Research Article GU J Sci 37(2): 606-620 (2024) DOI: 10.35378/gujs.998964

Gazi University

Journal of Science

http://dergipark.gov.tr/gujs

An Empirical Evaluation of Feature Selection Stability and Classification Accuracy

Mustafa BUYUKKECECI[1, *], Mehmet Cudi OKUR[2]

[1]*Univerlist, İzmir, Türkiye*
[2]*Yasar University, Faculty of Engineering, Software Engineering Department, İzmir, Türkiye*

Highlights

- This paper examines the relationship between selection stability and classification accuracy.
- Experiments were conducted using five filter and two wrapper methods and twelve stability metrics.
- The correlation between selection stability and classification accuracy was measured using Pearson's *r*.

Article Info

Received: 22 Sep 2021
Accepted: 28 Aug 2023

Keywords

Feature selection
Selection stability
Classification accuracy
Filter methods
Wrapper methods

Abstract

The performance of inductive learners can be negatively affected by high-dimensional datasets. To address this issue, feature selection methods are used. Selecting relevant features and reducing data dimensions is essential for having accurate machine learning models. Stability is an important criterion in feature selection. Stable feature selection algorithms maintain their feature preferences even when small variations exist in the training set. Studies have emphasized the importance of stable feature selection, particularly in cases where the number of samples is small and the dimensionality is high. In this study, we evaluated the relationship between stability measures, as well as, feature selection stability and classification accuracy, using the Pearson's Correlation Coefficient (also known as Pearson's Product-Moment Correlation Coefficient or simply Pearson's *r*). We conducted an extensive series of experiments using five filter and two wrapper feature selection methods, three classifiers for subset and classification performance evaluation, and eight real-world datasets taken from two different data repositories. We measured the stability of feature selection methods using a total of twelve stability metrics. Based on the results of correlation analyses, we have found that there is a lack of substantial evidence supporting a linear relationship between feature selection stability and classification accuracy. However, a strong positive correlation has been observed among several stability metrics.

1. INTRODUCTION

Nowadays, it is quite common to encounter datasets with high dimensions. However, machine learning algorithms are not specifically designed to handle such datasets, and experience a decline in performance when confronted with them. To address this issue, dimensionality reduction techniques have been proposed. Feature selection is one of these techniques. The main objective of feature selection is to select relevant features without losing any useful information. Eliminating irrelevant[1] and redundant[2] features increases the accuracy and generalization capacity of machine learning algorithms, reduces computational costs, and helps to get simple models that are easy to interpret. All these reasons highlight the importance of feature selection and make it an integral part of the machine-learning process. Nonetheless, searching for relevant features requires time and adds an extra layer to the modelling task.

Stability is an important issue in feature selection which refers to the robustness of the selection algorithm against minor changes, i.e., perturbations, in training data. The feature preferences of a stable feature

[1] The term refers to non-informative features.
[2] The term refers to features providing similar or duplicate information as other features in the dataset.

*Corresponding author, e-mail: mustafa.buyukkececi@univerlist.com

selection algorithm should not be affected by changes made in the training set. Algorithms that exhibit fluctuations in feature preferences, i.e., unstable algorithms, can misguide users and erode their confidence in the algorithm. Therefore, stability is a token of the fitness of the feature selection algorithms. Stability metrics are used to quantify the stability of the selection algorithms. Studies on feature selection stability have mostly focused on proposing selection algorithms that are resistant to data perturbations [1], determining the sources of instability [2], and quantifying the stability [3].

The major objectives of this study were to test the relationship between stability metrics and feature selection stability and classification performance. To this end, we have conducted a literature review to identify the publications relevant to this research topic. Wang et al. [4] conducted experiments to evaluate the stability and model performance of various feature selection techniques in different scenarios. The study involved nine software metric datasets, seven filter-based feature selection techniques, four levels of dataset perturbation, and nine different numbers of selected features. Drotár and Smékal [5] assessed five commonly used feature selection techniques from two perspectives: stability and its impact on classification performance. The authors utilized a stability measure based on Hamming Distance and Matthews Correlation Coefficient (MCC) to assess the quality of the classification models. Han and Yu [6] presented a theoretical framework that explores the relationship between the stability and accuracy of feature selection using a formal bias-variance decomposition of feature selection error. Domingos [7], and Munson and Caruana [8] provided a comprehensive and structured analysis of the bias-variance trade-off in their studies. Turney [9] discussed the relationships between selection stability, accuracy, and bias. Alelyani et al. [10] conducted comprehensive experiments to demonstrate the relationship between data characteristics and the stability of the selection algorithms. Gulgezen et al. [11] introduced an entropy-based nonlinear correlation (Symmetrical Uncertainty) to measure the similarity of feature subsets and performed accuracy and stability measurements of the Minimum Redundancy Maximum Relevance (mRMR) algorithm. Chu et al. [12], Karabulut et al. [13], and Janecek et al. [14] demonstrated the effect of feature selection on the accuracy of classifiers with rigorous experimental setups.

The studies mentioned in the paragraph above have contributed to our understanding of the relationship between feature selection, feature selection stability, and predictive accuracy from different perspectives. The main contribution of this article is to investigate the relationship between these concepts statistically. By conducting experiments and analyzing the results, the paper demonstrates that there is a lack of substantial evidence supporting a linear relationship between feature selection stability and classification accuracy. This implies that just because a feature selection method is more stable does not necessarily guarantee improved classification performance. Additionally, the paper highlights an important observation of a strong positive correlation among several stability metrics. This finding suggests that different stability metrics tend to agree with each other, indicating that they capture similar aspects of feature selection stability. The results of the study can potentially lead to a reevaluation of current practices and provide insights for future research and development in the field of feature selection stability and classification.

The remaining part of this study has been organized as follows. The second section introduces the feature selection process. The next section formulates the problem of feature selection stability, explains how to measure stability, and briefly summarizes similarity-based and frequency-based stability measures. The fourth section describes the experimental framework and setup. The fifth section presents and evaluates the findings obtained from the experiments, and the final section concludes the paper with a summary and discussion.

2. FEATURE SELECTION

Datasets may contain redundant and irrelevant features. These features do not contribute to the machine learning task and negatively affect the analysis process. Feature selection aims to identify and choose pertinent features while preserving the essential characteristics of the data. This process is carried out to improve or maintain the accuracy of classification or the quality of clusters. However, selecting an optimal feature set is an NP-Hard problem since searching the whole feature space is computationally intractable [15]. Reducing the number of input variables by removing redundant and irrelevant features increases the accuracy and generalization capacity of machine learning algorithms, shortens training and utilization

times, facilitates model understanding, and defies the curse of dimensionality. Besides its advantages, selecting relevant features requires time and adds a layer of complexity to the modeling task.

Feature selection is classified into three categories: supervised, unsupervised, and semi-supervised. This article is specifically focused on supervised feature selection and stability. Therefore, unsupervised and semi-supervised feature selection methods are not discussed. However, readers who are interested in exploring all feature selection methods are directed to reference [16] for more comprehensive information. Supervised feature selection methods utilize labeled data and can be used for binary, multiclass, and multi-label classification and regression problems. For supervised feature selection, relevant features are the ones correlating with the class variable, i.e., the dependent variable. Supervised feature selection methods fall into five categories: filter, wrapper, embedded, hybrid, and ensemble.

1. **Filter methods:** These methods involve ranking all features based on an evaluation function that utilizes distance, information (entropy), accuracy (error rate), correlation, or consistency. They do not select features, the feature selection is done by the user, and do not use a classifier to assess the performance of the selected features. Filter methods are divided into two subcategories. **Univariate filter methods** evaluate features individually, which causes feature dependencies to be ignored. On the other hand, **multivariate filter methods** take the mutual relationship between the features into account [17].
2. **Wrappers:** Wrappers have three components: a search strategy, e.g., randomized search, a classifier that works as a black box, e.g., Naïve Bayes, and a stopping criterion, e.g., the maximum number of iterations. In general, wrapper methods search through the feature space and evaluate all possible feature subsets using an inductive learner until a stopping criterion is met [18]. They can search the feature space extensively and interact with the classifier to select the relevant features. On the other hand, they have high computational costs and require longer running times.
3. **Embedded methods:** Embedded methods select relevant features during the training phase of the learning algorithm. Thus, like wrappers, they perform classifier-dependent feature selection. However, in contrast to wrappers, they have a lower computational cost and running time since they do not call the classifier repeatedly. Embedded methods also capture feature dependencies [19, 20].
4. **Hybrid methods:** Hybrid methods combine different feature selection approaches. The hybrid method that is widely used is the combination of filter and wrapper methods. After a specific filter method ranks the features, the user generates a subset of features. Then, the selected features are given as input to the wrapper algorithm to generate the final feature subset [21]. Hybrid methods have better accuracy than filters and have better computational complexity than wrappers.
5. **Ensemble methods:** Ensemble methods are flexible and robust feature selection methods, based on the idea of repeating feature selection more than once to create a group of feature subsets and then aggregating them into a single feature subset. Ensemble feature selection can be applied in three different ways. **The data diversity method** is performed by sampling the original dataset and applying a single feature selection algorithm to each sample. **The functional diversity method** is performed by applying a set of different selection algorithms to the original dataset. Lastly, the **hybrid method** applies a set of different selection algorithms to different samples of the original dataset [22].

3. FEATURE SELECTION STABILITY

The quality of feature selection algorithms is assessed by the classification performance of the features they prefer and their stability. Any feature selection algorithm is called stable if it is not sensitive to slight changes in the training set and produces repetitive results, i.e., outputs. On the other hand, an unstable feature selection algorithm produces varying or inconsistent results, e.g., exhibits fluctuations in feature rankings, when faced with slight variations in the training data. Training sets can be perturbed by using a resampling technique, removing or adding samples to the training set, reordering samples and features, and adding noisy and discrete samples to the training set.

The stability of feature selection algorithms can be measured in two steps. In the first step, perturbations are applied to the training set, and after each perturbation, the output of the selection algorithm, e.g., a ranked feature vector, a feature weight vector, or an index of the features, is collected. In the second step, outputs are compared using either similarity-based or frequency-based stability measures. Similarity-based approaches use pairwise comparison, while frequency-based approaches use the frequency of occurrence of an attribute or set of attributes to evaluate stability [3]. Based on the output representation of the feature selection algorithm, similarity-based approaches are categorized into three classes [23, 24].

- **Stability by Rank:** The measures in this category evaluate the correlation, i.e., similarity ratio, between ranked feature vectors obtained after each perturbation. Spearman's Rank Correlation Coefficient (commonly referred to as Spearman's ρ), Kendal's Rank Correlation Coefficient (commonly referred to as Kendall's τ), Canberra Distance, and Weighted Canberra Distance are exemplars of ranked-based stability measures.
- **Stability by Weight:** The measures in this category evaluate the correlation between the weight vectors obtained after each perturbation. Pearson's Correlation Coefficient (PCC) is the only method to compute the association between feature weight vectors.
- **Stability by Index:** The measures in this category evaluate the correlation between vectors of feature indices or binary vectors. Most of the stability measures in this category assess the amount of overlap between the resulting feature subsets. Some examples of index-based stability measures are the Sørensen-Dice Coefficient, Kuncheva Index, Ochiai Index, Jaccard Index, and Hamming Distance.

One of the aims of this study is to compare various stability measures. Therefore, to fulfill this goal, a novel frequency-based measure, proposed by Nogueira [3], is incorporated into the empirical part. Nogueira's stability measure[3] is defined as follows:

$$NM = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d} s_f^2}{\frac{\bar{k}}{d}\left(1 - \frac{\bar{k}}{d}\right)} \tag{1}$$

where d is the number of features, s_f^2 is the sample variance of the selection of the f^{th} feature and $\bar{k}$ is the average number of selected features. It is important for readers to remember that when using frequency-based approaches, the selected subsets of features should be represented as a binary matrix.

4. EXPERIMENTAL SETUP

We carried out the empirical study in three phases. Initially, we assessed the classification accuracy[4] of three classifiers (Naïve Bayes, K-Nearest Neighbors, and Discriminant Analysis) trained on the entire feature set. In the second phase, we introduced perturbations at the instance level to the training set and examined their impact on the stability of seven feature selection methods. Subsequently, we retrained the classifiers using the selected feature sets. Finally, we evaluated the correlation between stability metrics and feature selection stability and classification performance, using Pearson's Correlation Coefficient. The structure of our empirical framework is depicted in Figure 1. All the code used in the experiments was implemented in MATLAB® 2021a, executed on a macOS 64-Bit operating system, and run on a computer with an 8-core Intel Core i9 CPU (3.6 GHz) and 64GB DDR4 RAM. The subsequent subsections provide a detailed description of the experimental framework.

[3] MATLAB® and Python™ scripts are available at http://www.cs.man.ac.uk/~gbrown/stability/

[4] Classification accuracy is determined by dividing the number of correct predictions by the total number of predictions. The outcome is then multiplied by 100 to represent the accuracy as a percentage.

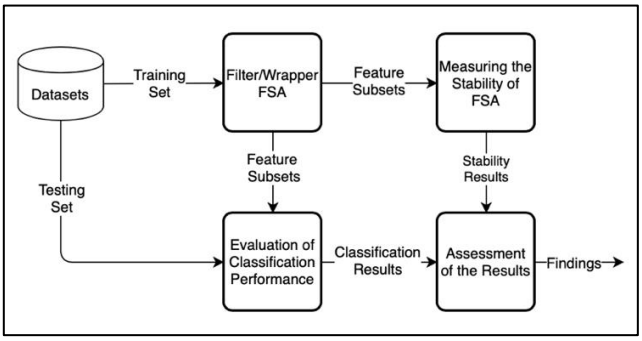

Figure 1. *The general framework of the empirical study*[5]

4.1. Datasets and Perturbation Methods

We conducted experiments on 8 publicly available real-world datasets taken from the UCI [25] and KEEL [26] machine learning repositories. Some characteristics of the datasets used in the empirical study and the names of the data repositories where datasets were collected are shown in Table 1. Missing values can be filled in by standard methods, such as the mean, median, or mode of the column, or even more sophisticated methods, such as shape-preserving piecewise cubic spline interpolation or cubic Hermite spline. In this study, missing values were filled by the moving (rolling) average.

Table 1. *Some characteristics of the datasets used in the empirical study and the name of the data repositories where datasets were collected*

Dataset	Instance	Attribute	Instance/Attribute Ratio	Number of Classes	Class Distribution		Reference
Australian Credit Approval	690	14	49.2857	2	383 307	55.51% 44.49%	UCI
Breast Cancer	569	32	17.781	2	212 357	37.26% 62.74%	UCI
Ionosphere	351	33	10.636	2	225 126	64.10% 35.90%	KEEL
Landsat Satellite	6435	36	178.750	7	1533 703 1358 626 707 0 1508	23.82% 10.92% 21.10% 9.73% 10.99% 0.00% 23.43%	UCI
QSAR Biodegradation	1055	41	25.731	2	356 699	33.74% 66.26%	UCI
SPECT Heart	267	44	6.068	2	212 55	79.40% 20.60%	KEEL
Sonar	208	60	3.466	2	97 111	46.63% 53.37%	KEEL
Vehicle	846	18	47	4	199 217 218 212	23.52% 25.65% 25.77% 25.06%	KEEL

Resampling techniques [27, 28], such as bootstrapping or *k*-fold cross-validation can be used as a perturbation method. During the experiments with filter methods, we applied a small amount of perturbation to the training sets using bootstrapping. We generated 10 bootstrap samples of the original data and then applied each filter method to these samples. Bootstrapping employs a technique called simple random

[5] FSA is the acronym for "feature selection algorithm".

sampling with replacement (SRSWR). Therefore, on average, each bootstrap sample consists of 63.2% of the original data, while the remaining 36.8% is left out to form the bootstrap test set.

Cross-validation is another resampling method used to evaluate the performance of supervised learning models. In *k*-fold cross-validation, the initial dataset is separated into *k* equal-sized[6] disjoint, i.e., mutually exclusive, subsets and the training and testing of the classifiers are repeated *k* times. In each iteration, one subset is used for testing and the others are used for training. During the experiments with wrapper methods, we applied a small amount of perturbation to the training sets using a 10-fold cross-validation procedure. Therefore, in each iteration, i.e., the experimental run, nine folds were used for training, and the remaining fold was used for testing.

4.2. Feature Selection Algorithms

The relevant features were selected using univariate parametric tests[7], such as the Two-Sample T-Test, Bhattacharyya Distance and Entropy, univariate nonparametric tests[7], such as the Wilcoxon Rank-Sum Test and ROC, Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS) wrappers [29-35]. SFS and SBS algorithms utilize a greedy search strategy, which means they make locally optimal decisions at each step without considering the global optimality of the solution. While this approach offers speed, simplicity, and ease of implementation, it does not guarantee to find the best possible solution. The quality of candidate feature subsets generated by the wrapper algorithms are evaluated using an induction algorithm. Therefore, we utilized Naïve Bayes, K-Nearest Neighbors, and Discriminant Analysis classifiers as feature subset evaluators and employed Bayes optimization, which is essentially a black-box optimization[8] method, to optimize the hyperparameters[9] of the classifiers.

4.3. Quantifying the Stability and Evaluation of Classification Performance

The stability performance of the filter algorithms was assessed using various measures, including Canberra and Weighted Canberra Distance, Spearman's Rank Correlation Coefficient, Kendall's Rank Correlation Coefficient, and Pearson's Correlation Coefficient. On the other hand, the stability of the wrapper algorithms was evaluated using the Hamming Distance, Jaccard Index, Cosine Index, Sorensen–Dice Coefficient, Lustgarten's Measure, Ward's Measure, and Nogueira's Measure. The stability metrics, excluding correlation-based measures and those by Lustgarten, Ward, and Nogueira, have values between 0 and 1, while others fall within the range of -1 to 1. Regrettably, due to limitations in available space, it is not possible to provide a comprehensive description of each stability measure in this context. However, readers who are interested in exploring detailed explanations of these metrics can refer to [3, 23, 24, 36] for more information.

5. RESULTS AND DISCUSSION

Table 2 compares the classification accuracies (in %) of the optimized Naïve Bayes, K-Nearest Neighbors, and Discriminant Analysis classifiers trained on the entire dataset (before employing any feature selection method to the datasets). The best and the worst accuracies were highlighted in bold and italic font, respectively. As we employed 10-fold cross-validation, the accuracy values correspond to the average cross-validation error. What can be seen in Table 2 is the dominant classification performance of the K-Nearest Neighbors classifier over the others on almost all datasets.

[6] If the dataset cannot be separated evenly, one subset can contain more samples than the other.

[7] Nonparametric (*or* distribution-free) tests are actually rank tests that use the count or ranking of the subjects on the dependent, i.e., response, variable.

[8] Black-box optimization is used in optimization problems where the structure of the objective function is unknown.

[9] Hyperparameter (*or* top-level parameter) is a user-defined variable, set before a learning algorithm is trained.

Table 2. *Classification accuracies (in %) of the classifiers trained on the entire dataset*

Dataset	**Classifier**[10]		
	NB	**K-NN**	**DA**
Australian Credit Approval	*80.0*	86.1	**87.1**
Breast Cancer	*94.4*	**97.2**	95.6
Ionosphere	90.9	**91.2**	*87.5*
Landsat Satellite	*82.3*	**90.8**	85.8
QSAR Biodegradation	*63.2*	**86.4**	85.6
SPECT Heart	*74.5*	**80.1**	79.4
Sonar	*76.4*	**87.0**	*76.4*
Vehicle	*62.2*	80.0	**85.1**

Each of the 10 bootstrap samples underwent the application of five different filter methods. The average stability performance of each filter method is listed in Table 3. The best and worst stability scores were highlighted in bold and italic font, respectively. The closer the stability score is to 1, the more stable the algorithm is. The top ⌊25%⌋ of the ranked feature vectors were used to compute the Weighted Canberra Distance. The results of the Canberra and Weighted Canberra Distances were divided by the total number of features and normalized to the interval [0,1]. Table 3 is quite revealing in several ways. First, a range of stability measures demonstrates the resilience of filter methods when faced with slight perturbations in the training data. The stability scores indicate that the variability in the rankings of features is minimal. Secondly, among the metrics utilized, the stability scores based on Kendal's Rank Correlation Coefficient are the lowest. Lastly, the stability scores based on Weighted Canberra Distance surpass other metrics in terms of being the highest. This means there is less variation in the upper positions of the ranked feature vectors.

Table 3. *Average stability performances of filter algorithms*

Dataset	**Filter Method**	**Stability Measure**[11,12]				
		CD	**WCD**	**PCC**	**SRCC**	**KRCC**
Australian Credit Approval	T-Test	0.936	**0.983**	**0.983**	0.943	*0.838*
	Entropy	0.905	0.960	**0.995**	0.903	*0.762*
	Bhatt.	0.945	**0.982**	0.979	0.948	*0.858*
	ROC	0.954	0.975	**0.976**	0.940	*0.841*
	Wilcoxon	0.915	**0.952**	0.948	0.926	*0.798*
Breast Cancer	T-Test	0.961	0.987	**0.992**	0.987	*0.926*
	Entropy	0.933	0.979	**0.980**	0.967	*0.869*
	Bhatt.	0.919	**0.995**	0.992	0.978	*0.895*
	ROC	0.955	0.989	**0.994**	0.989	*0.932*
	Wilcoxon	0.935	**0.989**	0.977	0.966	*0.868*
Ionosphere	T-Test	0.825	**0.976**	0.857	0.763	*0.584*
	Entropy	0.895	0.978	**1.000**	0.876	*0.729*
	Bhatt.	0.891	0.972	**1.000**	0.878	*0.720*
	ROC	0.790	**0.970**	0.731	0.705	*0.524*
	Wilcoxon	0.762	**0.964**	0.721	0.503	*0.353*
Landsat Satellite	T-Test	0.948	**0.990**	0.988	0.985	*0.914*
	Entropy	0.957	**0.995**	0.994	0.985	*0.917*
	Bhatt.	0.962	**0.994**	**0.994**	0.989	*0.932*

[10] Throughout this paper, the terms NB, K-NN, and DA refer to Naive Bayes, K-Nearest Neighbors, and Discriminant Analysis respectively.

[11] Throughout this paper, the terms CD, WCD, PCC, SRCC, and KRCC refer to Canberra Distance, Weighted Canberra Distance, Pearson's Correlation Coefficient, Spearman's Rank Correlation Coefficient, and Kendal's Rank Correlation Coefficient respectively.

[12] The correlation coefficients were calculated by employing the 95% confidence interval and underwent an averaging process in three steps. Initially, they were transformed into *z*-scores using the Fisher transform. Subsequently, the *z*-scores were averaged. Finally, the resulting means were converted back to correlation coefficients.

	ROC	0.924	**0.991**	0.971	0.961	*0.849*
	Wilcoxon	0.934	0.978	**0.994**	0.982	*0.905*
QSAR Biodegradation	T-Test	0.888	**0.992**	0.956	0.943	*0.810*
	Entropy	0.884	**0.998**	0.952	0.911	*0.777*
	Bhatt.	0.879	**0.998**	0.921	0.893	*0.758*
	ROC	0.923	**0.988**	0.974	0.946	*0.815*
	Wilcoxon	0.918	**0.992**	0.970	0.947	*0.829*
SPECT Heart	T-Test	0.849	**0.996**	0.888	0.877	*0.698*
	Entropy	0.817	**0.994**	0.802	0.824	*0.632*
	Bhatt.	0.818	**0.986**	0.828	0.827	*0.634*
	ROC	0.826	**0.992**	0.841	0.831	*0.637*
	Wilcoxon	0.764	**0.953**	0.660	0.640	*0.460*
Sonar	T-Test	0.798	**0.995**	0.784	0.757	*0.561*
	Entropy	0.797	**0.994**	0.721	0.780	*0.586*
	Bhatt.	0.770	**0.993**	0.675	0.725	*0.528*
	ROC	0.797	**0.993**	0.773	0.755	*0.558*
	Wilcoxon	0.810	**0.990**	0.790	0.680	*0.501*
Vehicle	T-Test	0.916	**0.988**	0.971	0.968	*0.874*
	Entropy	0.919	**0.982**	0.956	0.953	*0.850*
	Bhatt.	0.885	**0.968**	0.914	0.926	*0.799*
	ROC	0.861	0.922	**0.957**	0.914	*0.776*
	Wilcoxon	0.917	**0.986**	0.981	0.937	*0.824*

Filter methods calculate a relevance score for each feature and rank the features according to these scores. Feature subset generation is generally based on either a user-specified score threshold (features exceeding the threshold are selected) or selecting the top *n* features of the ranked list. In this study, for each dataset, we first averaged all relevance scores and sorted them in decreasing order. Next, we examined abrupt changes (sudden decreases) in the scores to establish a threshold. We also aimed to generate feature sets with the lowest possible cardinality. Table 4 compares the average classification accuracies (in %) of the classifiers on the selected features. The best and the worst accuracies were highlighted in bold and italic font, respectively. From Table 4, it can be concluded that K-NN, and Discriminant Analysis classifiers are better in building models with high accuracy. Furthermore, when the Ionosphere and SPECT Heart datasets were subjected to the Wilcoxon method for feature ranking, it was found that the classification performance of the first six features in Ionosphere and the first ten features in SPECT Heart outperformed the classification performance of the entire original feature set. Similarly, the top five features identified through the ROC and T-Test rankings for the Australian dataset exhibited equivalent classification performance to the entire feature set. Moreover, feature subsets derived from the Breast Cancer, Landsat Satellite, QSAR Biodegradation, and Sonar datasets yielded classification performances that were very close to those achieved by the entire feature set. The worst classification performances were observed in the Vehicle dataset.

Table 4. *Average classification accuracies of the classifiers on the selected features*

Dataset	**Number of Selected Features**	**Subset Evaluator**	**Accuracy of the Feature Subsets Ranked by…**				
			T-Test	**Entropy**	**Bhatt.**	**ROC**	**Wilcoxon**
Australian Credit Approval	5	NB	86.4	*77.8*	*77.8*	*81.7*	*82.0*
		K-NN	*85.2*	84.9	84.9	**87.1**	85.1
		DA	**86.5**	**85.7**	**85.7**	86.4	**85.7**
Breast Cancer	7	NB	*94.0*	*92.8*	*94.4*	*94.0*	*92.1*
		K-NN	**94.9**	93.7	95.1	94.9	**93.1**
		DA	*94.0*	**94.4**	**95.3**	**95.3**	**93.1**
Ionosphere	6	NB	*84.9*	**90.3**	**90.3**	*84.9*	*85.8*
		K-NN	**88.9**	89.7	89.7	**90.6**	**92.3**
		DA	86.3	*87.2*	*87.2*	87.2	86.9
Landsat Satellite	10	NB	*80.7*	*81.8*	*81.5*	*82.2*	*73.5*

		K-NN	**86.6**	**89.8**	**89.0**	**89.2**	**83.3**
		DA	84.3	86.4	85.4	86.4	79.9
QSAR Biodegradation	7	NB	*74.1*	*55.1*	*75.7*	*77.8*	*74.1*
		K-NN	**81.3**	**66.3**	**79.7**	**84.4**	**81.3**
		DA	75.6	**66.3**	76.4	79.2	75.6
SPECT Heart	10	NB	*73.0*	*74.9*	*75.7*	*73.0*	*74.2*
		K-NN	79.8	**80.9**	**80.5**	79.8	**81.6**
		DA	**80.1**	80.1	79.8	**80.1**	80.9
Sonar	10	NB	76.0	76.4	72.1	76.9	*78.4*
		K-NN	**78.4**	**77.4**	**78.4**	**77.4**	**82.2**
		DA	*72.1*	*71.2*	*71.6*	*74.5*	79.3
Vehicle	6	NB	*53.3*	59.1	61.2	*60.3*	63.7
		K-NN	**57.9**	**70.0**	**68.6**	**72.9**	**70.8**
		DA	57.4	*59.0*	*59.9*	67.4	*63.2*

The average classification and stability performances of SFS and SBS wrappers are listed in Table 5. The best and the worst classification accuracies and stability scores were highlighted in bold and italic font, respectively. According to the observations presented in Table 5, the feature subsets generated by the SFS and SBS wrapper methods generally exhibit higher classification accuracies compared to the accuracy achieved using the entire feature set. For instance, in the case of the Australian dataset, the feature subsets obtained through the SBS method, with the Discriminant Analysis classifier as the subset evaluator, demonstrated an average accuracy rate of 87.7%, which was 87.1% before feature selection. Similarly, for the Sonar dataset, the feature sets generated by the SBS method using the K-NN classifier as the subset evaluator achieved an average accuracy of 94.6%. This corresponds to a 7.6% increase in accuracy following feature selection.

Table 5. *Average classification accuracies and stability performances of SFS and SBS methods*

Dataset	**Search Direction**	**Subset Evaluator**	**Avg. Accuracy**	**Stability Measure[13]**						
				HD	**JI**	**CI**	**SDC**	**LM**	**WM**	**NM**
Australian Credit Approval	Forward	NB	**86.3**	*0.803*	*0.478*	*0.629*	*0.600*	*0.481*	*0.717*	*0.509*
		K-NN	85.9	0.946	0.807	0.892	0.880	0.759	**1**	0.848
		DA	*85.7*	**1**	**1**	**1**	**1**	**0.857**	**1**	**1**
	Backward	NB	87.6	**0.811**	0.751	**0.862**	**0.854**	**0.520**	**0.885**	**0.536**
		K-NN	*87.4*	0.779	**0.752**	0.857	**0.854**	0.419	0.597	0.242
		DA	**87.7**	*0.759*	*0.737*	*0.848*	*0.846*	*0.294*	*0.419*	*0.073*
Breast Cancer	Forward	NB	*97.0*	0.952	0.697	0.803	0.798	0.746	0.864	0.784
		K-NN	97.4	*0.780*	*0.281*	*0.418*	*0.415*	*0.253*	*0.322*	*0.303*
		DA	**97.5**	**1**	**1**	**1**	**1**	**0.833**	**1**	**1**
	Backward	NB	*96.4*	**0.701**	*0.573*	*0.741*	*0.721*	**0.348**	**0.734**	**0.391**
		K-NN	**97.9**	*0.682*	0.643	0.785	0.778	*0.181*	0.299	*0.124*
		DA	97.0	0.689	**0.652**	**0.789**	**0.785**	0.194	*0.287*	0.133
Ionosphere	Forward	NB	**94.6**	*0.890*	0.570	0.716	0.713	*0.561*	*0.710*	0.655
		K-NN	94.4	0.923	*0.568*	*0.697*	*0.692*	0.619	0.711	*0.649*
		DA	*89.2*	**0.969**	**0.769**	**0.867**	**0.863**	**0.822**	**0.930**	**0.845**
	Backward	NB	**93.7**	0.687	0.599	0.754	0.742	**0.324**	**0.603**	0.309
		K-NN	93.2	*0.630*	*0.545*	*0.707*	*0.702*	*0.157*	*0.287*	*0.186*
		DA	*90.2*	**0.754**	**0.654**	**0.790**	**0.788**	0.321	0.586	**0.487**
Landsat Satellite	Forward	NB	*83.6*	**0.867**	**0.577**	**0.696**	**0.692**	**0.499**	**0.664**	**0.628**
		K-NN	**90.6**	*0.629*	*0.422*	*0.590*	*0.588*	*0.150*	*0.292*	*0.255*
		DA	87.3	0.781	0.466	0.625	0.624	0.353	0.510	0.478
	Backward	NB	*83.4*	0.777	0.748	0.857	0.854	**0.368**	**0.522**	**0.316**

[13] Throughout this paper, the terms HD, JI, CI, SDC, LM, WM, and NM refer to Hamming Distance, Jaccard and Cosine Index, Sorensen–Dice Coefficient, Lustgarten's, Ward's and Nogueira's metrics respectively.

		K-NN	**91.4**	**0.812**	**0.805**	**0.892**	**0.891**	0.188	*0.230*	*0.031*
		DA	87.0	*0.667*	*0.608*	*0.758*	*0.754*	*0.176*	0.294	0.211
QSAR Biodegradation	Forward	NB	85.7	0.736	0.340	0.498	0.493	0.268	0.380	0.329
		K-NN	**87.5**	**0.810**	**0.473**	**0.633**	**0.628**	**0.416**	**0.577**	**0.507**
		DA	*84.3*	*0.650*	*0.315*	*0.467*	*0.464*	*0.155*	*0.239*	*0.214*
	Backward	NB	*84.6*	*0.701*	*0.648*	*0.788*	*0.783*	*0.260*	0.420	0.258
		K-NN	**87.9**	**0.762**	**0.720**	**0.836**	**0.835**	**0.333**	**0.479**	**0.372**
		DA	86.7	0.739	0.711	0.832	0.829	0.261	*0.378*	*0.213*
SPECT Heart	Forward	NB	82.0	0.913	0.179	0.279	0.264	0.302	0.327	0.235
		K-NN	**82.4**	*0.846*	*0.089*	*0.154*	*0.145*	*0.097*	*0.109*	*0.090*
		DA	*79.4*	**1**	**1**	**1**	**1**	**0.977**	**1**	**1**
	Backward	NB	**85.5**	0.736	*0.613*	*0.758*	*0.754*	**0.300**	**0.588**	**0.466**
		K-NN	*82.3*	**0.805**	**0.800**	**0.888**	**0.887**	*0.127*	*0.152*	*- 0.086*
		DA	83.4	*0.693*	0.650	0.790	0.786	0.212	0.334	0.200
Sonar	Forward	NB	*81.4*	**0.940**	**0.428**	**0.560**	**0.555**	**0.540**	**0.584**	**0.539**
		K-NN	**85.6**	*0.826*	0.210	0.349	0.326	0.315	0.372	0.241
		DA	82.9	0.837	*0.178*	*0.310*	*0.291*	*0.287*	*0.334*	*0.209*
	Backward	NB	*82.3*	*0.661*	*0.591*	*0.748*	*0.740*	**0.217**	**0.391**	**0.229**
		K-NN	**94.6**	0.689	0.650	0.789	0.783	0.209	0.324	0.160
		DA	89.2	**0.692**	**0.665**	**0.798**	**0.797**	*0.143*	*0.197*	*0.109*
Vehicle	Forward	NB	*63.8*	*0.711*	*0.481*	*0.637*	*0.626*	*0.283*	*0.526*	0.408
		K-NN	77.5	0.828	0.688	0.802	0.792	0.452	**0.824**	**0.650**
		DA	**86.2**	**0.841**	**0.824**	**0.901**	**0.899**	**0.553**	0.702	*0.319*
	Backward	NB	*64.7*	*0.662*	*0.608*	*0.757*	*0.751*	*0.199*	*0.344*	0.143
		K-NN	82.0	0.796	0.778	0.877	0.872	0.395	0.558	*0.129*
		DA	**86.6**	**0.869**	**0.857**	**0.924**	**0.922**	**0.580**	**0.730**	**0.309**

We used Pearson's Correlation Coefficient (see Formula 2) with a confidence interval of 95% to statistically measure the relationship between feature selection stability and classification accuracy. Pearson's Correlation Coefficient, generally abbreviated as r, measures the linear association between two continuous random variables, often referred to as zero-order correlation[14] and is defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2}$$

where, n is the sample size, x_i and y_i are the values of the x and y variables indexed with i, and $\bar{x}$ and $\bar{y}$ are the means of the values of the variables x and y respectively.

In this study, the correlation results were visualized using correlation heat maps, in which correlation coefficients are represented as colors. The heat maps below use a color palette with shades of blue. The correlation coefficients are indicated by the lightness and darkness of the shade, so the darker the color, the stronger the relationship between the two variables. The correlation heat maps are constructed using the stability scores and accuracies presented in Tables 3, 4, and 5. Before evaluating the results, it is important to keep in mind that correlation does not imply causation.

The correlation heat maps of filter methods are shown in Figure 2. Each correlation heat map shows the relationship between stability metrics and classification accuracy. The first column of each correlation heat map exhibits the correlation between stabilities and classification accuracy, while the remaining columns represent the correlation between the stability metrics. Upon examining the results, it was observed that only a few methods, such as ROC and Wilcoxon, displayed a moderately positive, i.e., weak, relationship

[14] Zero-degree correlation is the correlation between two variables without considering the effect of other variables.

between feature selection stability and classification performance. Additionally, we also observed that these associations vary depending on the classifier employed.

The correlation heat maps of wrapper methods are shown in Figure 3. For SFS and SBS methods, the correlations between the stability scores and accuracies are predominantly weak negative. This suggests that variables tend to move in opposite directions from one another, and the relationship between them is not strong. However, in certain cases, e.g., the SFS method using NB and DA as subset evaluators, a moderate positive correlation is observed between stability and classification accuracies. This indicates that variables tend to move in tandem but the relationship between them is not strong. The correlation coefficients that are close to zero indicate a negligible correlation. Hereby, considering the results of correlation analyses, there is no strong evidence of a linear relationship between feature selection stability and classification accuracy. Nevertheless, a strong positive correlation has been found between several stability metrics, such as Canberra Distance, Spearman's Rank Correlation Coefficient, Kendall's Rank Correlation Coefficient, and Pearson's Correlation Coefficient.

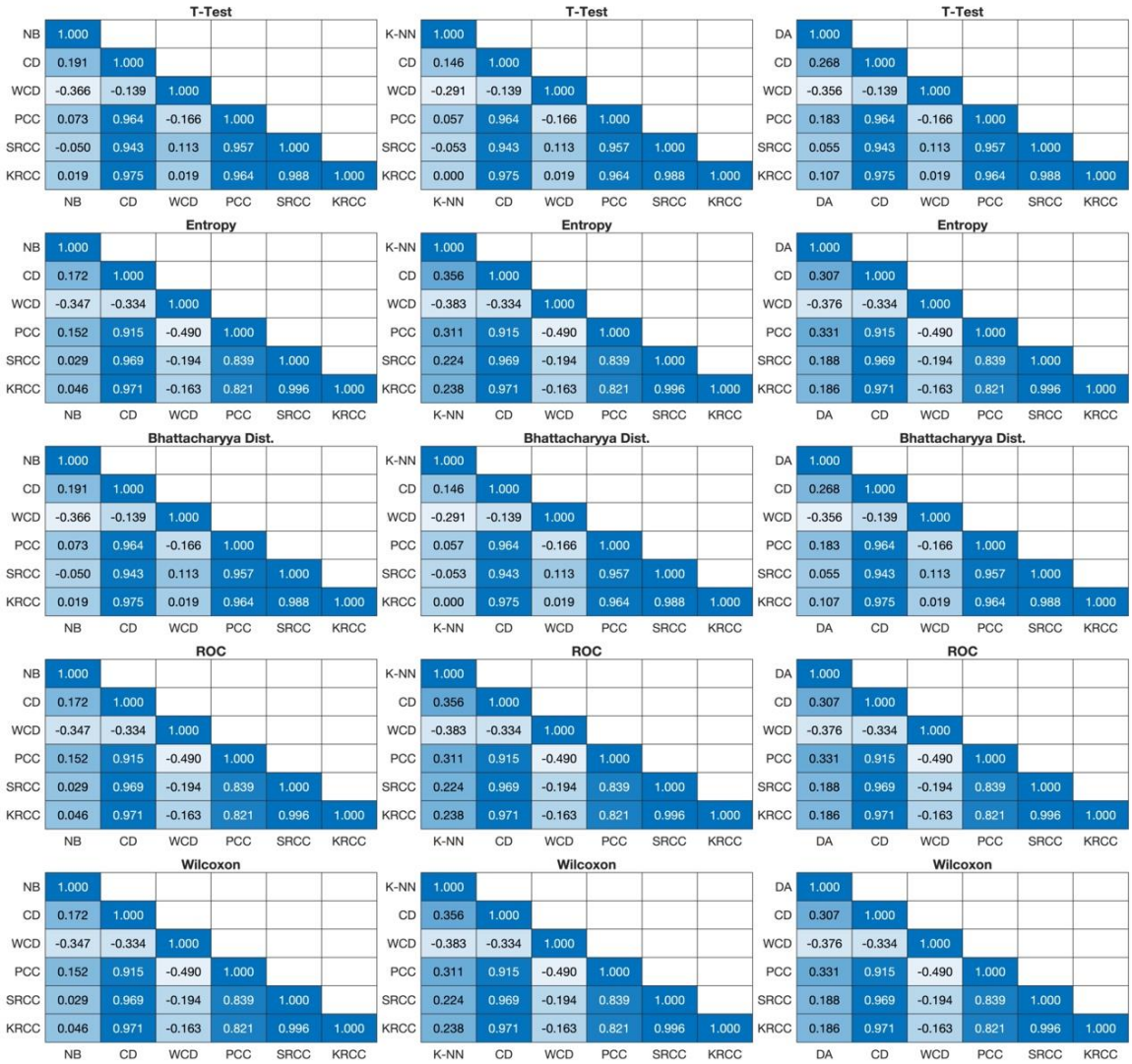

Figure 2. *Correlation heat maps of the filter methods*

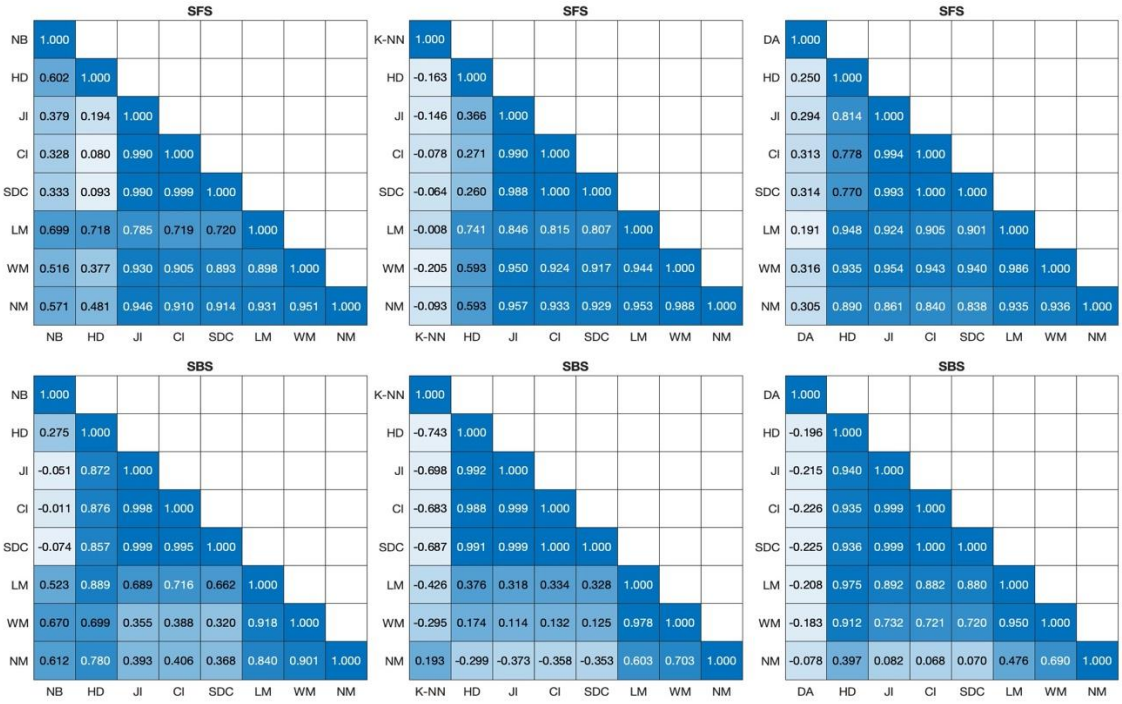

Figure 3. *Correlation heat maps of the SFS (top) and SBS (bottom) methods*

6. CONCLUSION

Supervised feature selection algorithms are evaluated according to their stability and the effect of feature preferences on prediction accuracy. In recent years, there has been a growing interest in these issues. For example, various empirical studies, such as those conducted by Wang et al. [4], Gulgezen et al. [11], González et al. [37], Wang et al. [38], and Deraeve and Alexander [39], have proposed novel selection methods that prioritize both stability and accuracy. This research study aimed to investigate the relationship between stability metrics, feature selection stability, and classification performance statistically. Krizek et al. [40] emphasized that the classification performance of the selected features and feature selection stability are two different concepts. The results obtained in this study mostly support the authors' assertions. However, our findings do not definitively establish the presence or absence of a relationship between these variables. In some instances, there is a moderately positive relationship between the stability and classification performance of the selected feature subsets. Furthermore, the majority of correlation coefficients fail to provide substantial evidence for the hypothesis that an algorithm with high selection stability produces subsets of features with high accuracy. On the other hand, the analysis revealed a strong positive correlation among several stability metrics. As is known, correlation does not imply causation. It means that just because two variables are correlated, it does not necessarily mean that one variable directly causes the other to occur. The correlation only measures the degree of association or relationship between variables, but it does not provide evidence of a cause-and-effect relationship. In other words, even if two variables show a strong correlation, it is possible that their relationship is coincidental or influenced by other factors. Additional research and evidence are required to establish a causal relationship between variables. Therefore, we will extend the empirical work to establish the cause-effect relationship and to gain further insights into the subject.

CONFLICTS OF INTEREST

No conflict of interest was declared by the authors.

REFERENCES

[1] Loscalzo, S., Yu, L., Ding, C., "Consensus group based stable feature selection", Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 567-576, Paris, France, (2009).

[2] Kalousis, A., Prados, J., Hilario, M., "Stability of feature selection algorithms: a study on high-dimensional spaces", Knowledge and Information Systems 12: 95-116, (2007).

[3] Nogueira, S., "Quantifying the stability of feature selection", Ph.D. Thesis, University of Manchester, Manchester, United Kingdom, 21-67, (2018).

[4] Wang, H., Khoshgoftaaar, T.M., Liang, Q., "Stability and classification performance of feature selection techniques", 2011 10th International Conference on Machine Learning and Applications and Workshops, 151-156, Honolulu, HI, USA, (2011).

[5] Drotár, P., Smékal, Z., "Stability of feature selection algorithms and its influence on prediction accuracy in biomedical datasets", TENCON 2014 - 2014 IEEE Region 10 Conference, 1-5, Bangkok, Thailand, (2014).

[6] Han, Y., Yu, L., "A variance reduction framework for stable feature selection", 2010 IEEE International Conference on Data Mining, 206-215, Sydney, NSW, Australia, (2010).

[7] Domingos, P., "A unified bias-variance decomposition and its applications", Proceedings of the 17th International Conference on Machine Learning, 231-238, Stanford, CA, USA, (2000).

[8] Munson, M.A., Caruana, R., "On feature selection, bias-variance, and bagging", ECML PKDD '09: Machine Learning and Knowledge Discovery in Databases, 144-159, (2009).

[9] Turney, P, "Technical note: bias and the quantification of stability", Machine Learning 20: 23-33, (1995).

[10] Alelyani, S., Liu, H., Wang, L., "The effect of the characteristics of the dataset on the selection stability", 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence, 970-977, Boca Raton, FL, USA, (2011).

[11] Gulgezen, G., Cataltepe, Z., Yu., L., "Stable and accurate feature selection", ECML PKDD '09: Machine Learning and Knowledge Discovery in Databases, 5781: 455-468, (2009).

[12] Chu, C., Hsu, A.-L., Chou, K.-H., Bandettini, P., Lin, C., "Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images", NeuroImage, 60(1): 59-70, (2012).

[13] Karabulut, E., Ozel, S., Turgay, I., "Comparative study on the effect of feature selection on classification accuracy", Procedia Technology, 1: 323-327, (2012).

[14] Janecek, A., Gansterer, W., Demel, M., Ecker, G., "On the relationship between feature selection and classification accuracy", Journal of Machine Learning Research, 4: 90-105, (2008).

[15] Amaldi, E., Kann, V., "On the approximation of minimizing non-zero variables or unsatisfied relations in linear systems", Theoretical Computer Science, 209(1-2): 237-260, (1998).

[16] Ang, J.C., Mirzal, A., Haron, H., Hamed, H.N.A., "Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 13(5): 971-989, (2015).

[17] Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., de Schaetzen, V., Duque, R., Bersini, H., Nowé, A., "A survey on filter techniques for feature selection in gene expression microarray analysis", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9(4): 1106-1119, (2012).

[18] Kohavi, R., John, G.H., "Wrappers for feature selection", Artificial Intelligence, 97(1-2): 273-324, (1997).

[19] Chandrashekar, G., Sahin, F., "A survey on feature selection methods", Computers and Electrical Engineering, 40(1): 16-28, (2014).

[20] Lal, T.N., Chapelle, O., Weston, J., Elisseeff, A., "Embedded methods", Feature Extraction, Studies in Fuzziness and Soft Computing, 207: 137-165, (2006).

[21] Cateni, S., Colla, V., Vannucci, M., "A hybrid feature selection method for classification purposes", 8th European Modeling Symposium on Mathematical Modeling and Computer Simulation EMS2014, 39-44, Pisa, Italy, (2014).

[22] Saeys, Y., Abeel T., Peer, V.Y., "Robust feature selection using ensemble feature selection techniques", ECML PKDD 2008: Machine Learning and Knowledge Discovery in Databases, 5212: 313-325, (2008).

[23] Khoshgoftaar, T.M., Fazelpour, A., Wang, H., Wald, R., "A survey of stability analysis of feature subset selection techniques", 2013 IEEE 14th International Conference on Information Reuse & Integration (IRI), 424-431, San Francisco, CA, USA, (2013).

[24] Khaire, U.M., Dhanalakshmi, R., "Stability of feature selection algorithm: a review", Journal of King Saud University - Computer and Information Sciences, 34(4): 1060-1073, (2022).

[25] Dua, D., Graff, C., "The UCI Machine Learning Repository", University of California, School of Information and Computer Science, Irvine, CA, http://archive.ics.uci.edu/ml, (2019).

[26] Alcala-Fdez, J., Fernandez, A., Luengo, J., Derrac, J., Garcia, S., Sanchez, L., Herrera, F., "KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework", Journal of Multiple-Valued Logic and Soft Computing, 17: 255-287, (2011).

[27] Berrar, D., "Cross-Validation", Encyclopedia of Bioinformatics and Computational Biology, 3: 542-545, (2018).

[28] Som, R.K., "Practical Sampling Techniques", Second Edition, United Kingdom, CRC Press, Taylor & Francis Group, 389-423, (1996).

[29] Wang, D., Zhang, H., Liu, R., "T-Test feature selection approach based on term frequency for text categorization", Pattern Recognition Letters, 45: 1-10, (2014).

[30] Reyes-Aldasoro, C.C., Bhalerao, A., "The Bhattacharyya space for feature selection and its application to texture segmentation", Pattern Recognition, 39(5): 812-826, (2006).

[31] Largeron, C., Moulin, C., Gery, M., "Entropy based feature selection for text categorization", SAC '11: Proceedings of the 2011 ACM Symposium on Applied Computing, Taichung, Taiwan, 924-928, (2011).

[32] Shilaskar, S., Ghatol, A., "Feature selection for medical diagnosis evaluation for cardiovascular diseases", Expert Systems with Applications, 40(10): 4146-4153, (2013).

[33] Serrano-Lopez, A., Olivas, E.S., Martín-Guerrero, J.D., Magdalena, R., Gómez-Sanchís, J., "Feature selection using ROC curves on classification problems", IJCNN '10: International Joint Conference on Neural Networks, 1-6, Barcelona, Spain, (2010).

[34] Theodoridis, S., Koutroumbas, K., "Pattern Recognition", 4th ed., USA: Academic Press, 261-322, (2009).

[35] Aha, D.W., Bankert, R.L., "A comparative evaluation of sequential feature selection algorithms", Lecture Notes in Statistics, Learning from Data, 112: 199-206, (1996).

[36] Alelyani, S., "On feature selection stability: a data perspective", Ph.D. Thesis, Arizona State University, Phoenix, USA, 10-40, (2013).

[37] González, J., Ortega, J., Damas, M., Martín-Smith, P., Gan, J.Q., "A new multi-objective wrapper method for feature selection – accuracy and stability analysis for BCI", Neurocomputing, 333: 407-418, (2019).

[38] Wang, A., Liu, H., Liu, J., Ding, H., Yang J., Chen, G., "Stable and accurate feature selection from microarray data with ensembled fast correlation based filter", 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2996-2998, Seoul, South Korea, (2020).

[39] Deraeve, J., Alexander, W.H., "Fast, accurate, and stable feature selection using neural networks", Neuroinformatics, 16(2): 253-268, (2018).

[40] Krizek, P., Kittler, J., Hlavac, V., "Improving stability of feature selection methods", 12th International Conference on Computer Analysis of Images and Patterns (CAIP), 929-936, Vienna, Austria, (2007).