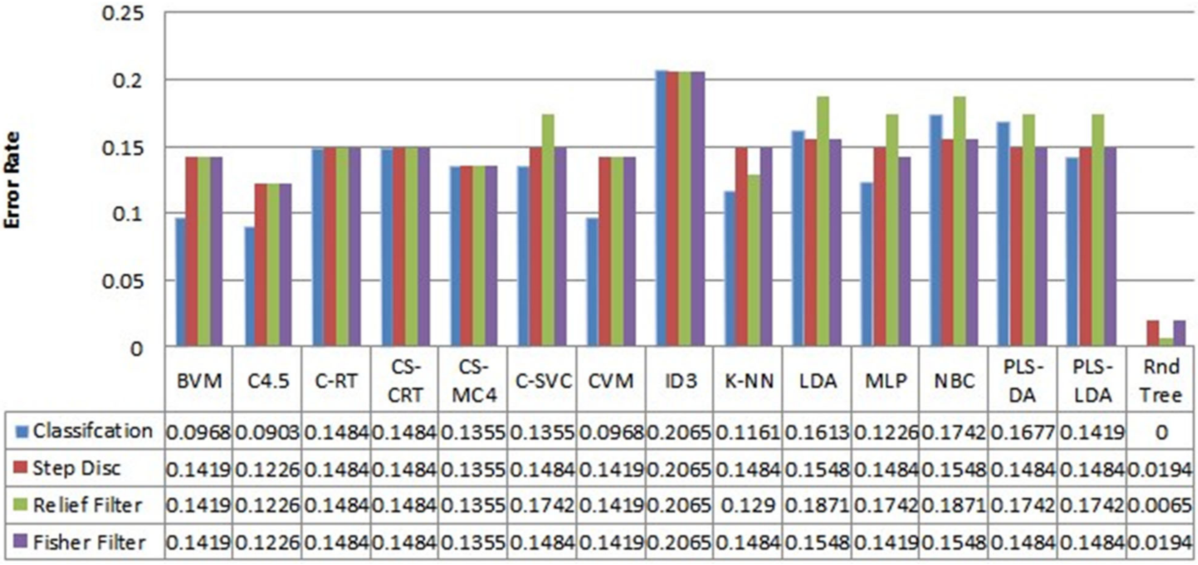
Where: CA = classification algorithm/classifier; Err = error rate; Acc = accuracy; FSFA = feature selection filter algorithm

Figure 5
Graphical representation of only using classification algorithm



	BVM	C4.5	C-RT	CS-CRT	CS-MC4	C-SVC	CVM	ID3	K-NN	LDA	MLP	NBC	PLS-DA	PLS-LDA	Rnd Tree
■ Series1	0.0968	0.0903	0.1484	0.1484	0.1355	0.1355	0.0968	0.2065	0.1161	0.1613	0.1226	0.1742	0.1677	0.1419	0

Figure 6
Representation of classifications with feature selection filter algorithms in bar form



	BVM	C4.5	C-RT	CS-CRT	CS-MC4	C-SVC	CVM	ID3	K-NN	LDA	MLP	NBC	PLS-DA	PLS-LDA	Rnd Tree
🟦 Classifcation	0.0968	0.0903	0.1484	0.1484	0.1355	0.1355	0.0968	0.2065	0.1161	0.1613	0.1226	0.1742	0.1677	0.1419	0
🟥 Step Disc	0.1419	0.1226	0.1484	0.1484	0.1355	0.1484	0.1419	0.2065	0.1484	0.1548	0.1484	0.1548	0.1484	0.1484	0.0194
🟩 Relief Filter	0.1419	0.1226	0.1484	0.1484	0.1355	0.1742	0.1419	0.2065	0.129	0.1871	0.1742	0.1871	0.1742	0.1742	0.0065
🟪 Fisher Filter	0.1419	0.1226	0.1484	0.1484	0.1355	0.1484	0.1419	0.2065	0.1484	0.1548	0.1419	0.1548	0.1484	0.1484	0.0194

discussed in detail in Figure 5 For example, BVM displays a result of 0.0968 below 0.1 on the vertical side, indicating the error rate. C4.5 displays 0.0903, C-RT and CS-CRT display point 0.1484 in graph, CS-MC4 and C-SVC display point 0.1355, CVM indicates point 0.0968, ID3 displays point 0.2065, K-NN displays point 0.1161 in graph, LDA displays point 0.1613, MLP displays 0.1226, NBC displays point 0.1742, PLS-DA displays 0.1677, PLS-LDA displays 0.1419, and Rnd tree displays.

5.5. Graphical view with using classification algorithm and FSFA

In Figure 6 below, we used three feature selection filter algorithms which are, respectively, step disc, relief, and Fisher filter FS algorithm along with classification algorithm.

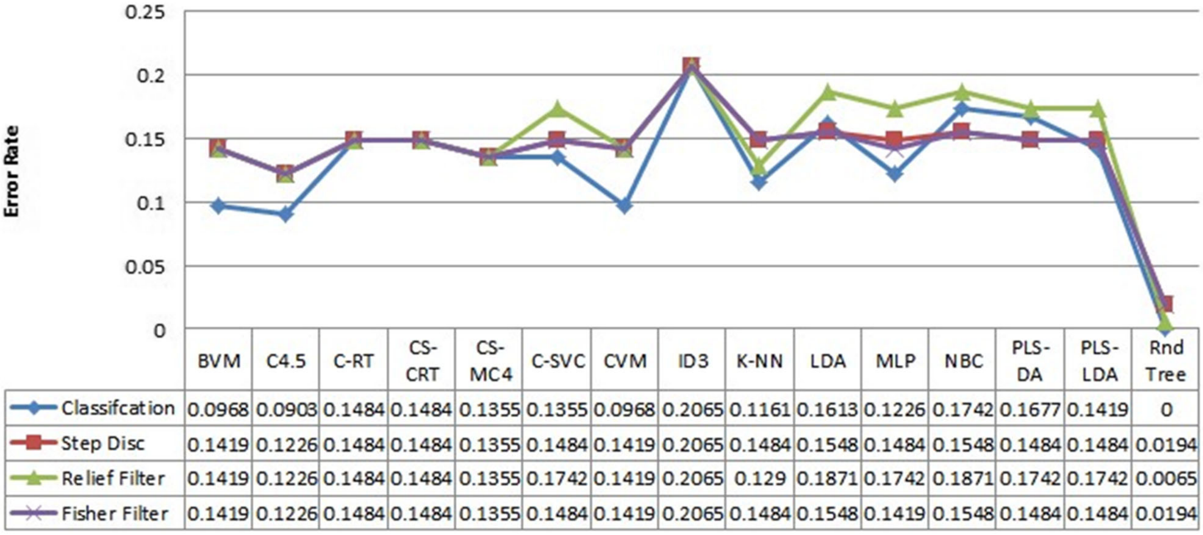
Below figure shows separately graphical representation of execution of data according to error rate of each combination (CA and FSFA).

In Figure 6, we can observe four vertical bar charts for each classification algorithm on horizontal line.

For instance, where the BMV algorithm having an error rate is 0.1419, C4.5 time is stopped on 0.1226, C-RT and CS-CRT time is stopped on 0.1484, CS-MC4 time is stopped on 0.1355, C-SVC time is stopped on.

Figure 7 shows classification algorithm along with FSFA. It shows that classification algorithm where having a blue line which is below then other lines that indicates minimum error rate than other filters methods (Fisher filter, step disc and relief filter) classification algorithm in form of linear line which starts from BVM where BVM point is stopped at 0.0968, C4.5 at

Figure 7
Classification and features selection filter algorithms showing in line bar



	BVM	C4.5	C-RT	CS-CRT	CS-MC4	C-SVC	CVM	ID3	K-NN	LDA	MLP	NBC	PLS-DA	PLS-LDA	Rnd Tree
◆ Classifcation	0.0968	0.0903	0.1484	0.1484	0.1355	0.1355	0.0968	0.2065	0.1161	0.1613	0.1226	0.1742	0.1677	0.1419	0
■ Step Disc	0.1419	0.1226	0.1484	0.1484	0.1355	0.1484	0.1419	0.2065	0.1484	0.1548	0.1484	0.1548	0.1484	0.1484	0.0194
▲ Relief Filter	0.1419	0.1226	0.1484	0.1484	0.1355	0.1742	0.1419	0.2065	0.129	0.1871	0.1742	0.1871	0.1742	0.1742	0.0065
✕ Fisher Filter	0.1419	0.1226	0.1484	0.1484	0.1355	0.1484	0.1419	0.2065	0.1484	0.1548	0.1419	0.1548	0.1484	0.1484	0.0194

0.0903, C-RT at 0.1484, CS-CRT at 0.1484, CS-MC4 at 0.1355, C-SVC at 0.1355, CVM at 0.0968, ID3 at0.2065, K-NN at 0.1161, LDP at 0.1613, MLP at 0.1226, NBC at 0.1742, PLSDA at 0.1677, PLS-LDA at 0.1419, and Rnd Tree at 0.0000.

6. Conclusion and Future Work

Nowadays, FSFAs combined with classification methods in the data mining area provide the promising results in terms of prediction and improve the performance of the classification task.

The main goal of this study is to propose an approach that predicts patient survival by the degree of accuracy and efficiency. Besides, we aim to demonstrate the importance of using classification and FS algorithms to obtain the best results in the shortest amount of time as it is a crucial factor in a patient's survival. After conducting experiments and analyzing results obtained in terms of error rate and accuracy, it was discovered that the classification algorithm produces better results without combining it with the FSFA. So, instead of using FSFAs, it is more efficient in terms of accuracy to apply a classification method (Driss et al., 2020).

A case study is presented that compares the performance and computational cost of several FS approaches. ReliefF and the Fisher filter are used as filtering methods. Filter methods have the advantage of being quick and simple to understand. The following are the characteristics of filter methods: Each feature is considered separately, and features that are redundant may be included. Some features with high discriminatory power as a group but low discriminatory power as individual features will be omitted. The filtering method is unrelated to the classification approach.

Furthermore, the classification algorithm is more efficient in retrieving results than the FSFA, and it has shown promising experimental results in terms of accuracy. However, the dimensional reduction is also necessary since the more significant the dimension, the more time it takes to complete. This study found that the approach used in this research work had the lowest error rate. The use of the classification algorithm is the best choice to reduce computational memory usage to eliminate irrelevant and redundant features. Furthermore, the experimental results show that many features play an essential role in improving classification accuracy.

The researcher may use more FSFAs to test results based on accuracy measures and error rate in upcoming research work to improve the process. This approach can also be applied to visualize features in images and analyze noisy network data (Driss et al., 2020). In the future, more FSFAs will be used to classify sensitive patients' information and increase the accuracy and reduce the error rate. Moreover, this scheme may be expanded in the future to include more effective combined classification algorithms. In future this research may be used to apply it on FE Feature Extraction that only considers dimensionality and also use the wrapper method for comparative analysis with fewer features is still an open research area. Research could be expended by employing other clustering techniques along with associated rules of mining for very large data regarding patients.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

References

Akhiat, Y., Chahhou, M., & Zinedine, A. (2019). Ensemble feature selection algorithm. *International Journal of Intelligent Systems and Applications*, *11*(1), 24. https://doi.org/10.5815/ijisa.2019.01.03.

Alirezanejad, M., Enayatifar, R., Motameni, H., & Nematzadeh, H. (2020). Heuristic filter feature selection methods for medical datasets. *Genomics*, *112*(2), 1173–1181. https://doi.org/10.1016/j.ygeno.2019.07.002.

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., & Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis*, *143*, 106839. https://doi.org/10.1016/j.csda.2019.106839.

Cherrington, M., Thabtah, F., Lu, J., & Xu, Q. (2019). Feature selection: filter methods performance challenges. In *2019 International Conference on Computer and Information Sciences*, 1–4.

Deng, X., Li, Y., Weng, J., & Zhang, J., (2019). Feature selection for text classification: A review. *Multimedia Tools and Applications*, *78*(3), 3797–3816. https://doi.org/10.1007/s11042-018-6083-5.

Driss, K., Boulila, W., Batool, A., & Ahmad, J. (2020). A novel approach for classifying diabetes' patients based on imputation and machine learning. In *2020 International Conference on UK-China Emerging Technologies*, 1–4.

Driss, K., Boulila, W., Leborgne, A., & Gançarski, P. (2020). Mining frequent approximate patterns in large networks. *International Journal of Imaging Systems and Technology*, *31*(3), 1265–1279. https://doi.org/10.1002/ima.22533

Hoseininejad, F. S., Forghani, Y., & Ehsani, O. (2019). A fast algorithm for local feature selection in data classification. *Expert Systems*, *36*(3), e12391. https://doi.org/10.1111/exsy.12391.

Hui, K. H., Ooi, C. S., Lim, M. H., Leong, M. S., & Al-Obaidi, S. M. (2017). An improved wrapper-based feature selection method for machinery fault diagnosis. *PloS One*, *12*(12), e0189143. https://doi.org/10.1371/journal.pone.0189143.

Karunakaran, V., Rajasekar, V., & Joseph, S. I. T. (2021). Exploring a filter and wrapper feature selection techniques in machine learning. In *Computational Vision and Bio-Inspired Computing*, 2020, 497–506.

Liu, C., Wang, W., Zhao, Q., Shen, X., & Konan, M. (2017). A new feature selection method based on a validity index of feature subset. *Pattern Recognition Letters*, *92*, 1–8. https://doi.org/10.1016/j.patrec.2017.03.018.

Mandal, M., Singh, P. K., Ijaz, M. F., Shafi, J., & Sarkar, R. (2021). A tri-stage wrapper-filter feature selection framework for disease classification. *Sensors*, *21*(16), 5571. https://doi.org/10.3390/s21165571.

Manikandan, G., Susi, E., & Abirami, S. (2017). Feature selection on high dimensional data using wrapper-based subset selection. In *2017 Second International Conference on Recent Trends and Challenges in Computational Models*, 320–325.

Masoudi-Sobhanzadeh, Y., Motieghader, H., & Masoudi-Nejad, A. (2019). FeatureSelect: A software for feature selection based on machine learning approaches. *BMC Bioinformatics*, *20*(1), 170. https://doi.org/10.1186/s12859-019-2754-0.

Qu, Y., Fang, Y., & Yan, F. (2019). Feature selection algorithm based on association rules. *Journal of Physics: Conference Series, 1168*(5), 052012.

Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. *Computers in*

Biology and Medicine, *112*, 103375. https://doi.org/10.1016/j.compbiomed.2019.103375.

Solorio-Fernández, S., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. (2020). A systematic evaluation of filter Unsupervised Feature Selection methods. *Expert Systems with Applications*, *162*, 113745. https://doi.org/10.1016/j.eswa.2020.113745.

Venkatesh, B., & Anuradha, J. (2019). A review of feature selection and its methods. *Cybernetics and Information Technologies*, *19*(1), 3–26. https://doi.org/10.2478/cait-2019-0001.

Visalakshi, S., & Radha, V. (2017). A hybrid filter and wrapper feature-selection approach for detecting contamination in drinking water management system. *Journal of Engineering Science and Technology*, *12*(7), 1819–1832.

Zhang, J., Xiong, Y., & Min, S. (2019). A new hybrid filter/wrapper algorithm for feature selection in classification. *Analytica Chimica Acta*, *1080*, 43–54. https://doi.org/10.1016/j.aca.2019.06.054.

How to Cite: Masood, F., Masood, J., Zahir, H., Driss, K., Mehmood, N., & Farooq, H. (2023). Novel Approach to Evaluate Classification Algorithms and Feature Selection Filter Algorithms Using Medical Data. *Journal of Computational and Cognitive Engineering* 2(1), 57–67, https://doi.org/10.47852/bonviewJCCE2202238