# Stability and Accuracy of Feature Selection Methods on Datasets of Varying Data Complexity

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

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

***Abstract*— One widespread criterion used to evaluate feature selection techniques is the classifier performance of the selected features. Another criterion that has recently drawn attention in the feature selection community is the stability of feature selection techniques. Our study indicates that using feature selection techniques with different data characteristics may generate different subsets of features under variations to the training data. Our study motivation is that there are significant contributions in the research community from examining the effect of complex data characteristics such as class overlap on classification algorithms performance; however, relatively few studies have investigated the stability and the accuracy of feature selection methods with complex data characteristics. Accordingly, this study aims to conduct empirical study to measure the interactive effects of the class overlap with different data characteristics so we will provide meaningful insights into the root causes for feature selection methods misdiagnosing the relevant features among different data challenges associated with real world data in which will guide the practitioners and researchers to choose the correct feature selection methods that are more appropriate for particular dataset. Also, in this study we will provide a survey on the current state of research in the feature selection stability context.***

***Keywords—Stability of Feature Selection, Class Overlapping, Data Characteristics, Complex Data.***

## 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 [1][2][3]. 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 each time, even following small perturbations of the dataset [4][3][5] [6][1]. On the other hand, as these techniques were not intentionally developed to produce stable features thus, stability was not analysed and was generally neglected until recently [6][5][7]. For example, 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 [8][9]. Domains such as 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 the specific features lead to a disease, or why they would be instrumental in the treatment). Other practical problems occur in the microarray dataset such as high dimensionality (in most cases exceeding the number of samples) and low sample size (often less than a hundred)[10]. Such data characteristics will add challenges to the feature selection learning performance, making it highly sensitive to data variations since not all the features contribute to the class due to the small sample size [11]. Using feature selection techniques in such cases may generate different subsets of features under variations to the training data [2] [10]. Such a situation will confuse the domain experts and reduce their confidence in the validation of 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, in the literature, there are relatively few criteria to evaluate the efficiency of feature selection outputs [6]. 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 [2][10].

Comparing both evaluation criteria (of predictive capability 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 [2][10].

The remainder of this paper is presented as follows. Section II explains Our Contribution. Section III provides a brief description of Stability of feature selection. Section IV discusses the Related Works in the stability context. In section V, describes the study
Methodology. Section VI presents The Result and VII Section provides the study
Conclusion.

## OUR CONTRIBUTION

Our study provides a survey on the current state of research in the feature selection stability context. From our review of the literature, it can be realised that the stability behavior of feature selection methods is strongly dependent on the data characteristics or data quality. Despite the different

types of proposed solutions as covered in Related Works section, the researchers were mainly trying to tackle or mitigate the effect of the issues related to data characteristics such as: data variance resulting from small sample size with high dimensional data; noise; redundant (correlated) features; and imbalanced classes. However, our study assumes that the above problems do not necessarily impose serious difficulties in feature selection methods' stability and the accuracy of feature selections methods if the classes are linearly separable in the input feature space. In fact, the interactive effects of other complex data characteristics such as overlapping classes and non-linearly separable data problems such as those associated with complex data shapes increase the chances of adverse effects on selection outcomes; for example, V. H. Barella et al. (2018); Barella et al. (2021); Pascual-Triana et al. (2021), Fu et al., (2020) investigated the effect of the imbalance problem on the 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 are overlapping. Furthermore, the authors emphasised that geometric characteristics of the data, such as class overlapping and non-linear 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 [12][13][14][15].

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 the selection accuracy? Irrelevant features / high dimensionality, Noise, Small sample size, Imbalanced classes and Class Overlap.
2. Among these challenges, which most significantly impacts feature selection stability and the selection 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 particular dataset, if the qualities of the dataset are known, and give insight into when the methods will fail with real world datasets. Furthermore, it has been noticed from the literature that most of the empirical studies in the context of feature selection stability examined the behavior of filter methods with little focus on the embedded and wrapper methods due to the high computational cost foe the later. Thus, to meet this gap, this work conducts a comprehensive comparison study to explore the behavior of six commonly used feature selection techniques from the filter, wrapper, and embedded methods.

## 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 data set [16][17][18] [1][2]. In other words, it is the insensitivity of the feature selection outcomes to variations in the training data set [18]. Other researchers consider an algorithm unstable if a minor change in data causes substantial changes in the feature selection subset [19].

To measure the stability performance, many measurements/metrics have been proposed in the literture to quantify the similarity between the feature selection outputs. However, 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 approaches, the frequency of the feature occurrence is calculated by representing the selected features as a binary string [20][21][22][1].

Nogueira et al. (2018), have stated five desirable properties of stability measure which are: fully defined, strict monotonicity, bounds, maximum stability and correction for chance, a full description about these properties can be found in [19][20].g Based on the literature, the stability measures/metrics can be categorized according to the type of feature selection outputs, where it has three different representations[20][21][1][22]:

### *A. 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. In order to find the similarity between the subsets, the index measurements assess the amount of overlap between the resulting subsets and measure the stability accordingly. Some examples for this measurement are Jaccard Index Dice's Coefficient, Tanimoto Distance and Kuncheva Index.

### *B. 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.

### *C. 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.

## RELATED WORKS

During the last decade, the stability issue has started to gain the attention of the feature selection community[2][3][23]. 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 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, which are: Dataset Perturbation Technique; Ensemble Feature Selection Technique; Group-Based Feature Selection Technique; and Data Characteristic Analysis. 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.

### *A. Dataset Perturbation Technique*

The researchers in the literature have proposed different data perturbation approaches to enhance the stability performance, which are 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, where the main concern is more on proposing new ensemble methods to boost stability which will be covered in this study later. The following sub-sections will show the studies conducted in this context.

#### *(i) 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 causes 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[16][27][24]. Some works using this approach can be found in[16][27].

#### (ii) *Data Sampling Techniques*

The basic concept of this approach is to generate a sub-sample from the original dataset (usually 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[10][24][25][27][26]. Some works using this approach can be found in[24][25].

#### *(iii) 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 [28][8]. 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 on the learning process[28][8][29]. Although researchers have used many measures to calculate the local profile, the standard measures used in the literature are based on the Sample Margin and Hypothesis of Margin [28][30]. Some works using this approach can be found in [8][29].

### *B. Ensemble Feature Selection Techniques*

Recently researchers showed more attention to ensemble feature selection techniques by proposing frameworks that combined multiple feature selection methods and aggregate its several outcomes into a single one; in machine learning, this combination is called ensemble learning [3][2][26]. However, the researchers assumed that using such a technique will 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[3][2][26][28][31][25][32]. Compared to single-based learning, the authors in the literature emphasised that ensemble learning is a good tool for discovering hidden knowledge related to the important features. Since it creates several hypotheses that reduce the risk of choosing wrong and unstable feature subsets, in other words, producing different feature selection outputs creates different local optima in the feature space. Therefore, aggregating several feature selectors opinions will provide a more accurate estimation of the optimal feature's subset than a single selector opinion[4][7][33].

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. However, after applying one of these types, multiple ranking output lists will be produced. 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 [2][3][26][33][25][23][7]. The following sub-sections will discuss in more detail these types and show some recent studies conducted to tackle the instability issue.

#### *(i) 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[3][2]. However, to achieve the desired data diversity, the process starts by generating multiple random subsamples from the same original dataset. 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 [3]. 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 to a single result using the aggregation function[2][34]. Recent studies using this approach in the

context of feature selection stability can be found in [2][7][35].

*(ii) 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[2][32][34]. However, the heterogeneous ensemble technique is a good approach for evaluating the individual-based selectors' strengths and weaknesses[36]. Recent studies using this approach to tackle the stability issue can be found in[32][36][31].

*(iii) Hybrid Approaches*

Based on the study done by Seijo-Pardo et al. (2015), 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 [31][4][34].

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 by 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, recently 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[3][2][25][34][7]. Recent studies using this approach in the context of the stability can be found in[4][3].

### *C. 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[37][38][17]. This method follows the principle of group-based learning[37], which involves two stages: Feature Group Generation and Feature Group Transformation[26]. 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[26][23][37]. 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[37].Recent studies using this approach in the context of the stability can be found in[37][39].

### *D. Data Characteristic Analysis*

Generally, the common 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 so that an appropriate feature selection method is used for the particular data problem. This section will present the studies that aim to assess the stability of feature selection behaviour against different data characteristics. A recent study by Ramezani et al. (2020) 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[5]. Similar to the above study work by Altidor et al. (2012), and Shanab, A. A. et al.(2012), reached a similar conclusion where their study aims to understand how combinations of different noise levels and specific data characteristics auchas sample size and class imbalance, affect the feature selection stability[42][11]. Another interesting work done by S. Alelyani et al. (2011) has examined the stability behaviour of several well-known feature selection algorithms under various datasets characteristics: dimensionality, the absolute 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. To investigate the most significant between the sample size and dimensionality, the study showed that the sample size has more influence than high dimensionality on the stability[43].

Based on the above studies, it can be realised that the stability behaviour of feature selection methods is strongly dependent on the data characteristics or data quality. Moreover, the researchers were mainly trying to tackle or mitigate the effect of the issues related to the data characteristics, which are the data variance resulting from the small sample size , high dimensional data, the noise, the redundant (correlated) features, and the imbalanced classes. 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 feature 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. Thus, due to a limited number of studies measuring the interactive effects of classes overlapping with other data characteristics in feature selection context, this study aims to investigate this issue. We believe that exploring the relationship between the classes overlapping with the small sample size, high dimensionality, imbalance classes, and noise will help describe 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**.**

## Methodology

### *A. Experiment Strategy*

In real-world problems, the datasets are normally associated with overlapping classes and complex decision boundary shapes (non-linear separability amongst classes) and also high data sparsity resulting from the small size, high dimensionality and the presence of noise[45][47]. Furthermore, the ambiguity regarding the structural and geometric properties of the data distribution might mislead the practitioners about the actual causes of any misclassification errors, which are usually accrued in the overlapping regions[45]. Based on that, the experiment strategy in this paper is designed to be in five levels of difficulty according to the degree of class overlap. Thus, the experiment strategy aims to simulate real-world problems using different difficulty level starting from easy level (no overlap between the classes) to harder levels (classes overlapping to varying degrees). We believe that using the difficulty gradient levels will help in covering common scenarios in real-world problems and in turn will allow precisely the identification of the factor(s) that have the most significant impact on the stability and accuracy of the tested feature selection methods.

### *B. Dataset*

To gain a better understanding of the effects of the different factors on the stability and accuracy of the selection of 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 is the actual relevant features in the real-world datasets are often unknown, which can make analysis and comparison of the results of the feature selection methods difficult. The following sections describe the steps for generating the data and the experimental procedures of each level.

#### *(i) Generating Synthetic Dataset:*

The computer programming language used in this study is Python. Hence, to generate the synthetic dataset, the scikit-learn library in Python was used, which includes a set of data generation functions that allow the users to create simulated datasets of specified properties which can be used to investigate algorithms behaviour. However, in this study, a four multiclass dataset was generated using the make_blobs() function to generate blobs of samples of a Gaussian distribution. The reasons for using this function are that it provides greater control over: (1) the number of centres (clusters/blobs/classes), (2) the degree of overlapping of classes by specifying the standard deviations of the (clusters/blobs), (3) the number of relevant features, and other properties[44].

The aim of this study is to explore the relationship between the overlapping classes with the other data characteristics, which are: small sample size, high dimensionality, imbalanced classes, and noise; therefore, the synthetic datasets are generated to include all these data properties assessed across five levels of difficulty based on the degree of classes overlap as shown in the Fig. 1 & Fig. 2

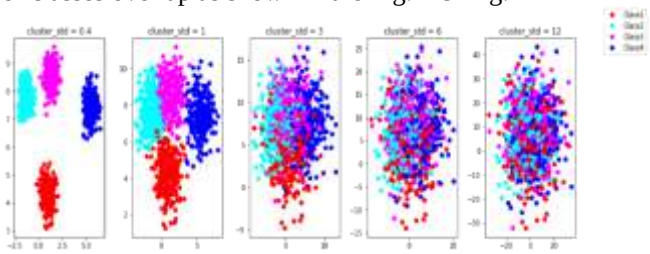

Fig. 1. The Graphical Representation of Classes Distribution of the Generated Synethtic Datasets (1000 smaple size) Across Different Difficulty Levels.

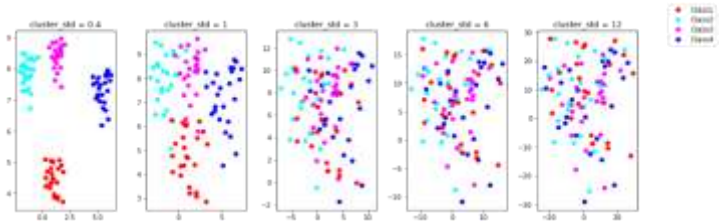

Fig. 2. The Graphical Representation of Classes Distribution of the Generated Synethtic Datasets (100 smaple size) Across Different Difficulty Levels.

#### *(ii) Dataset description*

In our empirical study, several levels are generated, with Level One the easiest level having no class overlap, and then four further levels having increasing class overlap. In each level we have generated four synthetic datasets based on the following properties:

- *Degree of class overlap*: to control the spread of the samples in each cluster we injected Gaussian noise distribution in the samples using the probability density for the Gaussian distribution:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (1)$$

where $\mu$ is the mean and $\sigma$ the standard deviation[48]. So, the spread of the samples in each cluster is controlled by changing the standard deviation of the clusters across different levels while maintaining $\mu$ =0. Level One has $\sigma$= 0.4, Level Two $\sigma$ = 1, Level Three $\sigma$ = 3, Level Four $\sigma$= 6 and Level Fve $\sigma$= 12.

- *Sample size*: to measure the effect of the small sample size in the feature selection methods, the datasets are generated in two scenarios, with 100 samples across different levels. Then the same data generation procedure is repeated with 1000 samples scenario; so we can examine the impact of the different sample size on the feature selection methods performance. For the details about the data characteristics of 100 and 1000 sample size datasets see TABLE I. 0

TABLE I. DATASETS CHARACTERISTICS FOR 100 SAMPLES DATASETS

| Datasets | The Diff. Levels | No. of Sample | No. of Relevant Feat. | Total No. of Feat. | The Classes Ratio |
|---|---|---|---|---|---|
| Dataset_1 | Easy Level 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 | Easy Level 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 | Med. Level 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 | Diff. Level 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 | Diff. Level 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 II. DATASETS CHARACTERISTICS FOR 1000 SAMPLES DATASETS

| Datasets | The Diff. Levels | No. of Sample | No. of Relevant Feat. | Total No. of Feat. | The Classes Ratio |
|---|---|---|---|---|---|
| Dataset_21 | Easy Level 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 | Easy Level 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 | Med. Level 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 | Diff. Level 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 | Diff. Level 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 |

- *Relevant feature*: the datasets are generated with six features relevant to the target classes as they contribute directly to the shape of the clusters.

- *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- by generating samples from the standard normal distribution of $\mu$ mean=0 and $\sigma^2$ =1[49], of shape 100 samples & 44 features for the (Dataset_1 & Dataset_2) and shape of 100 samples & 994 features for the (Dataset_3 & Dataset_4) , following the same procedure in the 1000 sample size datasets scenario see TABLE I. TABLE II. .

- *Imbalanced Classes*: to assess the effect of the imbalanced classes on the feature selection methods, 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. To perform the under-sampling technique, we follow the procedures below:

- Let the classes ratios be $alpha_{us}$ as defined by

$$N_{rm} = alpha_{us} * N_m \quad (2)$$

where $N_{rm}$ and $N_m$ are the number of samples in the majority class after resampling and the number of samples in the minority class respectively[50].

*(i) The Data Complexity Measure:*

To measure the complexity of the generated synthetic datasets across different levels, one of the complexity measures is used which proposed by Hoekstra, A. & Duin, R. (1996) is used [51]. This metric is often used as supporting pre-processing data tasks that measure to what extent the problem is complex, 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 class overlap with other characteristics thus, we used Neighborhood Measure (N4) to capture the shape of the decision boundary and to characterise the class overlap; more details about these measures can be found in[52]. However, N4 produces a value is in the range [0, 1]; the low value indicates that the dataset is linearly separable, which is considered an easy problem while higher value indicated that the problem is more complex [47]. Hence, we categorised the level of difficulty based on this value as shown below in the 0

TABLE III. THE DIFFICULTY LEVELS

| Level | N4 | The Cluster Std | Difficulty Degree |
|---|---|---|---|
| Level One | 0 | 0.4 | Easy |
| Level Two | 0 | 1 | Easy |
| Level Three | 0.02 | 3 | Medium |
| Level Four | 0.20 | 6 | Difficult |
| Level Five | 0.40 | 12 | Very Difficult |

According to T. R. Fraça et al. (2020) study, they consider the value of 0.35 as a very difficult level.

*(ii) The Stability Measure:*

To measure the stability behaviour of feature selection methods, we used the stability measure proposed by Nogueira et al., 2018. 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 and allows the development of a statistical framework for quantifying stability which are not possessed by other stability measures proposed in the literature. According to Nogueira et al., 2018 the proposed stability measure is:

$$\hat{\Phi}(Z) = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d} s_f^2}{\mathbb{E}\left[\frac{1}{d}\sum_{f=1}^{d} s_f^2 | \mathrm{H}_0\right]} = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d} s_f^2}{\frac{k}{d}(1-\frac{k}{d})} \quad (3)$$

Where $\hat{\Phi}$ defines the stability measure and Z defines all collections of feature selection methods outputs, more details
about the other parameters can be found in [19]. 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 levels.

- Next let $z_0$ be the subset of predetermined top k ranked features obtained by applying the feature selection methods

on the (Dataset_1) the clean of noise dataset (balanced classes and no overlap between the classes) see TABLE I.

- Then, let $Z^N = \{z_1, z_2, \dots, z_n\}$ be the subsets of predetermined top k ranked features obtained in the perturbed datasets $(L^N)$.
- Finally, a single stability index measure applied for each feature selection method output ( $Z^N$) of the in the perturbed datasets $(L^N)$ and compared with the feature selection methods output ( $z_0$ ) of the claen of noise ,balanced dataset (Dataset_1) using the equation (3).

Worth noting that the stability metric produces a value in the range [0, 1]; the low value indicates the feature selection method provide unstable outcomes, whereas the high value indicates that the method has stable outcomes.

### *E. Feature Selection Methods*

As mentioned in Our Contribution section, another contribution of this study is to explore the behaviour of a combination from 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 univariate feature selection methods which are ANOVA(F-test)[53] and Mutual Information (MI)[54]. From Wrapper Methods Recursive Feature Elimination Cross-Validation (RFECV) with Support Vector Machine estimator[55] and Genetic Algorithm (GA) with Support Vector Machine estimator[56] are used, whereas in the Embedded Method Tree-Based feature selection[57] and LASSOCV [58] are used.

## THE RESULT

The following section will present the study experiment results of the feature selection methods stability and the selection performance across different difficulty levels. Each difficulty level has two scenarios according to the sample sizes (small sample size=100 and large sample size= 1000) associated with different data characteristics as mentioned in B section. Worth noting that since we know the relevant features thus, evaluating the feature selection methods using classification algorithms will be skipped. Instead, we will evaluate the feature selection method performance based on its ability to identify all the six relevant features correctly.

*- Level One &Two:*

Both levels showed almost identical feature selection and stability performance since they are categorised as an easy level problem according to the Complexity Metrix (N4) where the classes are not overlapping, and the decision boundaries are linearly separable. The experiment results indicated that most feature selection methods used in both levels have correctly identified all six relevant features across different characteristics in both scnarios small&large datasets see Fig. 3, Fig. 4, Fig. 5,Fig. 6,Fig. 7Fig. 8Fig. 9Fig. 10. However, the only method that failed to identify all six relevant features in both scenarios across different characteristics and produced unstable results are GA and LASSOCV. Except in the large sample size scenario LASSOCV showed a good performance in identifying all the six relevant features across both levels, see Fig. 7&Fig. 15.

In terms of the stability performance, the methods showed similar results to the feature selection performance where most of the method have produced stable results in both levels see Fig. 11,Fig. 12, Fig. 13,Fig. 14,Fig. 15,Fig. 16,Fig. 17,Fig. 18. Moreover, LASSOCV shares the same GA stability performance.

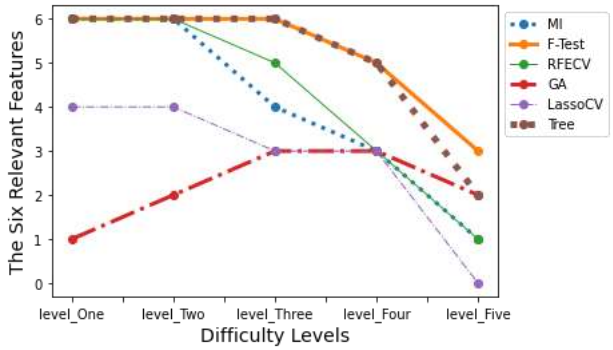


Fig. 3. The Selection Performance of the Feature Selection Methods in The Balanced Dataset (Sample Size=100 & Feature Size=50) Across Different Difficulty levels.

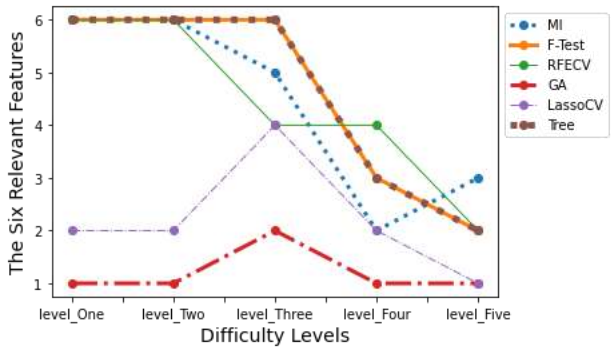


Fig. 4. The Selection Performance of the Feature Selection Methods in The Imbalanced Dataset (Sample Size=60 & Feature Size=50) Across Different Difficulty levels

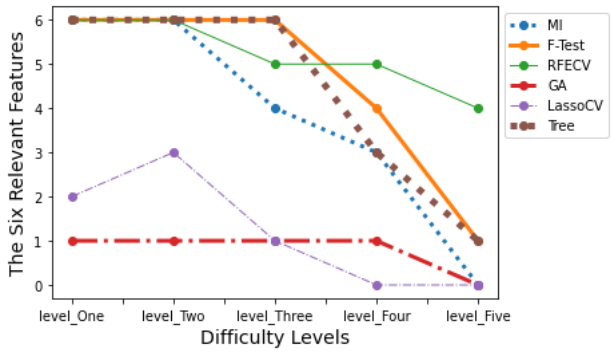


Fig. 5. The Selection Performance of Feature Selectiom Methods in Balanced Dataset (Sample Size=100 & Feature Size=1000) Across Different Difficulty levels.

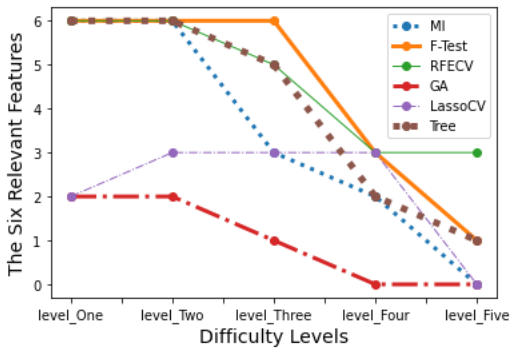


Fig. 6. The Selection Performance of Feature Selection Methods in Imbalanced Dataset (Sample Size=60 & Feature Size=1000) Across Different Difficulty levels.

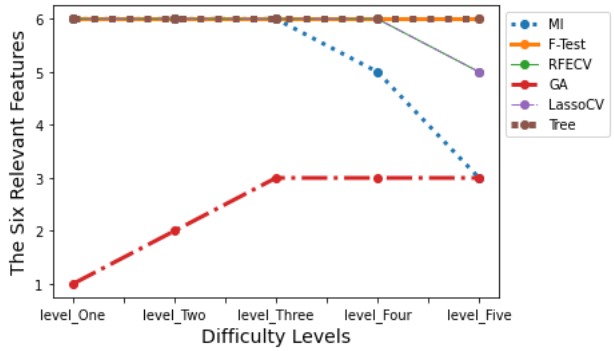


Fig. 7. The Selection Performance of Feature Selection Methods in Balanced Dataset (Sample Size=1000 & Feature Size=50 ) Across Different Difficulty levels.

- *Level Three:*

Based on the Complexity Metrix (N4), this level is considered a medium-level difficulty where the classes are partially overlapped in decision boundary regions of the clusters/classes see Fig. 1&Fig. 2. Generally, the results indicated that the feature selection methods started to miss some of the relevant features and added the irrelevant ones. However, the methods showed different behaviour in both small and large sample sizes scenarios. In the small sample size scenario, most of the feature selection methods failed to identify all the six relevant features except in the case of the balanced dataset, where F-test and Tree-Based methods are the only methods that correctly identified all the six relevant features in both cases (50 &1000 features sizes) at this level, as shown inFig. 3&Fig. 5. In contast, most of the methods showed better performance in the large sample size scenarios (balanced case) in both cases (50 &1000 features sizes), where it have correctly identified all six relevant features at this level see Fig. 7&Fig. 9, except GA, which showed poor performance. In terms of the large sample size (imbalanced case), F-test, RFECV and Tree-Based methods are the only methods that have correctly identified all six relevant features in case of 1000 features dataset and F-test and Tree-Based methods in case of 50 features dataset see Fig. 8 & Fig. 10.

In terms of stability performance, it almost has samiler behaviour to the selection performance where most feature selection methods have produced stable results, specifically in the large size balanced datasets dataset except GA see the cases Fig. 15& Fig. 17. However, in the case of large size imbalanced datasets LASSOCV, MI, RFECV and GA showed unstable behaviour in case of 50 features dataset and LASSOCV, MI and GA in case of 1000 features datasets see Fig. 16&Fig. 18.On the other hand, with related to the small sample size balanced datasets scenarios the only methods that have stable outputs are F-test and Tree-Based methods in both cases (50 & 1000features) see Fig. 11& Fig. 13. For the small sample size imbalanced datasets, the F-test and Tree-Based are the only methods have produced stable results in case of 50 features and only F-test method in case of 1000 features dataset as shown in Fig. 12& Fig. 14.

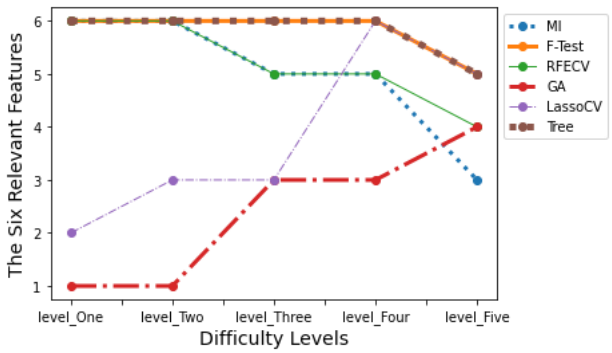


Fig. 8. The Selection Performance of Feature Selection Methods in Imbalanced Dataset (Sample Size=600 & Feature Size=50) Across Different Difficulty levels.

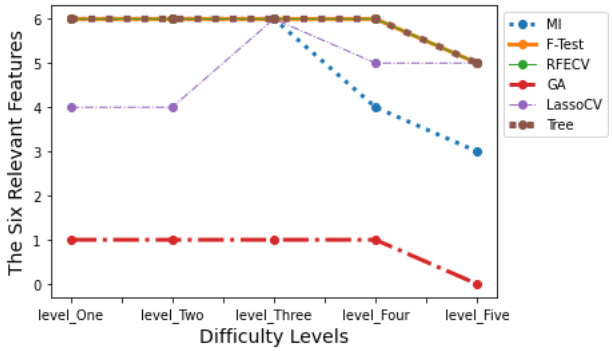


Fig. 9. The Selection Performance of Feature Selection Methods in Balanced Dataset (Sample Size=1000 & Feature Size=1000) Across Different Difficulty levels.

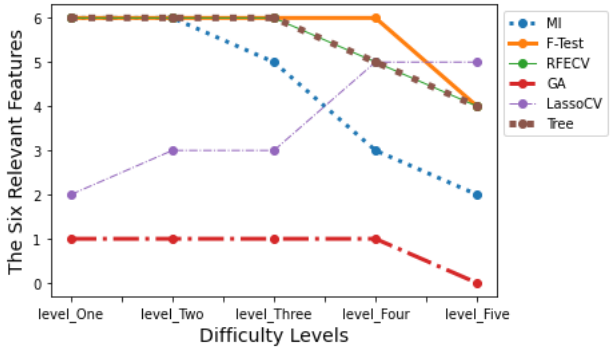


Fig. 10. The Selection Performance of Feature Selection Methods in Imbalanced Dataset (Sample Size=600 & Feature Size=1000) Across Different Difficulty levels.

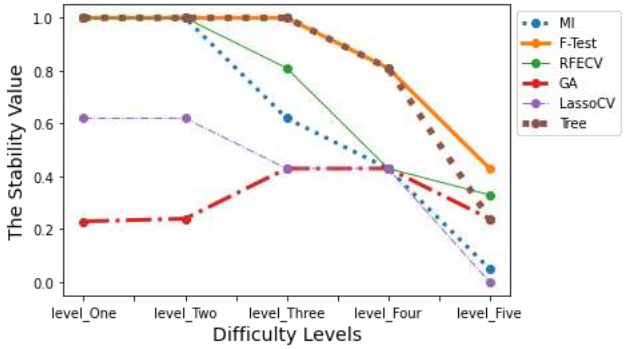


Fig. 11. The Stability Performance of the Feature Selection Methods in The Balanced Dataset (Sample Size=100 & Feature Size=50) Across Different Difficulty levels.

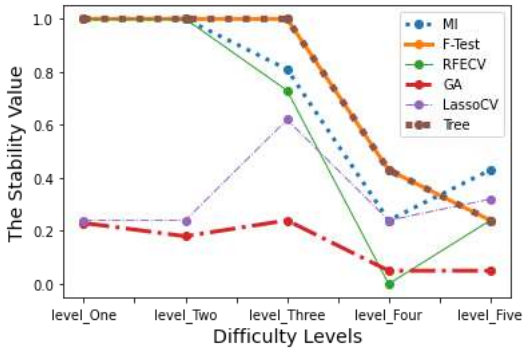


Fig. 12. The Stability Performance of the Feature Selection Methods in The Imbalanced Dataset (Sample Size=60 & Feature Size=50) Across Different Difficulty levels

*Level Four & Five:*

These levels are considered the most difficult levels based on the Complexity Metrix (N4). The results show that all the feature selection methods failed to identify all six relevant features specifically in all small samples size scenarios (balanced & imbalanced) see Fig. 3, Fig. 4, Fig. 5 and Fig. 6. However, in the case of large sample size balanced datasets scenarios, the only methods that correctly identified all relevant features are F-test and Tree-Based methods in case of 50 features (in both levels four & five) with LASSOCV and RFECV case of 50 features at level four only see Fig. 7. In the large sample size scenario imbalanced case, F-test, LASSOCV and Tree-Based methods are the only method that identified all the relevant features in case of 50 features at level four only see Fig. 8. Whereas in the case of large sample size scenario imbalanced case of the 1000 features, all the methods failed to identify all six relevant features except F-test at level four only see Fig. 10. In terms of the stability performance, the feature selection methods have produced unstable results across all the cases in both levels except the F-test and Tree-Based methods in large sample size balanced case in both levels four and five with RFECV and LASSOCV at level four only see Fig. 15, Fig. 16 and Fig. 17.

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 been affected by the existence of small sample size, high dimensionality, noise, and imbalanced classes when the classes are linearly separable (no classes overlapping) in the easy levels (level one & two). However, the methods started to misdiagnose some relevant features and added irrelevant ones when the classes started to overlap (level three), where they continued degrading in missing more relevant features and added irrelevant ones as they are moving to the upper level, especially in the difficult and very difficult levels (level four & five). Concerning the stability performance across different levels, it is likely to have similar selecting performance where the feature selection methods started to produce unstable output when the classes start to overlap.

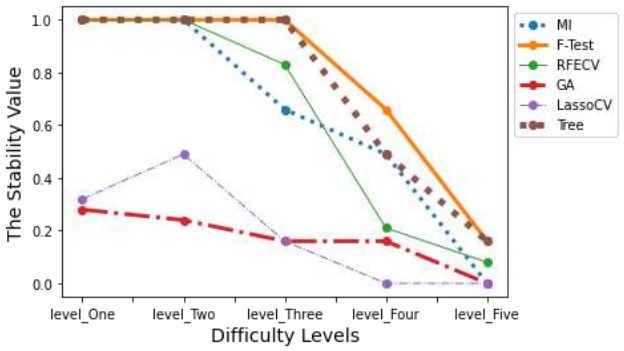


Fig. 13. The Stability Performance of Feature Selectiom Methods in Balanced Dataset (Sample Size=100 & Feature Size=1000) Across Different Difficulty levels.

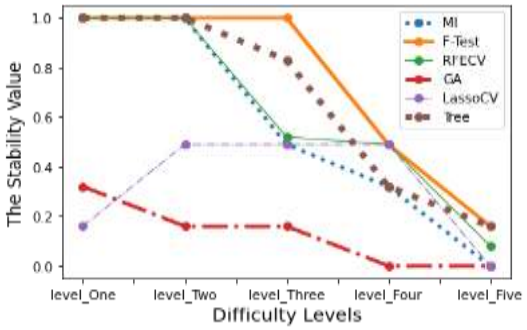


Fig. 14. The Stability Performance of Feature Selection Methods in Imbalanced Dataset (Sample Size=60 & Feature Size=1000) Across Different Difficulty levels.

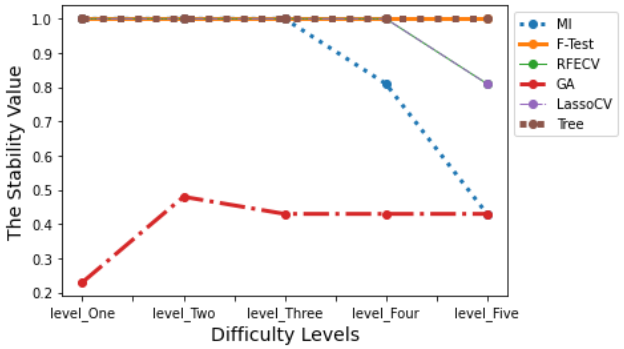


Fig. 15. The Stability Performance of Feature Selection Methods in Balanced Dataset (Sample Size=1000 & Feature Size=50 ) Across Different Difficulty levels.

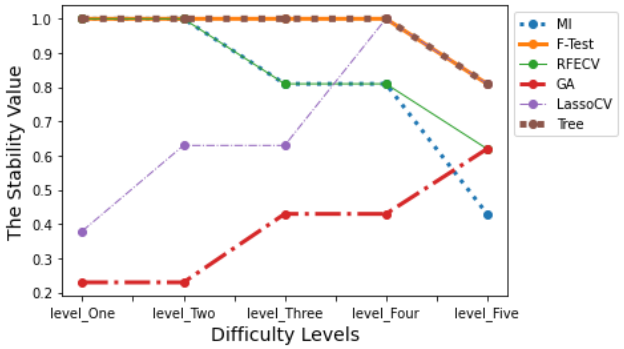

Fig. 16. The Stability Performance of Feature Selection Methods in Imbalanced Dataset (Sample Size=600 & Feature Size=50) Across Different Difficulty levels.

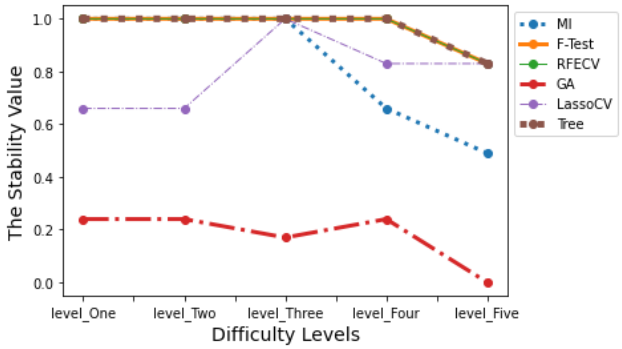


Fig. 17. The Stability Performance of Feature Selection Methods in Balanced Dataset (Sample Size=1000 & Feature Size=1000) Across Different Difficulty levels.

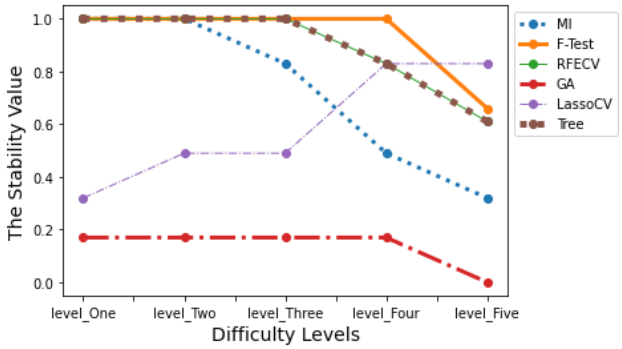


Fig. 18. The Stability Performance of Feature Selection Methods in Imbalanced Dataset (Sample Size=600 & Feature Size=1000) Across Different Difficulty levels

## CONCLUSION

This paper investigates several challenges - high dimensionality/ irrelevant features, small sample size, noise, imbalanced dataset, and class overlap – and their influence on the stability and accuracy of selection performance. The results showed that if the noise is 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 classes on the feature selection methods are relatively low. This outcome proved our assumption that the noise, high dimensionality, imbalanced classes issues are not necessarily imposing severe difficulties in the stability and accuracy of the selection performance if the classes are linearly separable. On the other hand, the interactive effects of the class overlap, and non-linear separability increase the chances of adverse effects on selection outcomes. Furthermore, this study showed that the small sample size and overlapping classes have a high impact on the feature selection performance. 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 results. 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 the stability and selection accuracy are the Tree and 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.

Future work will investigate the performance on further nonlinear relationships and how the feature selection performance can be improved so that real world data can be analysed with confidence.

## REFERENCES


[1] P. Mohana Chelvan and K. Perumal, "A comparative analysis of feature selection stability measures," 2017 International Conference on Trends in Electronics and Informatics (ICEI), 2017, pp. 124-128, doi: 10.1109/ICOEI.2017.8300901.

[2] Alelyani, S., 2021. Stable bagging feature selection on medical data. Journal of Big Data, 8(1).

[3] Salman, R., Alzaatreh, A., Sulieman, H., & Faisal, S. (2021). A Bootstrap Framework for Aggregating within and between Feature Selection Methods. Entropy, 23(2), 200. doi: 10.3390/e23020200.

[4] Colombellia, F., Woycinck Kowalskib, T. and Recamonde-Mendozaa, M., 2021. A Hybrid Ensemble Feature Selection Design for Candidate Biomarkers Discovery from Transcriptome Profiles.

[5] Ramezani, I, Niaki, MK, Dehghani, M & Rezapour, M 2020, '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

[6] Sechidis, K, Papangelou, K, Nogueira, S, Weatherall, J & Brown, G 2019, On The Stability of Feature Selection in the Presence of Feature Correlations. in European Conference on Machine Learning. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Wurzburg, Germany, 16/09/19.

[7] Pes, B. Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. Neural Comput & Applic 32, 5951–5973 (2020). https://doi.org/10.1007/s00521-019-04082-3.

[8] Han, Y., & Yu, L. (2012). A variance reduction framework for stable feature selection. Statistical Analysis And Data Mining: The ASA Data Science Journal, 5(5), 428-445. doi: 10.1002/sam.11152

[9] Gulgezen G., Cataltepe Z., Yu L. (2009) Stable and Accurate Feature Selection. In: Buntine W., Grobelnik M., Mladenić D., Shawe-Taylor J. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2009. Lecture Notes in Computer Science, vol 5781. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04180-8_47

[10] W. Awada, T. M. Khoshgoftaar, D. Dittman, R. Wald and A. Napolitano, "A review of the stability of feature selection techniques for bioinformatics data," 2012 IEEE 13th International Conference on Information Reuse & Integration (IRI), 2012, pp. 356-363, doi: 10.1109/IRI.2012.6303031.

[11] Shanab, A. A. et al. “Impact of noise and data sampling on stability of feature ranking techniques for biological datasets.” 2012 IEEE 13th International Conference on Information Reuse & Integration (IRI) (2012): 415-422.

[12] V. H. Barella, L. P. F. Garcia, M. P. de Souto, A. C. Lorena and A. de Carvalho, "Data Complexity Measures for Imbalanced Classification Tasks," 2018 International Joint Conference on

Neural Networks (IJCNN), 2018, pp. 1-8, doi: 10.1109/IJCNN.2018.8489661.

[13] Barella, Victor & Garcia, Luís Paulo & de Souto, Marcilio & Lorena, Ana & de Carvalho, Andre. (2020). Assessing the Data Complexity of Imbalanced Datasets. Information Sciences. 553. 10.1016/j.ins.2020.12.006.

[14] Pascual-Triana, José & Charte, David & Andrés Arroyo, Marta & Fernández, Alberto & Herrera, Francisco. (2021). 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. 63. 10.1007/s10115-021-01577-1.

[15] Fu, GH., Wu, YJ., Zong, MJ. *et al*. Hellinger distance-based stable sparse feature selection for high-dimensional class-imbalanced data. *BMC Bioinformatics* **21,** 121 (2020). https://doi.org/10.1186/s12859-020-3411-3

[16] H. Wang, T. M. Khoshgoftaaar and Q. Liang, "Stability and Classification Performance of Feature Selection Techniques," 2011 10th International Conference on Machine Learning and Applications and Workshops, 2011, pp. 151-156, doi: 10.1109/ICMLA.2011.133.

[17] Lei Yu, Chris Ding, and Steven Loscalzo. 2008. Stable feature selection via dense feature groups. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08). Association for Computing Machinery, New York, NY, USA, 803–811.

[18] Li, Y., Si, J., Zhou, G., Huang, S. and Chen, S., 2015. FREL: A Stable Feature Selection Algorithm. IEEE Transactions on Neural Networks and Learning Systems, 26(7), pp.1388-1402.

[19] Nogueira, S., Sechidis, K. and Brown, G., 2018. On the Stability of Feature Selection Algorithms. Journal of Machine Learning Research,.

[20] Naik, A., Kuppili, V. and Edla, D., 2020. A new hybrid stability measure for feature selection. Applied Intelligence, 50(10), pp.3471-3486.

[21] Mungloo-Dilmohamud Z., Jaufeerally-Fakim Y., Peña-Reyes C. (2020) Stability of Feature Selection Methods: A Study of Metrics Across Different Gene Expression Datasets. In: Rojas I., Valenzuela O., Rojas F., Herrera L., Ortuño F. (eds) Bioinformatics and Biomedical Engineering. IWBBIO 2020. Lecture Notes in Computer Science, vol 12108. Springer, Cham. https://doi.org/10.1007/978-3-030-45385-5_59

[22] P, M. and K, P., 2017. On Feature Selection Algorithms and Feature Selection Stability Measures : A Comparative Analysis. International Journal of Computer Science and Information Technology, 9(3), pp.159-168.

[23] Liu Y., Diao X., Cao J., Zhang L. (2017) Evolutionary Algorithms' Feature Selection Stability Improvement System. In: He C., Mo H., Pan L., Zhao Y. (eds) Bio-inspired Computing: Theories and Applications. BIC-TA 2017. Communications in Computer and Information Science, vol 791. Springer, Singapore. https://doi.org/10.1007/978-981-10-7179-9_6

[24] Haury, A., Gestraud, P. and Vert, J., 2011. The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures. PLoS ONE, 6(12), p.e28210.

[25] H. Wang, T. M. Khoshgoftaar, R. Wald and A. Napolitano, "A novel dataset-similarity-aware approach for evaluating stability of software metric selection techniques," 2012 IEEE 13th International Conference on Information Reuse & Integration (IRI), 2012, pp. 1-8, doi: 10.1109/IRI.2012.6302983.

[26] Khaire, U., & Dhanalakshmi, R. (2019). Stability of feature selection algorithm: A review. Journal Of King Saud University - Computer And Information Sciences. doi: 10.1016/j.jksuci.2019.06.012

[27] Alelyani, S. and Liu, H., 2013. Supervised Low Rank Matrix Approximation for Stable Feature Selection. In: 2012 11th International Conference on Machine Learning and Applications. Boca Raton, FL, USA: IEEE.

[28] Li, Y., Li, T., & Liu, H. (2017). Recent advances in feature selection and its applications. Knowledge And Information Systems, 53(3), 551-577. doi: 10.1007/s10115-017-1059-8

[29] Lei Yu, Yue Han, & Berens, M. (2012). Stable Gene Selection from Microarray Data via Sample Weighting. IEEE/ACM Transactions On Computational Biology And Bioinformatics, 9(1), 262-272. doi: 10.1109/tcbb.2011.47

[30] Schleif, F., Hammer, B., & Villmann, T. (2007). Margin-based active learning for LVQ networks. Neurocomputing, 70(7-9), 1215-1224. doi: 10.1016/j.neucom.2006.10.149

[31] Seijo-Pardo B., Bolón-Canedo V., Porto-Díaz I., Alonso-Betanzos A. (2015) Ensemble Feature Selection for Rankings of Features. In: Rojas I., Joya G., Catala A. (eds) Advances in Computational Intelligence. IWANN 2015. Lecture Notes in Computer Science, vol 9095. Springer, Cham. https://doi.org/10.1007/978-3-319-19222-2_3

[32] A. Noureldien, N. and A. Mohammed, E., 2020. Measuring Success of Heterogeneous Ensemble Filter Feature Selection Models. International Journal of Recent Technology and Engineering, 8(6), pp.1153-1158.

[33] Saeys Y., Abeel T., Van de Peer Y. (2008) Robust Feature Selection Using Ensemble Feature Selection Techniques. In: Daelemans W., Goethals B., Morik K. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2008. Lecture Notes in Computer Science, vol 5212. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-87481-2_21

[34] D. J. Dittman, T. M. Khoshgoftaar, R. Wald and A. Napolitano, "Comparing Two New Gene Selection Ensemble Approaches with the Commonly-Used Approach," 2012 11th International Conference on Machine Learning and Applications, 2012, pp. 184-191, doi: 10.1109/ICMLA.2012.175.

[35] A. Wang, et al., "Stable and Accurate Feature Selection from Microarray Data with Ensembled Fast Correlation Based Filter," in 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Korea (South), 2020 pp. 2996-2998. doi: 10.1109/BIBM49941.2020.931353

[36] Bania, R. and Halder, A., 2021. R-HEFS: Rough set based heterogeneous ensemble feature selection method for medical data classification. Artificial Intelligence in Medicine, 114, p.102049.

[37] Batur Şahin, Canan & Diri, Banu. (2020). ROBUST FEATURE SELECTION AND CLASSIFICATION USING HEURISTIC ALGORITHMS BASED ON CORRELATION FEATURE GROUPS.

[38] Z. Shang and M. Li, "Feature Selection Based on Grouped Sorting," *2016 9th International Symposium on Computational Intelligence and Design (ISCID)*, 2016, pp. 451-454, doi: 10.1109/ISCID.2016.1111.

[39] Bahekar K.B., Gupta A.K. (2018) Artificial Immune Recognition System-Based Classification Technique. In: Tiwari B., Tiwari V., Das K., Mishra D., Bansal J. (eds) Proceedings of International Conference on Recent Advancement on Computer and Communication. Lecture Notes in Networks and Systems, vol 34. Springer, Singapore. https://doi.org/10.1007/978-981-10-8198-9_65

[40] Corrales, D., Ledezma, A., & Corrales, J. (2018). From Theory to Practice: A Data Quality Framework for Classification Tasks. Symmetry, 10(7), 248. doi: 10.3390/sym10070248

[41] Cho, H., & Lee, S. (2021). Data Quality Measures and Efficient Evaluation Algorithms for Large-Scale High-Dimensional Data. Applied Sciences, 11(2), 472. doi: 10.3390/app11020472

[42] Altidor, W., Khoshgoftaar, T. and Napolitano, A., 2012. Measuring stability of feature ranking techniques: a noise-based approach. International Journal of Business Intelligence and Data Mining, 7(1/2), p.80.

[43] S. Alelyani, H. Liu and L. Wang, "The Effect of the Characteristics of the Dataset on the Selection Stability," 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence, 2011, pp. 970-977, doi: 10.1109/ICTAI.2011.167.

[44] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

[45] Lorena, A., Garcia, L., Lehmann, J., Souto, M., & Ho, T. (2019). How Complex Is Your Classification Problem?. ACM Computing Surveys, 52(5), 1-34. doi: 10.1145/3347711.

[46] Gudivada, Venkat & Apon, Amy & Ding, Junhua. (2017). Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations. International Journal on Advances in Software. 10. 1-20.

[47] Cano, José. (2013). Analysis of data complexity measures for classification. Expert Systems with Applications. 40. 4820–4831. 10.1016/j.eswa.2013.02.025.

[48] P. R. Peebles Jr., “Central Limit Theorem” in “Probability, Random Variables and Random Signal Principles”, 4th ed., 2001, pp. 51, 51, 125.

[49] Numpy.org. 2021. numpy.random.randn — NumPy v1.21 Manual. [online]Availableat:<https://numpy.org/doc/stable/reference/random/generated/numpy.random.randn.html#numpy.random.randn> [Accessed 9 September 2021].

[50] Imbalanced-learn.org. 2021. *How to use sampling_strategy in imbalancedlearnVersion0.8.0*.[online]Availableat:<https://imbalancedlearn.org/stable/auto_examples/api/plot_sampling_strategy_usage.html#sphx-glr-auto-examples-api-plot-sampling-strategy-usage-py> [Accessed 9 September 2021].

[51] Hoekstra, A., Duin, R.P.W.: On the non- linearity of pattern classifiers, In: Proc. 13th. Int. Conference on Pattern Recognition, Vienna, Austria (1996) 271-275.

[52] T. R. Fraça, P. B. C. Miranda, R. B. C. Prudêncio, A. C. Lorenaz and A. C. A. Nascimento, "A Many-Objective optimization Approach for Complexity-based Data set Generation," *2020 IEEE Congress on Evolutionary Computation (CEC)*, 2020, pp. 1-8, doi: 10.1109/CEC48606.2020.9185543.

[53] Kuhn, M. and Johnson, K., 2020. *Feature Engineering and Selection A Practical Approach for Predictive Models*. 1st ed. 2021: Chapman and Hall/CRC, p.242.

[54] Ross, B., 2014. Mutual Information between Discrete and Continuous Data Sets. *PLoS ONE*, 9(2), p.e87357.

[55] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V., “Gene selection for cancer classification using support vector machines”, Mach. Learn., 46(1-3), 389–422, 2002.

[56] Li Zhuo, Jing Zheng, Xia Li, Fang Wang, Bin Ai, and Junping Qian "A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine", Proc. SPIE 7147, Geoinformatics 2008 and Joint Conference on GIS and Built Environment: Classification of Remote Sensing Images, 71471J (7 November 2008); https://doi.org/10.1117/12.813256

[57] P. Geurts, D. Ernst., and L. Wehenkel, “Extremely randomized trees”, Machine Learning, 63(1), 3-42, 2006.

[58] Kim, S., Koh, K., Lustig, M., Boyd, S. and Gorinevsky, D., 2007. An Interior-Point Method for Large-Scale -Regularized Least Squares. *IEEE Journal of Selected Topics in Signal Processing*, 1(4), pp.606-617.