Assesing the Stability and Selection Performance of Feature Selection Methods Under Different Data Complexity

Omaimah Al Hosni
School of Engineering,
University of Aberdeen,
UK
o.alhosni.19@abdn.ac.uk

Andrew Starkey
School of Engineering,
University of Aberdeen,
UK
a.starkey@abdn.ac.uk

Abstract: *Our study aims to investigate the stability and the selection accuracy of feature selection performance under different data complexity. The motivation behind this investigation is that there are significant contributions in the research community from examining the effect of complex data characteristics such as overlapping classes or non-linearity of the decision boundaries on the classification algorithm's performance; however, relatively few studies have investigated the stability and the selection accuracy of feature selection methods with such data characteristics. Also, this study is interested in investigating the interactive effects of the classes overlapped with other data challenges such as small sample size, high dimensionality associated with irrelevant features, and imbalance classes to provide meaningful insights into the root causes for feature selection methods misdiagnosing the relevant features among different real-world data challenges. This analysis will be extended to real-world data to guide the practitioners and researchers in choosing the correct feature selection methods that are more appropriate for a particular dataset. Our study outcomes indicate that using feature selection techniques with datasets of different characteristics may generate different subsets of features under variations to the training data showing that small sample size and overlapping classes have the highest impact on the stability and selection accuracy of feature selection performance, among other data challenges that have been investigated in this study. Also, in this study, we will provide a survey on the current state of research in the feature selection stability context to highlight the area that requires more attention for other researchers.*

Keywords: *Stability of feature selection, class overlapping, data characteristics, complex data.*

Received April 10, 2022; accepted April 28, 2022
https://doi.org/10.34028/iajit/19/3A/4

1. Introduction

In the context of feature selection, the main concern in using feature selection techniques is to improve the generalisation capabilities of the machine learning algorithms [4, 13, 43]. A wide range of feature selection algorithms have been developed in various application areas and proved to boost prediction accuracy. However, little attention has been paid to their stability, which is defined as the ability of the feature selection technique to produce the same results at each run, even following small perturbations of the dataset [13, 15, 39, 43, 44]. However, as these techniques were not intentionally developed to produce stable features thus, stability was not analysed and was generally neglected until recently [44, 39, 38]. The importance of having stable feature selection outcomes comes from the fact that there are some domains where feature selection is not used only to improve classification performance; more importantly, feature selection techniques are used as a knowledge discovery tool to identify the characteristic(s) of the observed event [20, 22]. For example, the medical domain encompassing bioinformatics, genetics, and medicine, require an understanding and identification of the relevant features as this is essential for discovering new hidden knowledge within the DNA (genes); this can guide the genetic analysis to pinpoint the critical biomarkers that help to diagnose a disease or its medication (i.e. they help to understand why they specific cause a disease, or why they would be instrumental in the treatment) [2]. However, having different subsets of features at each run under variations to the training data [4, 7] will confuse the domain experts and reduce their confidence in validating selected features. Furthermore, the practitioners mostly assume that if the data-target concept is fixed, the relevant features are also fixed and expect that the feature selection algorithms behave the same across different dataset's properties. So, to obtain accurate and stable feature selection outputs, it is necessary to explore the dataset's properties and use it as a guide to select the proper method for a given problem and enhance model interpretability.

However, there are relatively few criteria in the literature to evaluate the efficiency of feature selection outputs [44]. One widespread criterion used to evaluate feature selection techniques is the prediction performance of the selected features, which can only ever be an indirect evaluation of the feature selection

method. Another criterion that has recently drawn attention in the feature selection community is the stability of feature selection techniques. The researchers argued that besides performance accuracy, obtaining stable feature selection outcomes is vital to building a reliable and transparent model [4, 7]. Comparing both evaluation criteria (of predictive accuracy and stability) to assess the feature selection outputs, it has been found that the former depends on the inductive learning algorithms and the generalisation ability of feature selection methods, while the latter is dependent on the characteristics of the data [4, 7].

The remainder of this paper is presented as follows: section 2 explains our contribution. Section 3 provides a brief description of the stability of feature selection. Section 4 discusses the related works in the stability context. Section 5 describes the study methodology. Finally, section 6 presents study's conclusion.

2. Our Contribution

Our study provides a survey on the current state of research in the feature selection stability context as covered in the related works section. Another motivation for the research presented in this paper is that there are significant contributions in the research community from examining the effect of complex data characteristics on classification algorithms performance; however, relatively few studies have investigated the stability and the selection accuracy of feature selection methods with complex data characteristics. Accordingly, this study conducts an empirical study to validate this assumption by answering the following questions:

1. Do the following challenges affect feature selection stability and selection accuracy? (Irrelevant features /high dimensionality, noise, small sample size, imbalanced classes, class overlap and non-linearity of the decision boundaries).
2. Among these challenges, which most significantly impacts feature selection stability and accuracy?
3. Is the stability performance data-dependent or algorithm-dependent?
4. Is there a relationship between stability and the subsequent selection accuracy?

Answering the above questions will provide meaningful insights for the practitioners and researchers to choose the correct feature selection methods that are more appropriate for a particular dataset, if the qualities of the dataset are known, and give insight into when the methods fail with real-world datasets. Furthermore, the literature has noticed that most of the empirical studies in the context of feature selection stability examined the behaviour of filter methods with little focus on the embedded and wrapper methods due to the high computational cost for the latter. Thus, to meet this gap, this work conducts a comprehensive comparison study to explore the behaviour of six commonly used feature selection techniques from the filter, wrapper, and embedded methods.

3. Stability of Feature Selection

The stability of a feature selection method is defined as the degree of agreement between its outputs when applied to randomly selected subsamples from the same dataset [4, 30, 13, 49]. In other words, it is the insensitivity of the feature selection outcomes to variations in the training data set [30]. Other researchers consider an algorithm unstable if a minor change in data causes substantial changes in the feature selection subset [35]. However, many measurements/metrics have been proposed in the literature to quantify the similarity between the feature selection outputs to measure the stability performance. According to the literature, these measurements/metrics are constructed based on two concepts: either similarity-based or frequency-based. In the similarity-based concept, the similarity between different feature sets is computed, and the average similarity over all pairs of feature subsets is calculated. Whereas in the frequency-based approach, the frequency of the feature occurrence is calculated by representing the selected features as a binary string. Nogueira *et al*. [35] have stated five desirable properties of stability measure, which are: fully defined, strict monotonicity, bounds, maximum stability and correction for chance; a full description of these properties can be found in [19, 20].

However, based on the literature, the stability measures/metrics can be categorised according to the type of feature selection outputs, where it has three different representations [13, 14, 33, 34]:

3.1. Stability by Index

This measurement is proposed to handle a subset of features outputs where it represents the features as a binary vector with cardinality equal to the total number of features. To find the similarity between the subsets, the index measurements assess the amount of overlap between the resulting subsets and measure the stability accordingly. Examples for this measurement are Jaccard index, Dice's coefficient, and Kuncheva index.

3.2. Stability by Rank

This measurement is proposed to handle the ranking feature selection output; unlike the index measure, it assesses stability by evaluating the correlation between ranking outputs; an example of this method is
Spearman's Rank Correlation Coefficient (SRCC)

3.3. Stability by Weight

Similar to the rank method, this method assesses selection stability by evaluating the correlation between

two sets of weighted features outputs; an example of this method is Pearson's Correlation Coefficient (PCC).

4. Related Works

During the last decade, the stability issue has started to gain the attention of the feature selection community [4, 43]. Generally, researchers in the literature handled the stability issues differently; some studies examined the stability from a data perspective, while others investigated stability from the learning algorithm perspective. In the following section, we will cover the existing studies that focus on the stability issues; we have categorised the researchers' contributions into four groups based on the strategy adopted to tackle the stability issues, as shown in Figure 1. Worth noting that there might be additional studies in the literature that help indirectly to tackle the stability issues. However, our primary focus in this work is to present the existing studies that aim mainly to examine the feature selection stability.

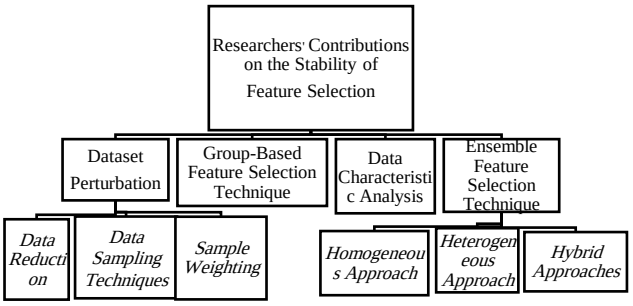

Figure 1.Taxonomy of researchers' contributions on the stability of feature selection.

4.1. Dataset Perturbation Technique

The researchers in the literature have proposed different data perturbation approaches to enhance the stability performance, which is usually implemented before applying any feature selection methods. So, the feature selection methods are applied to the perturbed data instead of the original dataset. However, current research shows less attention on this topic. The main concern is more on proposing ensemble methods to boost stability, which will be covered later in this study. The following sub-sections will show the studies conducted with this technique.

4.1.1. Data Reduction

Some researchers adopted a variance reduction approach to tackle the instability of feature selection outputs by perturbing the original dataset and creating new sub-samples from it. The researchers argued that one of the causes of instability is the impact of the high variance caused by the noise/outliers on the feature selection learning performance; hence creating several new reduced datasets by removing the outliers from the original dataset may help in reducing the adverse impact of the variance [3, 23]. Some works using this approach can be found in [3].

4.1.2. Data Sub-Sampling Techniques

The basic concept of this approach is to generate a sub-sample from the original dataset (this technique is usually applied in the small sample size dataset) and assess the stability of feature selection methods in each sub-sample under different levels of overlap degree between the sub-samples (the similarity between the subsamples). The primary purpose of this technique is to mitigate the effect of the data variation by controlling the underlying similarity between different sub-samples in the dataset since the researchers argued that the degree of overlap between the samples impacts the stability of feature selection methods [3, 7, 23, 25, 48]. Some works using this approach can be found in [23, 48].

4.1.3. Sample Weighting

The basic idea behind this approach is to assign each sample in the training set different weights based on the sample's influence on the feature relevance. Then feature selection methods are applied in the weighted training set [22, 29]. However, the feature relevance is determined by the samples' view or local profile according to the training data variations. Thus, if a sample has a noticeably different local profile from other samples, its existence in the training data will significantly impact the feature selection outcome. The principle of the local profile is that the high-density region that contains most of the instances is more relevant in determining the important features than the low-density region - which may contain outliers that may affect the learning process in diagnosing the important features. Therefore, according to this principle, instances in the low-density region should have lower instance weights compared to the high-density region; thus, the adverse effect of the data variance will be reduced in the learning process [22, 28, 29]. Some works using this approach can be found in [22, 28].

4.2. Ensemble Feature Selection Techniques

Recently researchers showed more attention to ensemble feature selection techniques by proposing frameworks that generate multiple random subsamples from the same original dataset (Homogeneous Approach) or combine multiple feature selection methods and aggregate its several outcomes into a single one (heterogeneous approach), in machine learning, this combination is called ensemble learning [4, 25, 43]. However, the researchers assumed that using such a technique would provide more accurate and stable results than results produced by a single feature selection method as it generates and aggregates different perspectives about the relevant features [4, 25, 29, 36,

43, 45, 48]. Compared to single-based learning, the authors in the literature emphasised that ensemble learning is a good tool for discovering hidden knowledge related to important features. Since it creates several hypotheses that reduce the risk of choosing wrong and unstable feature subsets, in other words, aggregating several feature selector's opinions will provide a more accurate estimation of the optimal feature's subset than a single selector opinion [15, 38, 41].

In terms of ensemble feature selection, there are three main types of this technique proposed in the literature: data diversity (homogeneous approach), functional diversity (heterogeneous approach), and a hybrid approach. Next, multiple ranking output lists will be produced after applying one of these types. Then similar to the classification ensemble model, multiple lists will be aggregated into a single list by using one of the aggregation functions proposed in the literature, such as mean aggregation, median aggregation, exponential aggregation and threshold-based aggregation [4, 43, 48]. The following sub-sections will discuss in more detail these types and show some recent studies conducted to tackle the stability issue.

4.2.1. Data Diversity (Homogeneous Approach)

Current studies in the ensemble feature selection method showed more interest in the homogeneous approach than the other two types mentioned above [4, 43]. However, the process starts by generating multiple random subsamples from the same original dataset to achieve the desired data diversity. Although many standard sampling approaches can be used in this step, such as bootstrapping, data split, k-fold cross-validation and over-sampling, the bootstrapping method is commonly used in the ensemble feature selection approach [43]. In the second step, a single feature selection technique is used for each subsample. The final step is to aggregate the different results produced from each subsample into a single result using the aggregation function [4, 16]. Recent studies using this approach in the context of feature selection stability can be found in [4, 38, 47].

4.2.2. Functional Diversity (Heterogeneous Approach)

The heterogeneous methodology follows the opposite way of the homogeneous approach; it applies multiple feature selection techniques in the same (single) original dataset throughout the process. after that, a ranked list for each feature selection technique will be produced and then aggregated into a single feature ranking list once all chosen techniques have been implemented [4, 16, 36]. However, the heterogeneous ensemble technique is a good approach for evaluating the individual-based selectors' strengths and weaknesses [9]. Recent studies using this approach to tackle the stability issue can be found in [9, 36, 45].

4.2.3. Hybrid Approaches

Based on the study done by Seijo-Pardo *et al*. [45], their experiment results indicated that the homogeneous and heterogeneous approaches showed different behaviours under various data characteristics, which is undesirable. However, to take advantage of these approaches' strengths and aid their weaknesses, researchers in the literature proposed a hybrid approach that combines both concepts [15, 16, 45]. Generally, the hybrid approach starts with a homogeneous strategy by generating different subsamples from the original training set. The next step is to apply a heterogeneous strategy using multiple feature selection techniques in each subsample. Finally, following the same step of the homogeneous and heterogeneous approaches, the results are aggregated into a single final ranked list using any aggregation function. However, the hybrid ensemble feature selection method has gained the attention of researchers due to its superiority for any given situation; still, there are minimal studies conducted in the context of feature selection stability [4, 16, 38, 43, 48]. Recent studies using this approach in the context of stability can be found in [15, 43].

4.3. Group-Based Feature Selection Technique

The group-based feature selection method aims to select the features relevant to the label at both levels: group level and individual feature level as well [42, 46, 49]. This method follows the principle of group-based learning [42], which involves two stages: feature group generation and feature group transformation [25]. In the feature group generation stage, the features are partitioned according to their similarity and grouped into the same group based on the degree of similarity. The next stage is feature group transformation, where the original feature space is transformed into a new form, representing each feature group as a single entity. Finally, the selection process is applied to the transformed feature space [25, 31, 42] However, in the context of feature selection stability, group-based-feature selection has received less attention in the literature compared to others approaches mentioned in this work [42]. Recent studies using this approach in the context of stability can be found in [8, 42].

4.4. Data Characteristic Analysis

Generally, the shared data issues that have been covered in the literature in the feature selection context are noise, missing values, outliers, high dimensionality, imbalanced class, inconsistency, redundancy, and small sample size. This is summarised by assessing the data characteristics being undertaken to use an appropriate feature selection method for the particular data problem. This section presents the studies that aim to assess the stability of feature selection behaviour against different data characteristics. A recent study by Ramezani *et al*.

[49] investigated the stability behaviour of six commonly used feature selection techniques with class and attribute noise. The experiments were performed on a clean dataset and injected with combinations of different levels of the gaussian noise distribution. The finding of the results indicated that the noise affects the stability performance [39]. A Study done by Altidor *et al*. [6] and Abu Shanab *et al*. [1] has reached a similar conclusion where their study aims to understand how combinations of different noise levels and specific data characteristics such as sample size and class imbalance, affect the feature selection stability [1, 6]. Another exciting work done by Alelyani and Liu [3] has examined the stability behaviour of several well-known feature selection algorithms under various datasets characteristics: dimensionality, the sample size, and the variation of the underlying distribution of the dataset. In terms of algorithm perspective, they have investigated the stability performance under the different sizes of the feature subset selected. The finding of this study indicated that the stability behaviour is data characteristic dependent. However, among all examined factors that have proven their influence on stability, the authors found that high dimensionality and the sample size significantly impact selection stability compared to other factors [5].

From a review of the literature, the researchers were mainly trying to tackle or mitigate the effect of the issues related to the data challenges, such as data variance resulting from the small sample size, noise, high dimensional data/irrelevant features, redundant (correlated) features, and imbalanced classes which proved that the stability behaviour of feature selection methods is strongly dependent on the data characteristics or data quality. However, our study assumes that the above problems do not necessarily impose serious difficulties in feature selection methods' stability and accuracy of the selected features if the classes are linearly separable in the input space. In fact, the interactive effects of other complex data characteristics such as overlapping classes and non-linearly separable relationships increase the chances of adverse effects on selection outcomes. To the best of our knowledge, a limited number of studies measure the interactive effects of overlapping classes with other data characteristics in the feature selection context, where more of the attention is on the classification algorithm context. For example, in terms of the classification algorithm, Barella *et al*. [10] Pascual-Triana *et al*. [37] investigated the effect of the imbalance problem on classification accuracy with complex data properties. Their studies implied that the imbalance problem is not considered severe if the classes are perfectly separated, but the problem arises when classes overlap. Furthermore, the authors emphasised that geometric characteristics of the data, such as overlapping classes and nonlinear separability, are considered amongst the most significant difficulties in the machine learning field and have proven their impact in degrading the classification algorithms' accuracy since it is not easily measured [10, 11, 37].

Based on that, this study aims to investigate this issue. We believe that exploring the relationship between the overlapping classes with the small sample size, high dimensionality, imbalanced classes, and noise will help in describing the root causes of the feature selection methods in misdiagnosing the relevant features, particularly for real-world data**,** and so will provide meaningful insights for the practitioners and researchers to choose the correct feature selection methods that are more appropriate for datasets**.**

5. Methodology

In this study, the experiment strategy has been designed to be in two phases which are:

5.1. First Phase: The Synthetic Datasets Experiment

In real-world problems, the datasets usually are associated with overlapping classes, complex decision boundary shapes (nonlinear separability amongst classes), high data sparsity resulting from the small size, high dimensionality and the presence of noise [12, 32]. However, to better understand the effects of the different data challenges on the stability and selection accuracy of the feature selection methods, it is crucial to have a controlled environment that enables us to assess the effect of each factor across different difficulty levels. Since it is hard to find real-world datasets that meet the requirements described above, creating synthetic datasets with controlled characteristics is used in the experiments of this study. Another important reason to use synthetic data is that the actual relevant features in the real-world datasets are often unknown, making analysis and comparison of the results of the feature selection methods difficult.

Thus, to simulate real-world problems, the experiment strategy in this paper is designed to be at five levels of difficulty according to the degree of class overlap, starting from the easy level (no overlap between the classes) to more challenging levels (classes overlapping to varying degrees) as shown in Figures 2 and 3. Our aim of using difficulty gradient levels is to investigate the interactive effect of the class overlaps with other expected data challenges in the real-world problems on stability and selection performance of feature selection methods which are: small data size, noise, high dimensionality/irrelevant features, non-linearity of the decision boundaries, and Imbalanced classes.

Thus, using such a strategy will help cover common scenarios in real-world problems and precisely allow the identification of the factor (s) that have the most significant impact on the stability and accuracy of the

tested feature selection methods.

5.1.1. Datasets description

In our empirical study, several difficulty levels are generated, with level one the easiest level (no classes overlapping) and four further levels with increasing classes overlapping. to make sure that each level has covered all data challenges mentioned above, we have generated four synthetic multi-class datasets (four classes) at each level that includes the following data challenges:

1. Sample size: to measure the effect of the different sample sizes in the feature selection methods, the synthetic datasets are generated in two scenarios: a small sample size scenario with (100) samples across different difficulty levels see Table 1 and a large sample size scenario with (1000) samples see Table 2. Thus, we can examine the impact (if any) of the different sample sizes on the feature selection performance.

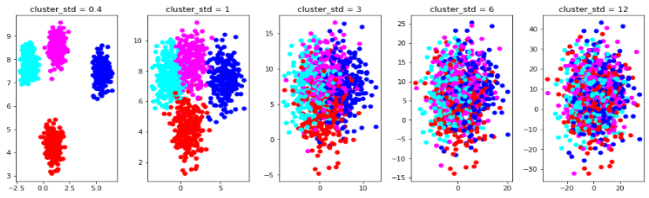

Figure 2.The graphical representation of classes distribution of the generated synthetic datasets (1000 sample size) across different difficulty levels.

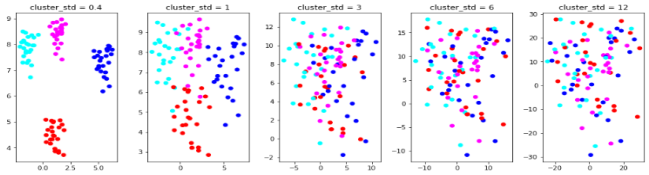

Figure 3. The graphical representation of classes distribution of the generated synthetic datasets (100 sample sizes) across different difficulty levels.

2. Relevant feature: the datasets are generated with six features relevant to the target classes as they contribute directly to the shape of the clusters. The same relevant features are used across different difficulty levels in both scenarios (small and large sample size) see Tables 1, and 2.
3. Imbalanced Classes: to assess the effect of the imbalanced classes on the feature selection methods, the controlled under-sampling technique in the balanced datasets is applied to generate new reduced imbalanced datasets* by eliminating several samples in the targeted classes based on the specified class ratio see Tables 1 and 2. Hence, the experiment will be repeated in both conditions (balanced and imbalanced classes) to examine the impact (if any) of the class imbalanced dataset on the stability and accuracy of feature selection performance.
4. Irrelevant features: to measure the effect of the irrelevant features, a number of irrelevant features are concatenated into the dataset (these features do not contribute information to the target classes). Here we aim to investigate the impact (if any) of the high dimensionality /irrelevant features on the stability and the selection performance of feature selection methods by examining its performance in different irrelevant feature sizes (44/994); see Table 1 for small sample size scenario and Table 2 for large sample size scenario.

Table 1. Dataset characteristics for (100) sample size dataset.

Datasets	The Diff. Levels	No. of Sample	No. of Relevant Feat.	Total No. of Feat.	The Classes Ratio
Dataset_1	Level One Cluster σ = 0.4	100	6	50	25:25:25:25
Dataset_2		60*			41:16:33:8
Dataset_3		100		1000	25:25:25:25
Dataset_4		60*			41:16:33:8
Dataset_5	Level Two Cluster σ = 1	100	6	50	25:25:25:25
Dataset_6		60*			41:16:33:8
Dataset_7		100		1000	25:25:25:25
Dataset_8		60*			41:16:33:8
Dataset_9	Level Three Cluster σ = 3	100	6	50	25:25:25:25
Dataset_10		60*			41:16:33:8
Dataset_11		100		1000	25:25:25:25
Dataset_12		60*			41:16:33:8
Dataset_13	Level Four Cluster σ = 6	100	6	50	25:25:25:25
Dataset_14		60*			41:16:33:8
Dataset_15		100		1000	25:25:25:25
Dataset_16		60*			41:16:33:8
Dataset_17	Level Five Cluster σ = 12	100	6	50	25:25:25:25
Dataset_18		60*			41:16:33:8
Dataset_19		100		1000	25:25:25:25
Dataset_20		60*			41:16:33:8

Table 2. Dataset characteristics for (1000) sample size.

Datasets	The Diff. Levels	No. of Sample	No. of Relevant Feat.	Total No. of Feat.	The Classes Ratio
Dataset_21	Level One Cluster σ = 0.4	1000	6	50	25:25:25:25
Dataset_22		600*			41:16:33:8
Dataset_23		1000		1000	25:25:25:25
Dataset_24		600*			41:16:33:8
Dataset_25	Level Two Cluster σ = 1	1000	6	50	25:25:25:25
Dataset_26		600*			41:16:33:8
Dataset_27		1000		1000	25:25:25:25
Dataset_28		600*			41:16:33:8
Dataset_29	Level Three Cluster σ = 3	1000	6	50	25:25:25:25
Dataset_30		600*			41:16:33:8
Dataset_31		1000		1000	25:25:25:25
Dataset_32		600*			41:16:33:8
Dataset_33	Level Four Cluster σ = 6	1000	6	50	25:25:25:25
Dataset_34		600*			41:16:33:8
Dataset_35		1000		1000	25:25:25:25
Dataset_36		600*			41:16:33:8
Dataset_37	Level Five Cluster σ = 12	1000	6	50	25:25:25:25
Dataset_38		600*			41:16:33:8
Dataset_39		1000		1000	25:25:25:25
Dataset_40		600*			41:16:33:8

5.1.2. The Data Complexity Measure

To measure the complexity of the generated synthetic datasets across different levels, the complexity measure proposed by Hoekstra, and Duin [24] is used. This metric is often used as a supporting pre-processing data task that measures to what extent the problem is complex for the classification algorithm, especially with

complex data characteristics such as overlapping classes or non-linearity of the decision boundaries. Since we are interested in examining the effect of the classes overlapped on the feature selection performance with the existence of other data challenges, thus, we used Neighborhood Measure (N4) to capture the shape of the decision boundary of the classes and to measure the class overlap's complexity; more details about these measures can be found in [17]. However, N4 produces a value in the range (0, 1), the low value indicates that the dataset is linearly separable, which is considered an easy problem, while a higher value indicates that the problem is more complex with a high degree of the classes overlapped [12]. Hence, we categorised the level of difficulty based on this value as shown below in Table 3.

Table 3. The difficulty levels.

Level	Noise Level	N4	Difficulty Degree
Level One	0.4	0	Easy
Level Two	1	0	Easy
Level Three	3	0.02	Medium
Level Four	6	0.20	Difficult
Level Five	12	0.40	Challenging

According to Fraça *et al*. [17] study, they consider the value of 0.35 as a challenging level.

5.1.3. The Stability Measure

To measure the stability behaviour of feature selection methods, we used the stability measure proposed by Nogueira *et al*, [35] the reasons for choosing this measure are that it attains all desirable properties of the stability measure mentioned in the stability of feature selection section. According to Nogueira *et al*, [35] the proposed stability measure is:

$$\hat{\Phi}(Z) = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d} s_f^2}{\mathbb{E}\left[\frac{1}{d}\sum_{f=1}^{d} s_f^2 | \mathrm{H}_0\right]} = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d} s_f^2}{\frac{k}{d}(1-\frac{k}{d})} \quad (1)$$

Where Φ defines the stability measure, and Z defines all collections of feature selection methods outputs, more details about the other parameters can be found in [35]. To implement the stability measure, we performed the following procedures:

- First, *let* $L^N = \{L_1, L_2, \ldots L_n\}$ be the synthetic datasets generated across different difficulty levels.
- Next, let Z_0 be the subset of predetermined top k ranked features obtained by applying the feature selection methods on the ideal datasets which are: (Dataset_1) see Table 1 and (Dataset_21) see Table 2. These datasets are clean of noise (no overlap between the classes), balanced classes and with low features size.
- Then, *let* $Z^N = \{Z_1, Z_2, \ldots, Z_n\}$ be the subsets of predetermined top k ranked features obtained in the perturbed dataset (L^N)
- Finally, a single stability index measure is applied for each feature selection method output (Z^N) of the perturbed datasets (L^N) and compared with the feature selection methods output (Z_0) of the clean of noise (no overlap between the classes), balanced classes, and low features size which are: (Dataset_1and Dataset_21) using the Equation (1).

Worth noting that the stability metric produces a value in the range (0, 1), the low value indicates the feature selection method provides unstable outcomes, whereas the high value indicates that the method has stable outcomes.

5.2. Feature Selection Methods

As mentioned in the our contribution section, another contribution of this study is to explore the behaviour of a combination of the filter, wrapper, and embedded feature selection methods due to little attention that has been paid to the embedded and wrapper methods in the context of feature selection stability. Thus, a comprehensive comparison has been conducted in this study that includes filter methods, which are ANOVA (F-Test) [27] and Mutual Information (MI) [40]. Furthermore, from Wrapper Methods Recursive Feature Elimination Cross-Validation (RFECV) with Support Vector Machine (SVM) estimator [21] and Genetic Algorithm (GA) with SVM estimator [50] are used, whereas in the embedded method Tree-Based feature selection [18] and LASSOCV [26] are used.

5.3. First Phase Experiment Results

The following section will present the study experiment results of the stability and selection performance of the methods across different difficulty levels in both scenarios (small sample size=100 and large sample size= 1000) associated with different data challenges, as mentioned in the methodology section. Worth noting that in this phase, since we already know what the relevant features are, thus evaluating the feature selection methods using classification algorithms will be skipped. Instead, we will evaluate the performance of the feature selection method based on its ability to correctly identify all the six relevant features and provide stable outcomes across different data challenges.

5.3.1. Level One and Two Outcomes

According to the N4, as shown in Table 3, levels one and two are categorised as an easy level problem where the classes are not overlapped, and the decision boundaries are linearly separable see Figures 2 and 3. The experiment results indicated that at both difficulty levels (one and two), most feature selection methods have correctly identified all six relevant features across different characteristics in both small and large sample sizes scenarios see Figures 4, 5, 6, 7, 8, 9, 10, and 11,

except GA and LASSOCV which are the only methods that failed to identify all relevant features at both levels.

Regarding the stability performance, the methods have shown similar behaviour to the feature selection performance. As most of the methods have produced stable results on both levels, see Figures 12, 13, 14, 15, 16, 17, 18, and 19, except for LASSOCV and GA, which have produced unstable outcomes in these difficulty levels.

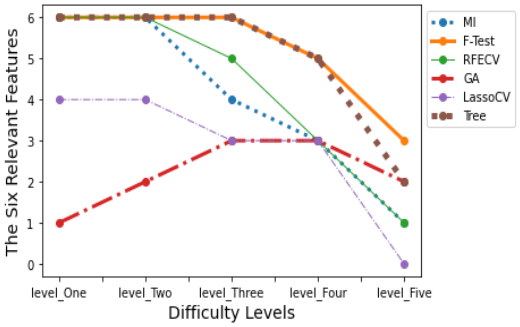

Figure 4. The selection performance of the small sample size scenario, balanced dataset (sample size=100 and feature size=50).

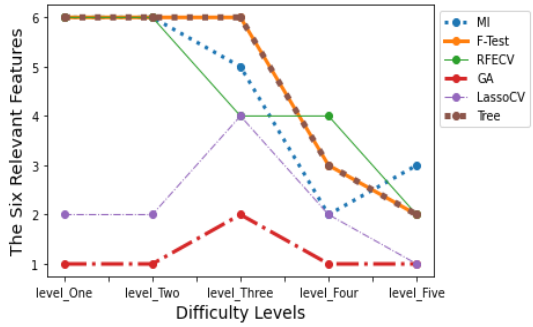

Figure 5. The selection performance of the small sample size scenario, imbalanced dataset (sample=60 and feature=50).

5.3.2. Level Three Outcomes:

Based on the complexity metric N4 see Table 3, this level is considered a medium-level difficulty in which the classes are partially overlapped at the decision boundary regions of the clusters/classes see Figures 1 and 2. Generally, at this level, the results indicated that the feature selection methods started to misdiagnose some of the relevant features and added the irrelevant ones. Furthermore, the methods have shown different behaviour in small and large sample sizes scenarios, performing the worst in small sample size scenario. In the small sample size scenario, most feature selection methods failed to identify all the six relevant features except ANOVA F-Test, and Tree-Based methods, which are the only methods that correctly identified all the six relevant features as shown in Figures 4, 5 and 6.

In contrast, in the large sample size scenario, most methods showed better performance in balanced classes datasets see Figures 8 and 10, except GA, which showed poor performance. Whereas, in the large sample size (imbalanced classes dataset), ANOVA F-Test, RFECV, and Tree-Based methods are the only methods that have correctly identified all six relevant features in the large feature size (1000) dataset see Figure 11. Whereas in the case of imbalanced classes of small features size (50) dataset, ANOVA F-Test and Tree-Based methods are the only methods that correctly identify relevant features see Figure9. In terms of stability performance, it almost has similar behaviour to the selection performance in this level; since the results indicated that some feature selection methods started to produce unstable results, specifically in the small sample size scenario. In this scenario the only methods that have produced stable outputs are ANOVA F-Test, and Tree-Based methods see Figures 12, 13, and 14, Except in an imbalanced classes dataset with a large feature size (1000), ANOVA F-Test is the only method that correctly identified all relevant features see Figure 15.

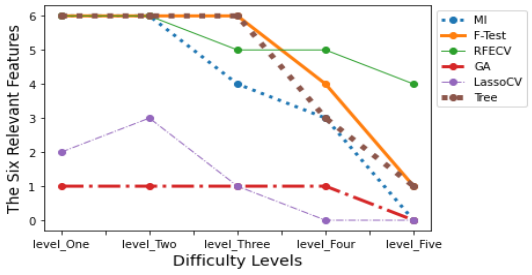

Figure 6. The selection performance of the small sample size scenario, balanced dataset (sample=100 and feature=1000).

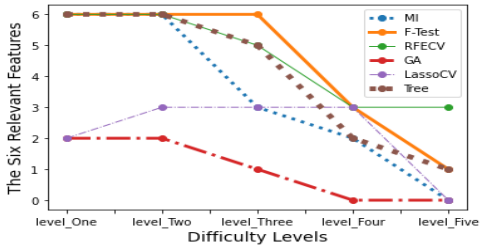

Figure 7. The selection performance of the small sample size scenario, imbalanced dataset (sample =60 and feature =1000).

In contrast, in the large sample size scenario, balanced classes dataset in both large (1000) and small (50) feature size, most methods produced stable results see Figures 16 and 18. except GA, which is the only method that produced unstable results in such dataset. However, in the case of the imbalanced classes dataset, LASSOCV, MI, RFECV, and GA showed unstable behaviour in the small feature size (50) dataset see Figure 17, and LASSOCV, MI and GA in the case of large features dataset (1000) see Figure 19.

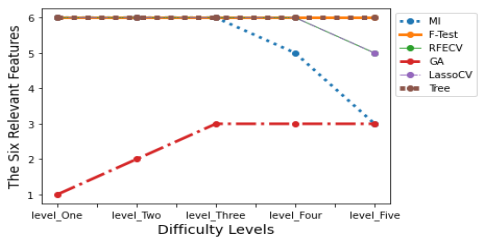

Figure 8. The selection performance of the large sample size scenario, balanced dataset (sample =1000 and feature =50).

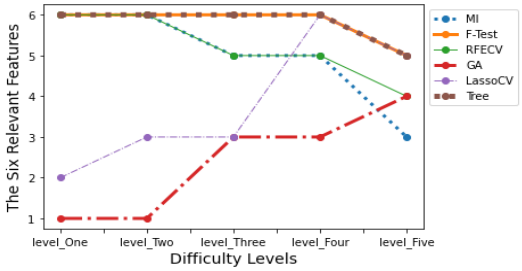

Figure 9. The selection performance of the large sample size scenario, imbalanced dataset (sample =600 and feature=50).

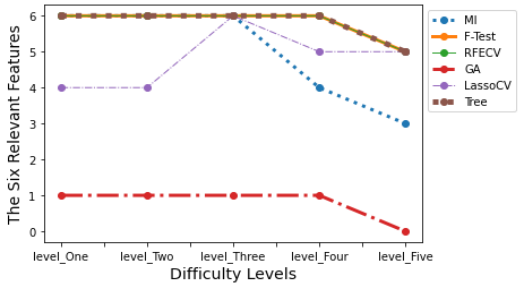

Figure 10. The selection performance of the large sample size scenario, in balanced dataset (sample =1000 and feature=1000).

5.3.3. Level Four and Five Outcomes

These levels are considered the most challenging level based on the complexity metrix (N4) as the classes are almost entirely overlapped, specifically in level five see Figures 2 and 3. The results show that all the feature selection methods failed to identify all six relevant features in all small samples size scenarios (balanced and imbalanced) in both (small and large feature size) see Figures 4, 5, 6 and 7. However, in the case of large sample size scenario balanced classes datasets, the only methods that correctly identified all relevant features are F-Test and Tree-Based methods only in the case of small feature size (50) dataset (at both levels four and five), and LASSOCV and RFECV at level four only see Figure 8. Whereas in the case of the large feature size (1000) dataset, all the methods failed to identify all six relevant features at level five, except at level four, where RFECV, ANOVA F-Test, and Tree-Based methods are the only methods that correctly identified all relevant features see Figure 10.

However, in the case of class imbalanced large sample size scenario, all methods failed to identify the relevant features at level five. In contrast, at level four, ANOVA F-Test, LASSOCV, and Tree-Based methods are the only methods that correctly identified all the relevant features in the small feature size (50) dataset see Figure 9. Whereas in the case of the large feature size (1000), all the methods failed to identify all six relevant features except F-Test see Figure 11.

Regarding the stability performance, the feature selection methods have produced unstable results in both levels except in large sample size balanced classes with the small feature size (50) dataset; here, ANOVA F-Test and Tree-Based methods are the only methods that produced stable results at both levels see Figure16. However, in the case of large feature size (1000), all methods have produced unstable results at level five. In contrast, at level four, ANOVA F-Test, Tree-Based and RFECV are the only methods that produced stable outcomes see Figures 18. For the large sample size imbalanced class dataset, all methods have produced unstable results at level five, except at level four, where ANOVA F-Test, Tree-Based and LASSOCV are the only methods that produced stable results in the case of the small feature size (50) see Figure 17. Whereas in the large feature size (1000) case, ANOVA F-Test is the only method that has produced a stable result at level four only see Figures 19.

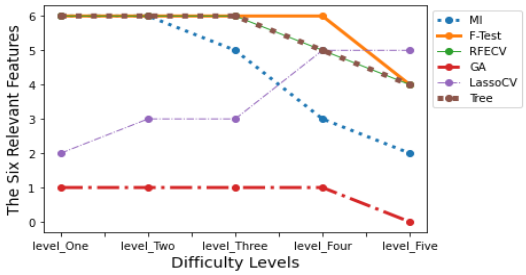

Figure 11. The selection performance of the large sample size scenario, in imbalanced dataset (sample=600 and feature=1000).

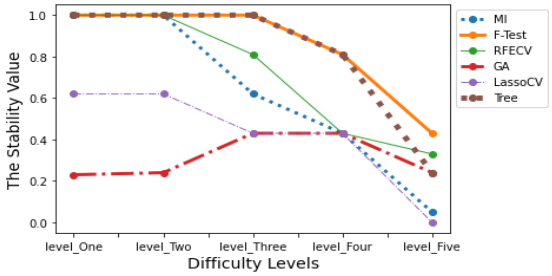

Figure 12. The stability performance of the small sample size scenario in the balanced dataset (sample=100 and feature=50).

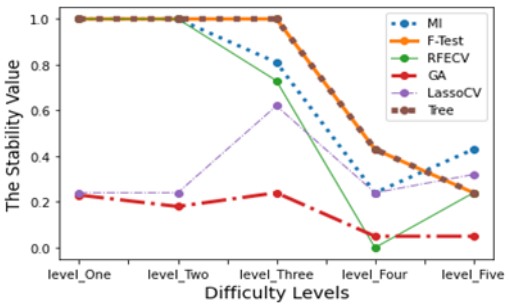

Figure 13. The stability performance of the small sample size scenario in the imbalanced dataset (sample=60 and feature=50).

It can be seen from the experiment results, as shown in the figures, that the overall feature selection methods performance showed good stability and selecting performance in identifying all the six relevant features without being affected by the existence of small sample size, high dimensionality, noise, and imbalanced classes when the classes are linearly separable (no classes overlapping) as it shown in the easy levels (level one and two). However, the methods started to misdiagnose some relevant features and added irrelevant ones when

the classes started to overlap (level three), and they continued degrading in missing more relevant features and added irrelevant ones as they moved to the upper level, especially in the challenging levels (level four and five). Concerning the stability performance across different levels, it is likely to have a similar selection performance as the methods started to produce unstable output when the classes started to overlap.

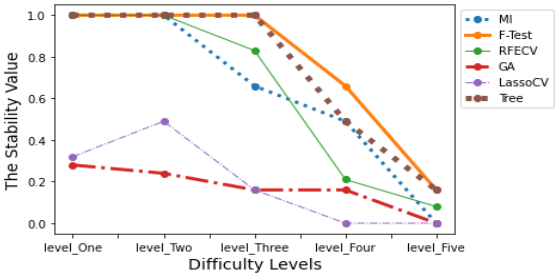

Figure 14. The stability performance of the small sample size scenario in the balanced dataset (sample=100 and feature=1000).

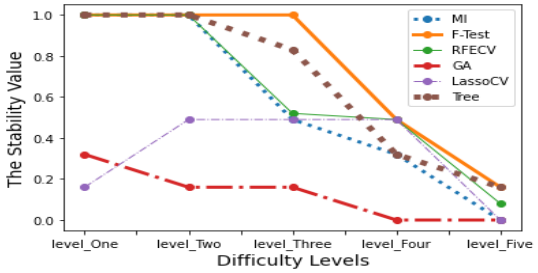

Figure 15. The stability performance of the small sample size scenario in the imbalanced dataset (sample=60 and feature=1000).

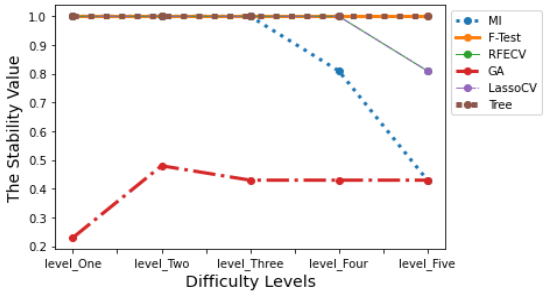

Figure 16. The stability performance of the large sample size scenario in the balanced dataset (sample=1000 and feature=50).

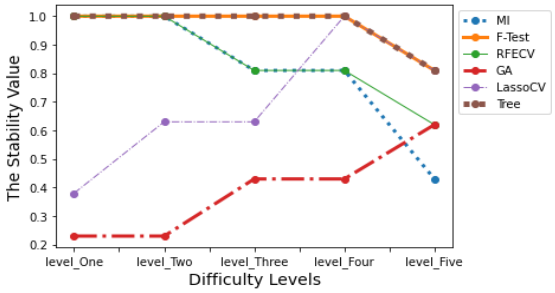

Figure 17. The stability performance of the large sample size scenario in the imbalanced dataset (sample=600 and feature=50).

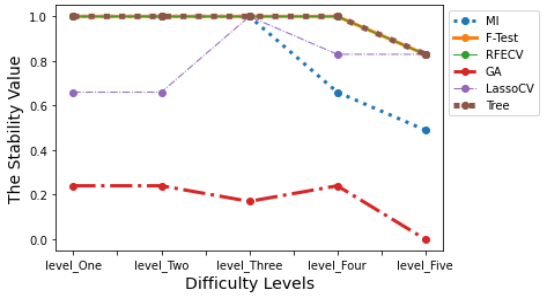

Figure 18. the stability performance of the large sample size scenario in the balanced dataset (sample=1000 and feature=1000).

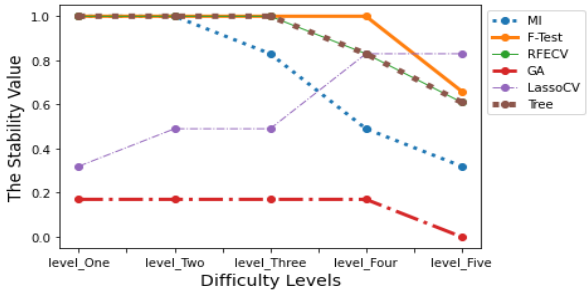

Figure 19. the stability performance of the large sample size scenario in the imbalanced dataset (sample=600 and feature=1000).

5.4. Second Phase: Real-World Dataset Experiment

Based on the first phase experiment outcomes, it has been proven that a small sample size and overlapped classes have the highest impact on the feature selection performance compared to the other data challenges that have been investigated in this study. So, in this phase, we have considered the following challenges when choosing the real-world dataset: small sample size, high dimensionality, overlapped classes and imbalanced classes' problem.

To meet these criteria, we have chosen the Gravier dataset [19], a microarray dataset with the properties described in Table 4. But in summary, it shows a challenging difficulty level according to complexity metrix N4 as the classes are overlapped see Figure 20 and it has unbalanced classes, a high number of features and a low number of samples.

Table 4. Gravier data description.

Sample Size	Number of Features	The classes ratio	Difficulty level
168	2905	(66%,33.9%)	Challenging

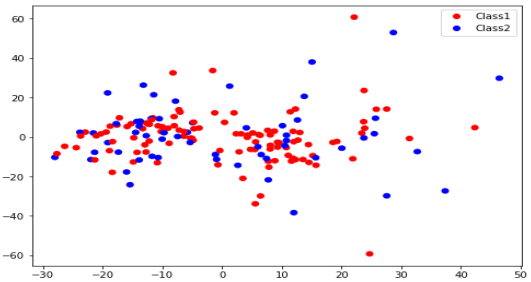

Figure 20. Graphical representation of the class distribution of the gravier dataset.

In the first phase experiment strategy, the results indicated that the Tree-Based and ANOVA F-Test methods had shown the best stability and selection accuracy performance compared to other methods. Accordingly, in the second phase experiment, we aim to validate this conclusion by investigating Tree-Based and ANOVA F-Test methods' performance in real-world datasets.

However, to assess the stability in the real-world dataset, the researchers in the literature have adopted different techniques, as we have covered in section 05.1. One of the techniques that have been widely used to assess the stability is splitting the original data into different sub-samples using different resampling techniques and then investigating the stability of feature selection methods in different sub-samples. Thus, in this study, we have used the k-fold cross-validation technique to assess the feature selection stability performance using two different k-folds cross-validation (k=10 and k=5). Hence after applying the feature selection methods in each k-fold, the feature selection outcomes of each k-fold will be compared to assess the feature selection methods' stability and accuracy selection in both k-fold datasets.

Since the relevant features are unknown in the real-world dataset, researchers and practitioners followed a standard procedure by specifying several k-subsets of features. Then, they evaluated the feature selection performance in different subsets relying on prediction accuracy to identify the relevant features. Based on that, this study has examined different feature subsets sizes in both datasets (k=5 and k=10 CV), as shown in Figures 21 and 22. To assess the accuracy, we have used SVM to evaluate the efficacy of the methods in identifying the relevant features. In terms of stability, we have followed the same stability measurement procedures used in the first phase.

5.4.1. The Second Phase Experiment Results

In this phase, we aim to investigate the stability and accuracy performance of the ANOVA F-Test and Tree-Based methods in the real-world dataset. The prediction accuracy for this dataset before the feature selection process is (0.57). however, from Figures 21 and 22 , it can be clearly seen that the prediction accuracy has increased after applying both feature selection methods (ANOVA F-Test and Tree-Based methods) across different feature subsets sizes and in both datasets (5 and 10-fold CV datasets).

However, in comparison to selection accuracy performance, both of them (ANOVA F-Test and Tree-Based methods) have shown similar behaviour in terms of the selection accuracy reaching the highest prediction accuracy in feature subset size = 30 in the 10-fold CV dataset; where the prediction accuracy for ANOVA F-Test is (0.75) and (0.78) for Tree-Based method see Figures 21, and 22.

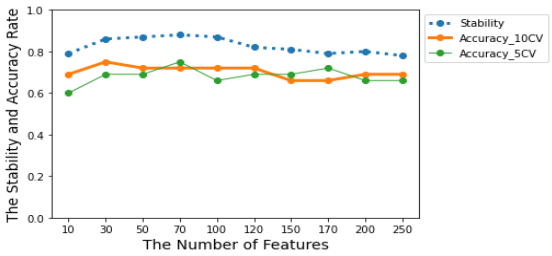

Figure 21. ANOVA F-TEST stability and accuracy performance.

In terms of stability performance, both methods have produced unstable outcomes across both k-fold datasets. However, the ANOVA F-Test method has produced more stable outcomes aligned with prediction performance across different feature subset sizes than the Tree-Based method.

On the other hand, comparing feature selection performance in both k-fold datasets, it can be seen from Figures 21 and 22 that both feature selection methods have better performance in the 10-fold CV datasets in terms of stability and selection accuracy than in the 5-fold datasets.

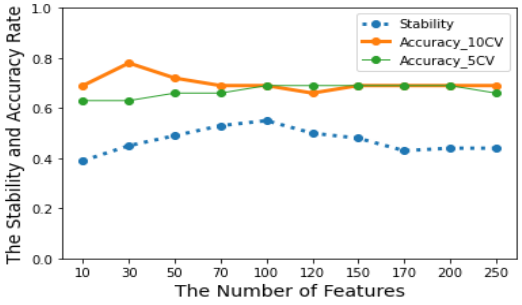

Figure 22. Tree-based method stability and accuracy performance.

6. Conclusions

This paper investigates the feature selection stability and accuracy performance against several real-world data challenges-high dimensionality/ irrelevant features, small sample size, noise, imbalanced dataset, and classes overlapped. The results showed that if the above issues are within the decision boundary of the class and the classes are linearly separated, the effect of the noise, irrelevant features/high dimensionality, and the imbalance of classes on the feature selection methods are relatively low. This outcome proved our assumption that the noise, high dimensionality, and imbalanced classes' issues are not necessarily imposing severe difficulties in the stability and accuracy of the selection performance if the classes are linearly separable. In fact, the interactive effects of the classes' overlapped, and nonlinear separability increase the chances of adverse effects on selection outcomes. Furthermore, this study showed that the small sample size and overlapping classes have the highest impact on the feature selection performance compared to other data challenges investigated in this study. In comparing both, the results indicated that class overlap has the most significant

effect on the stability and the accuracy of feature selection outputs since when the classes are linearly separable, the feature selection methods can identify the relevant features in both small and large sample sizes.

Related to the stability performance, the study gives a similar conclusion as other researchers in the literature. The results indicated that the stability is data-dependent since the feature selection methods produced unstable results across increasing difficulty levels. Also, from the study result, it can be noticed that there is a relationship between the feature selection accuracy and the stability performance as they are shown to have similar behaviour. Therefore, this paper shows that it is possible to use stability performance as an indicator to evaluate the efficiency of feature selection outputs with the classification algorithm prediction accuracy in case of real-world problems. In addition, overall, the result showed that the best performing feature selection methods in terms of stability and selection accuracy are the Tee-Based and ANOVA F-Test approaches, with the GA and LASSOCV the worst performing methods. however, LASSOCV performed poorly only in the cases of the small size datasets which proved that it is more sensitive to the variance caused by the small sample size datasets compared to other methods investigated in this study.

References

[1] Abu Shanab A., Khoshgoftaar T., Wald R., Napolitano A., "Impact of Noise and Data Sampling on Stability of Feature Ranking Techniques for Biological Datasets," *in Proceeding of the IEEE 13th International Conference on Information Reuse and Integration*, Las Vegas, pp. 415-422, 2012.

[2] Al Hosni O. and StarkeyA., "Stability and Accuracy of Feature Selection Methods on Datasets of Varying Data Complexity," *in Proceeding of the 22nd International Arab Conference on Information Technology*, Muscat, pp. 1-11, 2021.

[3] Alelyani S. and Liu H., "Supervised Low Rank Matrix Approximation for Stable Feature Selection, " *in proceeding of the 11th International Conference on Machine Learning and Applications*," Boca Raton, pp. 324-329, 202.

[4] Alelyani S., "Stable bagging feature selection on medical data," *Journal of Big Data*, vol. 8, no. 1, pp. 1-18, 2021.

[5] Alelyani S., Liu H., and Wang L., "The Effect of the Characteristics of the Dataset on the Selection Stability," *in Proceeding of the IEEE 23rd International Conference on Tools with Artificial Intelligence*, Boca Raton, pp. 970-977, 2011.

[6] Altidor W., Khoshgoftaar T., and Napolitano A., "Measuring Stability of Feature Ranking Techniques: A Noise-Based Approach," *International Journal of Business Intelligence and Data Mining*, vol. 7, no. 1-2, p. 80-115, 2012.

[7] Awada W., Khoshgoftaar T., Dittman D., Wald R., and Napolitano A., "A Review of the Stability of Feature Selection Techniques for Bioinformatics Data," *in Proceedings of IEEE 13th International Conference on Information Reuse and Integration*, Las Vegas, pp. 356-363, 2012.

[8] Bahekar K. and Gupta A., "Artificial Immune Recognition System-Based Classification Technique," *in Proceeding of the International Conference on Recent Advancement on Computer and Communication*. Bhopal, pp.629-635, 2018.

[9] Bania R. and Halder A., "R-HEFS: Rough Set Based Heterogeneous Ensemble Feature Selection Method for Medical Data Classification," *Artificial Intelligence in Medicine*, vol. 114, pp. 102049, 2021.

[10] Barella V., Garcia L., de Souto., M., Lorena A., and de Carvalho A., "Data Complexity Measures for Imbalanced Classification Tasks," *in Proceeding of the International Joint Conference on Neural Networks*, Rio de Janeiro, pp. 1-8, 2018.

[11] Barella V., Garcia L., de Souto M., Lorena A., and de Carvalho A., "Assessing the Data Complexity of Imbalanced Datasets," *Information Sciences*, vol. 553, pp. 83-109, 2021.

[12] Cano J., "Analysis of Data Complexity Measures for Classification," *Expert Systems with Applications*, vol 40, pp. 4820-4831, 2013.

[13] Chelvan P. and Perumal K., "A comparative Analysis of Feature Selection Stability Measures," *in Proceeding of the International Conference on Trends in Electronics and Informatics* , pp. 124-128, 2017.

[14] Chelvan P. and Perumal K., "On Feature Selection Algorithms and Feature Selection Stability Measures: A Comparative Analysis," *International Journal of Computer Science and Information Technology*, vol. 9, no. 3, pp. 159-168, 2017.

[15] Colombellia F., Kowalskib T., and Recamonde-Mendozaa M., "A Hybrid Ensemble Feature Selection Design for Candidate Biomarkers Discovery from Transcriptome Profiles," *arXiv preprint arXiv:2108.00290*, 2021.

[16] Dittman D., Khoshgoftaar T., Wald R., and Napolitano A., "Comparing Two New Gene Selection Ensemble Approaches with the Commonly-Used Approach," *in Proceeding of the 11th International Conference on Machine Learning and Applications*, Boca Raton, pp. 184-191, 2012.

[17] Fraça T., Miranda P., Prudêncio R., Lorenaz A., and Nascimento A., "A Many-Objective

Optimisation Approach for Complexity-Based Data Set Generation," *in Proceeing of the IEEE Congress on Evolutionary Computation*, Glasgow, pp. 1-8, 2020.

[18] Geurts P., Ernst D., and Wehenkel L., "Extremely Randomised Trees," *Machine Learning*, vol. 63, no. 1, pp. 3-42, 2006.

[19] GitHub. Gravier· ramhiser/datamicroarray Wiki. [online] Available at: https://github.com/ramhiser/datamicroarray/wiki/Gravier-%282010%29, Last Visited, 2022.

[20] Gulgezen G., Cataltepe Z., and Yu L., "Stable and Accurate Feature Selection," *in Proceeding of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Bled, pp. 455-468, 2009.

[21] Guyon I., Weston J., Barnhill S., and Vapnik V., "Gene Selection for Cancer Classification Using Support Vector Machines *Machine Learning*, vol. 46, no. 1, pp. 389-422, 2002.

[22] Han Y. and Yu L., "A Variance Reduction Framework for Stable Feature Selection," *Statistical Analysis And Data Mining: The ASA Data Science Journal*, vol. 5, no. 5, pp. 428-445, 2012.

[23] Haury A., Gestraud P., and Vert J., "The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures," *PLoS ONE*, vol. 6, no. 12, pp. e28210, 2011

[24] Hoekstra A. and Duin R., "On the Non-Linearity of Pattern Classifiers," *in Procceding of the 13th International Conference on Pattern Recognition*, Vienna, 271-275, 1996.

[25] Khaire U. and Dhanalakshmi R., "Stability of Feature Selection Algorithm: A review," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 4, 1060-1073, 2019.

[26] Kim S., Koh K., Lustig M., Boyd S., and Gorinevsky D., "An Interior-Point Method for Large-Scale -Regularized Least Squares," *in Proceeding of the IEEE Journal of Selected Topics in Signal Processing*, pp. 606-617, 2007.

[27] Kuhn M. and Johnson K., *Feature Engineering and Selection A Practical Approach for Predictive Models*, CRC press, 2019.

[28] Lei Y., Han Y., and Berens M., "Stable Gene Selection from Microarray Data via Sample Weighting," *IEEE/ACM Transactions on Computational Biology And Bioinformatics*, vol. 9, no. 1, pp. 262-272, 2012.

[29] LiY., Li T., and Liu H., "Recent Advances in Feature Selection and Its Applications," *Knowledge and Information Systems*, vol. 53, no. 3, pp. 551-577, 2017.

[30] Li Y., Si J., Zhou G., Huang S., and Chen S., "FREL: A Stable Feature Selection Algorithm," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 7, pp.1388-1402, 2014.

[31] Liu Y., Diao X., Cao J., and Zhang L., "Evolutionary Algorithms' Feature Selection Stability Improvement System," *in Proceeding of the International Conference on Bio-Inspired Computing: Theories and Applications*, Harbin, pp. 68-81, 2017.

[32] Lorena A., Garcia L., Lehmann J., Souto M., and Ho T. "How Complex is your Classification Problem," *ACM Computing Surveys*, vol. 52, no. 5, 2019.

[33] Mungloo-Dilmohamud Z., Jaufeerally-Fakim Y., and Peña-Reyes C., "Stability of Feature Selection Methods: A Study of Metrics Across Different Gene Expression Datasets," *in Proceeding of the International Work-Conference on Bioinformatics and Biomedical Engineering*, Granada, pp. 659-669, 2020.

[34] Naik A., Kuppili V., and Edla D., "A New Hybrid Stability Measure for Feature Selection," *Applied Intelligence*, vol. 50, no. 10, pp. 3471-3486, 2020.

[35] Nogueira S., Sechidis K., and Brown G., "On the Stability of Feature Selection Algorithms," *Journal of Machine Learning Research*, vol.18, no. 1, pp. 6345-6398, 2017.

[36] Noureldien N. and Mohammed E., "Measuring Success of Heterogeneous Ensemble Filter Feature Selection Models," *International Journal of Recent Technology and Engineering*, vol. 8, no. 6, pp.1153-1158, 2020.

[37] Pascual-Triana J., Charte D., Andrés Arroyo M., Fernández A., and Herrera F., "Revisiting Data Complexity Metrics Based on Morphology for Overlap and Imbalance: Snapshot, New Overlap Number of Balls Metrics and Singular Problems Prospect," *Knowledge and Information Systems*, vol. 63, no. 7, pp. 1961-1989 2021.

[38] Pes B., "Ensemble Feature Selection for High-Dimensional Data: A Stability Analysis Across Multiple Domains," *Neural Computing and Applications*, vol. 32, no. 10, pp. 5951-5973, 2020.

[39] Ramezani I., Niaki M., Dehghani M., and Rezapour M., "Stability Analysis of Feature Ranking Techniques in The Presence of Noise: A Comparative Study," *International Journal of Business Intelligence and Data Mining*, vol. 17, no. 4, pp. 413- 427, 2020.

[40] Ross B., "Mutual Information between Discrete and Continuous Data Sets," *PLoS ONE*, vol. 9, no. 2, pp. pp. e87357. 2014.

[41] Saeys Y., Abeel T., and Peer Y., "Robust Feature Selection Using Ensemble Feature Selection Techniques," *in Proceeding of the European Conference on Machine Learning and Knowledge Discovery in Databases*, Antwerp, pp. 313-325,

2008.
[42] Şahin C. and Diri B., "Robust Feature Selection And Classification Using Heuristic Algorithms Based On Correlation Feature Groups," *Balkan Journal of Electrical and Computer Engineering*, vol. 9, no. 1, pp. 23-32, 2021.
[43] Salman R., Alzaatreh A., Sulieman H., and Faisal, S., "A Bootstrap Framework for Aggregating within and between Feature Selection Methods," *Entropy*, vol. 23, no. 2, pp. 200, 2021.
[44] Sechidis K., Papangelou K., Nogueira S., Weatherall J., and Brown G., "On The Stability of Feature Selection in the Presence of Feature Correlations," *in Proceeding of the European conference on machine learning and knowledge discovery in databases*, Wurzburg, pp. 327-342, 2019.
[45] Seijo-Pardo B., Bolón-Canedo V., Porto-Díaz I., and Alonso-Betanzos A., "Ensemble Feature Selection for Rankings of Features," *in Proceeding of the International Work-Conference on Artificial Neural Networks*, Palma de Mallorca, pp. 29-42, 2015.
[46] Shang Z. and Li M., "Feature Selection Based on Grouped Sorting," *in Proceeding of the 9th International Symposium on Computational Intelligence and Design*, Hangzhou, pp. 451-454, 2016.
[47] Wang A., Liu H., Liu J., Ding H., Yang J., and Chen G., "Stable and Accurate Feature Selection from Microarray Data with Ensembled Fast Correlation Based Filter," *in Proceeding of the IEEE International Conference on Bioinformatics and Biomedicine*, Seoul, pp. 2996-2998, 2020.
[48] Wang H., Khoshgoftaar T., Wald R., and Napolitano A., "A Novel Dataset-Similarity-Aware Approach for Evaluating Stability of Software Metric Selection Techniques," *in Proceeding of the IEEE 13th International Conference on Information Reuse and Integration*, Las Vegas, pp. 1-8, 2012.
[49] Yu L., Ding C., and Loscalzo S., "Stable Feature Selection Via Dense Feature Groups," *in Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, pp. 803-81, 2008.
[50] Zhuo L., Zheng J., Li X., Wang F., Ai B., and Qian J., "A Genetic Algorithm Based Wrapper Feature Selection Method for Classification of Hyperspectral Images Using Support Vector Machine," *in Proceeding of the Geoinformatics and Joint Conference on GIS and Built Environment: Classification of Remote Sensing Images*, Guangzhou, pp. 503-511, 2008.

Omaimah Al Hosni is a lecturer at the University of Technology and Applied Sciences in Oman and a PhD Student at the University of Aberdeen. She received her master's degree in Computing from the University of Bradford, UK. Her research interests are in Artificial Intelligence, Feature Selection and Multi-label Classification.

Andrew Starkey is a Senior Lecturer at the University of Aberdeen and has research interests in Green AI, explainable AI, automated AI and autonomous learning methods. He has been awarded an Enterprise Fellowship from the Royal Society of Edinburgh and Scottish Enterprise. He is also responsible for Blueflow Ltd, a spin-out company from the University of Aberdeen that proposed a solution for a wide range of data analysis areas, such as financial, textual, and web data, such as blogs and discussion threads and condition monitoring.