T.C.

ALTINBAS UNIVERSITY

Institute of Graduate Studies

Electrical and Computer Engineering

UNDER-SAMPLING MODELS TO IMPROVE CLASSIFICATION OF RARE CLASS IN IMBALANCED DATASETS

Zina Al SHAMAA

Doctor of Philosophy

Supervisor

Asst. Prof. Sefer KURNAZ

Istanbul, 2020

UNDER-SAMPLING MODELS TO IMPROVE CLASSIFICATION OF RARE CLASS IN IMBALANCED DATASETS

by

Zina Zuhair Raof Al SHAMAA

Electrical and Computer Engineering

Submitted to the Institute of Graduate Studies

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

ALTINBAŞ UNIVERSITY

2020

The thesis titled “UNDER-SAMPLING MODELS TO IMPROVE LASSIFICATION OF RARE CLASS IN IMBALANCED DATASETS” prepared and presented by “Zina AL SHAMAA” was accepted as a Doctor of Philosophy Thesis in Electrical and Computer Engineering.

Asst. Prof. Dr. Sefer KURNAZ

Supervisor

Thesis Defense Jury Members:

Prof. Dr. Osman Nuri UÇAN.	School of Engineering and Natural Sciences, Altinbas University	____________
Asst. Prof. Dr. Sefer KURNAZ	School of Engineering and Natural Sciences, Altinbas University	____________
Asst. Prof. Dr. Oğuz KARAN	School of Engineering and Natural Sciences, Altinbas University	____________
Prof. Dr. Ahmet Mesut RAZBONYALI	Engineering and Natural Sciences, Maltepe University	____________
Prof. Dr. Adem KARAHOCA	Faculty of Engineering, MEF university	____________

I certify that this thesis satisfies all the requirements as a thesis for the degree of Doctor of Philosophy.

Approval Date of Institute of Graduate Studies: 17/12/2020

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Zina Zuhair Raof AL SHAMAA

DEDICATION

To my wonderful parents and family.

ACKNOWLEDGEMENTS

First and foremost, I thank and pray to God who gave me the health and the strength to do this research. Then, I would like to express my appreciation to the Iraqi ministry of education for awarding me the opportunity to pursue my postgraduate education and make different in my life. My deepest gratitude goes to my academic supervisor, Asst. Prof. Dr Sefer Kurnaz and Asst. Prof. Dr Adil Deniz Duru for their continuous encouragements, and advice during all stages of my work. Special thanks are extended to my brother, Dr Zaed Z.R. Hamady a Consultant surgeon from Southampton University Hospital, who provided me with the colorectal cancer dataset, also for his continuous support. Further, I am deeply indebted to thank my parents for their endless love and for being my first and greatest teachers. My deepest gratitude goes out to my husband, Fakhry A.Salih for his patience, understanding and for unconditional support during a hard time of my study period. Also, my sons Mahmood, Hassan and daughter Mariam have been my biggest source of strength in these tough years. I'd like to let them know that in my opinion, they are the most important achievement of my whole life. Last but not least, I'd like to thank all from the bottom of my heart without my loving family, friends and supervisors, I wouldn't be where I am.

ABSTRACT

UNDER-SAMPLING MODELS TO IMPROVE CLASSIFICATION OF RARE CLASS IN IMBALANCED DATASET

AL SHAMAA, Zina,

PhD, Electrical and Computer Engineering, Altınbaş University,

Supervisor: Asst. Prof. Dr. Sefer KURNAZ

Date: December/ 2020

Pages: 82

In classifying the unbalanced data, the accurate predictions are influenced by the characteristics of data distribution in feature space. The unequal class and overlapping between classes are essential features which have an impact on the efficiency of the classification of minor class instances. These problems occur in various realistic applications such as anomaly detection, predicting students drop out of school, disease diagnosis, etc., which are of immense interest in knowledge discovery. This study presented two under-sampling techniques to improve the classification efficiency of the minor class which is of importance to many applications.

For the first objective of this research, we investigated a new method that handles the unequal distribution problem by under-sampling the major class size to minimize the classification biases toward the major class. This method is named the Hellinger distance under-sampling model (HDUS). By using Hellinger distance, the model calculates the similarity between each major class samples and its neighboring minor class samples, then selects the highest resemblance major instances considering to keep the minor class without change. Under-sampling the major class led to better discriminates the minor class instances. The experiments show that HDUS improve the classification performance of minor class by providing high sensitivity, F1-Measure and balanced accuracy. Results also indicate that HDUS can outperform state-of-the-arts under-sampling models.

For the second objective of this research, we present a model that handle both unbalanced distribution and overlapping problems by combining the proposed HDUS with ANOVA feature selection (HDUS+FS). This model has been built by employing HDUS model to remove the instances which identified as a noise; then employing ANOVA feature selection to eliminate the features that can indicate a high overlap in the boundary region. The experiment demonstrates the robustness of our proposed HDUS+FS model which outperforms feature selection alone and the state-of-the-arts combined with FS models.

Keywords: Imbalanced Classification, Unequal Class Distribution, Overlapping, Under-Sampling, Hellinger Distance, Feature Selection

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

LIST OF ABBREVIATIONS

KDD	:	Knowledge Discovery in Databases
IR	:	Imbalanced Ratio
ROS	:	Random Over-Sampling
SMOT	:	Synthetic Minority Over-sampling Technique
RUS	:	Random Under-Sampling
TL	:	Tomek Link
ENN	:	Edited Nearest Neighbor
HD	:	Hellinger Distance
HDUS	:	Hellinger Distance Under-Sampling
FS	:	Feature Selection
KNN	:	K Nearest Neighbor
SVM	:	Support Vector Machine
DT	:	Decision Tree
MP	:	Minority Percentage
TPR	:	True Positive Rate
TNR	:	True Negative Rate
PPV	:	Positive Predictive Value
F1-M	:	F1- Measure
B-AC	:	Balanced ACcuracy

CRCD	:	Colorectal Cancer Dataset
PIMAD	:	PIMA Dataset
THSD	:	Thoracic Surgery Dataset
BCD	:	Breast Cancer dataset

1. INTRODUCTION

1.1 OVERVIEW

Classification is a common method used in data mining problems. It is conducted when an object needs to be classified into a predefined class based on the attribute of those objects [1]. Most classification algorithms were mainly built to classify the balanced dataset. While most datasets in real-life applications are imbalanced in nature. That is the distribution of class instances in the feature space is not uniform, where the class with a greater number of instances is the majority class and the class with fewer instances is the minority class (rare class) [2], [3].

The classification of unequal class distribution is one of the most complex problems in the data mining field and has recently attracted a great deal of interest. The conventional classifier algorithms tends to favor the majority class while failing to classify the rare class correctly. This results in high overall accuracy of predicting majority class but bad sensitivity towards classifying the minority class [4]. Thus, an optimal goal is to focus on optimizing the sensitivity of the abnormal class rather than relying on the overall accuracy[5].

Another important problem in the unbalanced classification is the overlapping between classes [6]. This fact is related to the features of the dataset and defined as the data points in different classes that interfere with identical data points in the feature space. It is well-known that a big number of attributes may weaken the identification of boundary region of the rare class [7] because some of these attributes may be redundant or noise or there is no strong correlation between them [8]. That will effect on the accurate identification of minority class, Therefore, it is important to remove the noise, redundant or irrelevant features [9].

Recently, the unbalanced problems have received much attention, since the unbalanced class datasets are contained in several realistic applications such as anomaly detection[10], predicting students drop out of school[11], bankruptcy prediction[12], Face recognition[13], text classification [14], fraudulent credit card transaction[15], disease diagnosis[16] and many others. Due to the rise in the number of such applications, learning in the existence of unbalanced data is becoming a vital subject of research. Previous studies have been presented with many techniques to solve class imbalance and overlapping problems. Pre-processing techniques are commonly used. These techniques mainly work on data level and categorized into Over-sampling, under-sampling and feature methods [17]–[19].

1.2 RELATED WORK

1.2.1 The Classification With Imbalanced Class

The unbalanced classification is a critical topic that needs an efficient solution for a variety of problems. These problems present challenges in the development of effective classification techniques because the traditional classification algorithms were not initially built to train the unbalanced datasets [20]-[24]. Such classification algorithms are designed to optimize the overall prediction accuracy. When working with the unbalanced datasets, the traditional approaches are strongly biased toward the majority class and tend to classify the majority class accurately but mostly misclassify the minority class which is the class of importance in many applications. Thus these approaches will lead to very low identification to the minor class when applied directly to the unbalanced datasets. Therefore, if the problems of unbalanced class distribution are not resolved before to implement the classification process, it will be weakened the identification power of the classifier to recognize the rare class correctly, and the model may be over-fitted with so many training cases from the majority class[25],[26].

Previous research experiments in[27], [28] showed that a greater number of data mining algorithms are based on the idea that training datasets are well balanced, which is frequently inaccurate. As stated in the previous section, there are several datasets in a real-life application having unbalanced class distribution. Many classification algorithms may not consider the shape of the unequal class distribution and thus produce an incorrect representation model in the learning tasks. Such an inappropriate action would lead to a decline in the classification performance of the minority class samples. Therefore, new methods are needed to ensure that a classification algorithm can efficiently identify these most important rare class samples [29]–[34].

Recently, the amount of data have increased to a higher level due to sophisticated computing technology. It is noticed that data display a larger size than before with an increasing number of features. Imbalanced distribution databases with a large number of features needed an efficient feature selection techniques to determine the consistency of the features, as it is well recognized that unnecessary, duplicate and noise existence in the feature space can degrade the efficiency of the classification algorithms [2],[19], [35]–[39].

1.2.2 Characteristics of Unbalanced Class Distribution

The imbalanced class exists because the quantity of one class outnumbered the quantity of the other class, which involves a variety of learning problem. In the classification of the unbalanced data, the right predictions are affected by the characteristics of the unequal distribution of data points in the feature space[40]–[42]. These characteristics are as follows:

1.2.2.1 The unbalance class distribution

The distribution of unbalanced class is calculated by the imbalanced ratio (IR) that is expressed as the number of major samples over the number of minor samples. Based on the IR rating, the unbalanced datasets are split into three groups: low imbalanced (IR is 1.5 – 3), medium imbalanced (IR is 3 – 9) and large imbalanced (IR is greater than 9)[1]. The research in [43] investigated the relationship between the IR in the training dataset and the classification outcomes utilizing decision tree classifier. However, the study in [40] stated that there is still no exact value of IR that will begin to a decline in the performance of the classification. The research in [44] showed that the equilibrium distribution between classes is not an assurance of enhancing the classifier performance because the equal class ratio is not always the appropriate distribution to learn from. This indicates that not only the unbalanced distribution effect on the classifier but other factors also have an impact. Figure 1.1 displays two various distributions under the same IR.

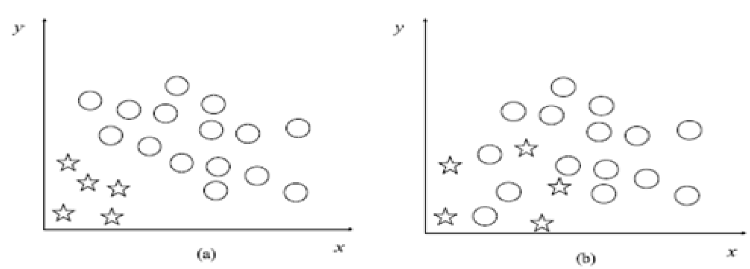


Figure 1.1: Two distinct distributions with the same IR. a) Simple case. b) Complex case.

1.2.2.2 Small sample sizes

Another difficulty to classify unbalanced data is the insufficient data due to the limited sample size in the training set. The insufficient number of instances may cause hardness to detect pattern uniformity particularly in the training dataset. The paper [23], showed that if the size of training instances grows, then the failure rate of the unbalanced class classification decreases. Similarly, the study [44], found similar results, There findings demonstrated that the classifier can build a stronger representation of classes with a more accessible training dataset

1.2.2.3 Overlapping

One of the important challenges with the unbalanced data classification is the appearance of overlapping in the dataset. Class overlapping refers to the data points in different classes which interfere with identical data points in the dataset [45]. In other meaning, it refers to how separable the classes in the dataset (as shown in figure 1.2.) The challenge is in distinguishing the minority class from the majority class, which have difficulty in distinguishing between various classes accurately, especially the smaller class[46]. Many studies [8], [47]–[51] have investigated the separation of classes and proved that overlapping of classes is a significant barrier to classify classes correctly, where traditional classifiers usually identify overlapping points as relating to the major class whereas considering the minor class as noise.

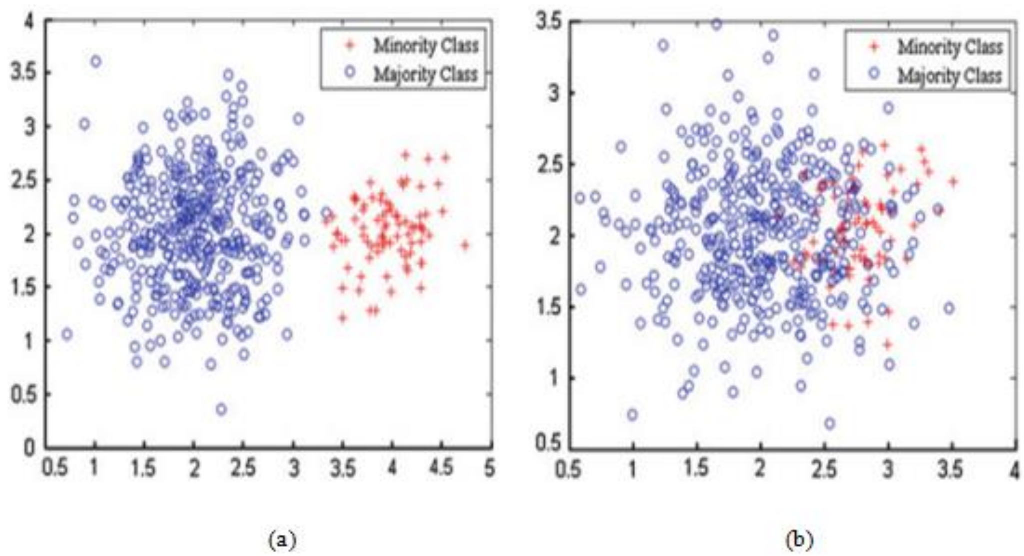


Figure 1.2: a) Example of data without class overlapping, b) Example of data with class overlapping.

1.2.2.4 Small disjuncts

It denotes to the few instances of the minor class which have been distributed separately in the space[52], [53] (as shown in figure 1.3). The study [54] investigated the class imbalanced and small disjuncts problems and suggested a method based up-sampling to improve classifying the minority class that caused by small disjuncts. The study in [40] stated that the small disjuncts in imbalanced class impact the efficiency of the classification due to 1) It challenges the classifier in the idea of minor class, 2) the appearance of the class problem is unconscious.

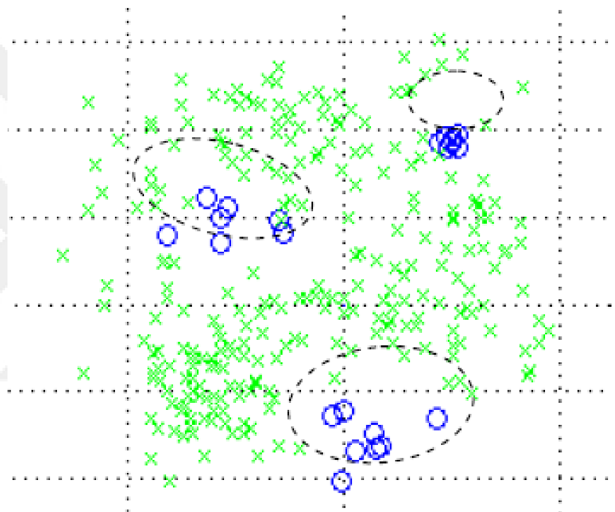


Figure 1.3: An example of small disjuncts.

1.2.3 Approaches to Handle Unbalanced Class Classification Problems

To solve the imbalanced classification problems, many methods have been introduced. These approaches are based on two main strategies: 1) algorithm level, 2) data level [6],[31], [55]–[58].

1.2.3.1 Algorithm Level

At the algorithm level approach, existing classification models are customized to address the issue of class imbalance between the minor and major classes [59]. In general, the algorithm level can be characterized as dedicated algorithms, Cost-sensitive, and Ensemble learning [60].

***a) Dedicated Algorithms*:**

In dedicated algorithms method, the classification algorithms are adjusted to suit the need for learning directly from the unbalanced class distribution. These algorithms can be learned about the spread of class imbalances prior to the extraction of essential information to improve a model according to the desired goal [61],[62]. There are many studies in the literature on improving the classifiers algorithm to handle the imbalanced class, such as improving support vector machine [63]–[66], improving Decision tree [67] and improving K nearest neighbor [68]–[71].

b) Cost-Sensitive:

The other algorithm level method is Cost-sensitive which works by defining various weights for different class cases [5], [72]. This method is built around the assumption that a high cost is forced on a classifier as misclassification occurs, e.g., a classifier applies higher costs to false negatives relative to false positives, thereby revealing accurate classification or misclassification the positive class. Many studies have investigated on the cost-sensitive method to improve classifying the imbalanced class distribution such as in [73] which introduced a cost-sensitive deep-confidence network to optimize the misclassification cost in an imbalanced dataset, [74]suggested combining pattern discovery with the cost-sensitive method, and [75] whose work designed coast-sensitive matrix with parallel learning framework. However, it is claimed that in most situations the exact cost is not known even in the situation of balanced distribution data[76], and the work [77] has shown that cost-sensitive learning can induce the over-fitting problems in preparation.

c) Ensemble Learning:

Another approach at algorithm level is Ensemble learning, which works by combining a variety of simple classifiers and then implements some collective strategies to integrate them to improve the efficiency of classifying imbalanced dataset [21], [78]–[82]. In general, Adaboost, bagging, voting and random forest are commonly used ensemble learning algorithms. Many studies have been used ensemble learning to fix the unbalanced class problems such as RUBoost [25] used boosting to address imbalanced problem, [83] utilized the voting strategy to enhance the weighted vector thereby optimize the overall classification performance of unbalanced medical dataset, and [84] used random forest (RF) to assign individual weights to each class rather than a single weight. Even though, ensemble techniques are often more flexible when compared to dedicated algorithms and cost-sensitive methods due to their isolation from the base classifier, but it still difficult for them to properly classify the minority class[85].

1.2.3.2 Data Level

The imbalanced classification problems can also be solved at data level through pre-processing techniques by changing or modifying the propensity of the class distribution on the data set, which is simple and efficient for unbalanced classification [1], [2], [86], [87]. The pre-processing step can be achieved by either sampling methods (using over-sampling or under-sampling) or feature selection technique to decrease the imbalanced rate of training data[3], [88].

***a) Over-Sampling Technique*:**

Over-sampling technique aims to generate instances for minor class by inserting copies of existing data from rare class samples [18], [89]–[92]. Many over-sampling techniques have been used earlier. One of the popular over-sample technique is Random over-sampling (ROS) which arbitrarily inserts samples to the minority class. Despite the fact that ROS changes the distribution of the class, it can raise the problem of over-fitting by creating identical copies of the minority class which impact the classification procedure [93]. Another traditional over-sample model is the synthetic minority over-sampling technique (SMOT) [94]. It is employed to produce artificial instances. Unlike ROS, SMOT avoided an over-fitting issue, but it can induce overlaps with the surrounding instances which maximize the total size of the training data and obstruct the training procedure [95], [96]. In general, the problem of the unbalanced class is reduced with the over-sampling method, but the training data is going to get overcrowded. Thus, the performance of the classification is affected [97].

b) Under-Sampling Technique:

Under-sampling is another rational method for sampling data that tries to minimize the number of instances in the major class. The idea of the under-sample approach is how to reduce the major class samples in a way that preserves the functional distinction amongst classes [19],[98],[99]. Various under-sample techniques have been applied and utilized earlier. Random under-sampling (RUS) is used to balance the distribution of classes by removing instances from major class randomly. This technique has been criticized in some studies as it may waste valuable knowledge which can be essential for the classification process [93], [100]–[102]. Another under-sample model is Tomek Link (TL), which utilized to tackle the overlap problems. It scans for pairs of instances belong to various classes but is the closest neighbor to each other and excludes the major instance of the pair [103],[104]. One other model is Edited nearest neighbor (ENN). It is used to eliminate majority class instances depending on the nearest K neighbor which belongs to the minority instances. If the number of neighbors in each majority class is larger in the minority class, the majority class instance may be excluded as borderline or noise [46].

Although several attempts have been made to handle class imbalanced problems by sampling methods, it should be noticed that all these methods make it possible to sampling to any required ratio and it is not essential to exactly equal the number of minority and majority classes. Some studies tried to determine the best sampling ratio for various problem settings [105]–[108]. Another work in [43] reported that there has been no optimum rule to obtain the best performance with under or over-sampling.

Some studies preferred to under-sample the major class for several reasons; As oversampling relies on the addition of a new copy of minority class samples already in the data set, it appears to over-fit of the data and it doesn't significantly improve minority class recognition[94], [109]. Besides, as the size of the data increase, the under-sampling model will be a better choice than the over-sampling model [110]. Also, authors in [55] discouraged to use over-sampling because the down-sampling model provides a fair sensitive value to adjustments in the cost of distribution and misclassification of classes, while over-sampling demonstrates little sensitivity, there are often very few variations in results when misclassification cost are adjusted.

In previous studies, the selection of instances was utilized to eliminate the outlier from the training data, which enable the classifier to achieve better than the initial data [111]–[113].

Nevertheless, the current models of instance selection have been designed to pick a portion of the original data set that cannot be utilized directly to pick samples from just one class of the data-set, like the choice from the majority class samples. The work in [114] presented a one-side instance selection that would eliminate some instances from the major class whereas preserving the initial instances relating to the rare class. They characterized the instances relating to the rare class into four groups: overlap samples, redundant samples, boundary samples and safe samples.

A lot of under-sampling models have currently been documented in the research to enhance the unbalanced data classification. The study [52] presented an under-sampling model by clustering the major class into groups of identical data instances and then selecting the unrepresentative instances from each group. The paper [46] introduced an under-sampling method beginning with defining the nearest major class neighbors to each rare class and then calculating the number of correlated of each neighbor from the majority class with the minority class instances. Then, the desired number of majority class is extracted from the correlated samples. In [115] the researchers employed the one-side under-sampling method. They suggested a model to reduce the size of the majority class that would adjust the distribution of the original unbalanced classes by comparing the similarity of each majority class with the related minority class. The model effectively separates the majority and minority class samples to improve the identification value for each class. The study [116], proposed an overlap based under-sampling technique, named (URNS) to improve the prediction of minor class samples in unbalanced datasets. By iteratively examining the nearest samples to the majority samples in the overlap area. Then, eliminate these samples to increase the recognition of rare class samples. The model implemented on real medical datasets. The proposed URNS have given a good value of f-measure and high sensitive value to the rare class, which is very needed in the medical area. The results of URUS were discussed with other state-of-the-arts and gave good solutions. The authors concluded that the sampling rate does not depend on imbalanced rate but on the amount of overlapping. In [117] authors proposed an over-sampling method based on Hellinger distance (HD) technique by generating synthetic instances for the probability outlines of instances with more neighbors in the same class, in order to balance the dataset classes, so can improve the minority class accuracy.

c) Feature Selection:

In addition to over/ under-sampling techniques, the selection of features is another pre-processing technique used recently to solve imbalanced class classification problems [45], and it is becoming a hot topic although it has a short history. In general, the selection of features is considered as an essential process for several data mining models, particularly if the data is high-dimensional. This technique aims to pick up n features from the original data without modifying the original features that help the classifier to achieve optimum efficiency. The data quality can be optimized by removing noisy, redundant and irrelevant features by choosing an appropriate subset of attributes. Moreover, the computing time can be minimized if a subset of attributes is used rather than all the original one. Since the topic of the imbalanced class is usually accompanied by the high-dimensionality problem [39], [118]–[120] implementing a feature selection technique is an essential course to take.

Recently, studies have investigated using feature selection [121]–[123] as a method to handle the imbalanced data problems. The reason is that feature selection techniques can move the emphasis to the features that enhance the distinction between classes instead of training instances [2],[124]. It implies distinguishing what another subset of attributes is more relevant to the information of the rare class. Comparing to the progress made by sampling techniques, substantially fewer studies developed feature selection approaches especially to handle the unbalanced class problems. The study [125] proposed two methodologies for feature selection to fix imbalanced binary class. The first method separated the major class into comparatively balance pseudo-subclasses and then evaluate the performance of the feature with the decomposed data to select a subset of high-quality features. The second method utilized Hellinger-distance to calculate the variance of an attribute in the instances of various classes. The research in [126] suggested a density-based feature selection approach to address the problems of the imbalanced dataset with high dimensions. This approach was calculated based on the probability distribution of the attributes. It calculated the recognition power of the attribute within classes based on the distribution of the value of the attribute. The paper [127] suggested a class-oriented feature selection based on the principle of information entropy. It has been inspected for a correlated subset of an attribute for each class. The chosen attributes are more balanced than those achieved by the conventional method. Authors in [39] proposed a new feature selection using a correlation metric to tackle the minority class problem in an unbalanced dataset. In [128] suggested a novel correlation metric called CFS to calculate the importance of a feature-subset depending on the degree of class correlations. Another research in [7] presented an attribute selection based on a straight-forward technique to compute the relevance of a feature depends on the average rate of the minor class. So a classifier is considered to be relevant if the minor class average rate is alike or equal to two standard deviations of the minor class rate. The work provides an efficient attribute selection model, named the neighborhood relationship, which maintains the value for multi-label classification.

In general, attribute selection methods are divided into four categories: filter, wrapper and embedded methods based on the strategies of search and evaluation measures. The filter technique is referred to models that used statistical measures to calculate the score of correlation among the individual attributes and the target class then select the highest attributes which applied to the predictor. This technique determines the correlation between variables depending on the intrinsic property of the data without relying on mining algorithms [129]–[131]. The most utilized filter techniques are Mutual information, Information gain, correlation-based feature selection, minimum redundancy maximum relevance, and chi-square [132]. The wrapper technique depends on the efficiency of the data mining algorithms as a criterion for evaluating good features or instances. In this approach, the predictor is wrapped on a search model for finding the features which achieve the best efficiency. The advantage of the wrapper method is to guarantee good results [133]–[135]. The third selection method is embedded. It is more dynamic than other methods because it embeds the search for the important features as part of the classifier construction [120], [130], [134], [136]. The work in [8], has experimented a backward removal method embedded with support vector machine to select specific attributes which are employed to distinguish amongst classes for an unbalanced dataset. The attribute selection may affect some classification models for achieving optimal efficiency. Besides, more meaningful and practical applications can be obtained after irrelevant attributes have been abandoned [120].

1.3 AIMS AND OBJECTIVES

This study aims to handle the skewed distribution and overlapping problems in imbalanced datasets to enhance the classification performance of rare class instances. The major contributions are listed here:

1- For the first objective, we present a new under-sampling model, called the Hellinger Distance Under-Sampling (HDUS), to address the issue of imbalanced class classification by employing the Hellinger distance to calculate the resemblance among each major class sample and its neighboring minor class samples, then selects a set of the highest similarity scores and adds them to obtain a resemblance score to every major sample. Lastly, the model chooses a set from the major samples with top resemblance scores and integrated with the original minor class samples. The HDUS model decreases the major class until rebalancing with the minor class to separate the two classes efficiently and increase the discrimination power for each class, thereby improving the classification sensitivity for the rare class which is the class of interest in many applications.

2- For the second objective, we introduce a hybrid under-sample and feature selection (FS) technique, named HDUS+FS, to handle both unequal class distribution and overlapping problems by combining our proposed HDUS model with ANOVA f-test feature selection in a cascade manner. The Under-sampling model will reduce the majority class, by removing the redundant majority samples having the highest similarity to rare class samples, to balance the instance distribution among the majority and minority classes. Feature selection will reduce the boundaries of the problem by restricting the effect of features which cause difficulties in the process of discrimination the minority class. The hybrid model will improve the classification performance for the important minority class.

3- A set of experiments are conducted on four unbalanced medical data sets using three classifiers. The performance of our two proposed methods are estimated through a standard metrics: Balanced accuracy, specificity, recall, precision and F1-Measure. We compared the HDUS model with a baseline model and three state-of-the-arts under-sampling techniques: Tomek Link (TL), random under-sampling (RUS) and edited nearest neighbor (ENN). And, we compared the HDUS+FS model with a baseline model, feature selection alone and the combination of the three state-of-the-arts under-sampling techniques with ANOVA+FS: TL+FS, RUS+FS and ENN+FS.

1.4 THE ORGANIZATION OF THE DISSERTATION

In this dissertation, we propose two models to enhance the classification performance of minority class samples in an unbalanced data sets. In chapter 2, we present the Preliminaries about basic classification algorithms, sampling methods and evaluation metrics; In chapter 3, we introduce the model named Hellinger distance under-sampling to tackle the imbalanced distribution problem; In chapter 4, we present a model that combine under-sampling and feature selection to address both unbalanced class and overlaps problems; In chapter 5 exhibit the conclusion and future work.

2. PRELIMINARIES: BASIC CLASSIFICATION ALGORITHMS, SAMPLING METHODS AND EVALUATION METRICS

In the imbalanced dataset, the class with a larger number of instances takes up most of the space. Unequal class distribution makes the classifier to be inadequately qualified to classify the smaller class instances, and the class with a larger number of instances overlaps the identification ability of the class with a smaller number of instances. In this case, the classifier would favor the majority class instances and scoring false high accuracy.

2.1 STANDARD CLASSIFIERS

In this study, we use three classification models with different characteristics: K Nearest Neighbor (KNN), Support Vector Machine (SVM) and Decision tree (DT), which are discussed in this section. The main reason to use these algorithms is to estimate the performance of the suggested models.

2.1.1 K Nearest Neighbor (KNN)

It is one of the simplest classification algorithms. It detects the unidentified data point using the previously defined data points depending on the nearest k training points in the multidimensional data space; each of the k neighbours has a class label and a label of a given point is specified by the major vote of the class labels of its k nearest neighbour. The successful choice of k relies on the data, although greater values of k usually minimize the influence of noise on the classification while making the boundary between classes fewer distinguished [137]. To identify new element x, the k items that are closest to it are specified. The class that utmost frequently occurs amongst these neighbouring elements is expected for x. While many classes meet for the greatest appearance, one of them is picked randomly.

The concept of closest is performed through a proper metric of distance or resemblance calculation. The closest neighbour to x is the k items having shortest distance or the greatest resemblance with x. The standard option of distance metric is the Euclidean distance specified as in equation (2.1)

$$d\_eucld(X,Y) = \sqrt{\sum_{i=1}^{n} |Xi - yi|^2} \tag{2.1 }$$

Where d_eucld is the Euclidean distance among x and y. When I variable minimal, the absolute value of the difference is not sufficient, since it is clearly not satisfied. In such a case, the term in summation should replace by (Xi ≠ yi), where it is recalled that the indicator function corresponds to 1 if the condition satisfied and to 0 otherwise.

To classify of a new element, the value of k is indicated by the user. The distance between the unclassified point and each of the training class point is measured. The k nearest neighbour is then selected on the bases of the distance at which the class label of the unclassified point is defined as a maximum repeat class labels of the k nearest neighbour [138]. The KNN classifier is adjustable where the user can use various values of k and optionally use various distance metrics. An example of KNN is presented in figure (2.1), which using the 7-NN rule. The dotted circle indicates the area that includes the 7 nearest neighbour of the unidentified data points. The unknown red point is classified for the black class because 5 of the 7 closest neighbours are of black class and 2 are of white class.

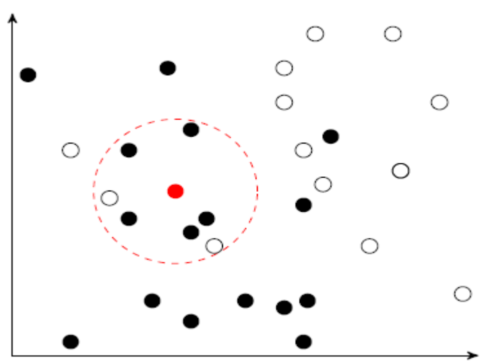


Figure 2.1: An example of KNN.

In the case of the unbalanced training dataset, instances of the minority class exist rarely in the data space. The limited relative existence of minority samples may lead the classifier to identify only a few samples as positives. The majority of its k neighbors will have to relate to the positive class before the sample itself is categorized as such, since it is more possible that the area includes a lot of negative items, precisely because they predominate the positive samples in the dataset [139],[140]. In the limiting case k=1, an item is wrongly classified when its closest neighbor relates to the opposite class. Generally, when contemplating the nature of class skewness and not the specific structure that may occur in a dataset, minority samples are also more likely to have the closest neighbor belonging to the opposite class, whereas negative instances are more frequently found in a dataset. Other nearest positive neighbors are not in a position to balance out the influence of such a negative neighbor.

2.1.2 Support Vector Machine (SVM)

SVM is a binary classifier that depends on the concept of the boundary. It has controlled power to separate the labels by hyper-plane. For a two-class classification problem, the separating hyper-plane is defined by the equation (2.2)

$$u(x) = w.x + b \tag{2.2}$$

Where w and b are the model parameters. Many possible hyper-planes could be chosen from equation (2.2). The main goal of SVM is to maximize the interval from the hyper-plane to the nearest data points. SVM aims to find an optimum hyper-plane separation that has a maximal margin. Optimizing the margin interval offers some strengthening so that expected data points are more precisely classified [141].

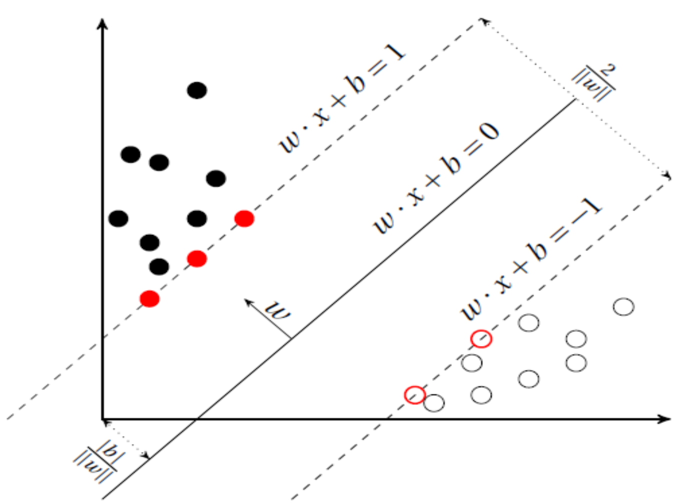


Figure 2.2: Illustration of a linearly separable two class problem with SVM. The source code used to make this figure is adapted from [142].

An example of linearly separable SVM for binary classification is presented in Figure (2.2). Let $\{x_i, y_i\}$, i=1, 2, 3,…., n, are the input/output cases in of the training dataset. suppose we have some separating hyper-plane (*Hp*), a point x that lies on (*Hp*) needs to satisfy the equation (2.2); where *w* is a weight vector that is normal to *Hp*. Then all observations from the training data need to fulfill the restrictions in the equations (2.3) and (2.4)

$$x_i.w + b \geq +1 \quad if \;\; y_i = +1, \tag{2.3}$$

$$x_i\,.w \;+\; b \leq -1 \quad if \;\; y_i = -1, \tag{2.4}$$

These may be merged into one set of inequality as in equation (2.5)

$$y_i\;(\,x_i.w + b) \geq 1, \tag{2.5}$$

Where i=1,2,….n,

Then the hyper-planes Hp1 and Hp2 are defined in equations (2.6) and (2.7), respectively.

$$x_i.w + b = \;+1, \tag{2.6}$$

$$x_i.w + b = -1, \tag{2.7}$$

So the points on *Hp1* and *Hp2* are the support vectors. Notice that *Hp1* and *Hp2* have similar normal vector so they are parallel to each other. Also, there should be no training points between them. Now, from equations (2.6) and (2.7), the distances from the origin to *Hp1* and *Hp2* are defined in equations (2.8) and (2.9), respectively.

$$\frac{|1 - \mathrm{b}|}{||\mathrm{w}||} \tag{2.8}$$

$$\frac{|-1 - \mathrm{b}|}{||\mathrm{w}||} \tag{2.9}$$

So, the margin between *Hp1*and *Hp2* is $\frac{2}{||w||}$. To increase the margin between *Hp1* and *Hp2,* ||w|| must be decreased. Integrating this with the condition in equation (2.5) we have a problem with optimization such as in equations (2.10) and (2.11)

$$min \ \frac{1}{2}||w||^2 \ , \tag{2.10}$$

$$y_i \ (\ x_i \ w + b) \geq 1, \tag{2.11}$$

Where i=1,2,…..n. Because the norm ||w|| has a square root, that renders optimization hard, we have change ||w|| with $\frac{||w||^2}{2}$. We now have a quadratic program that can be handled out using Lagrangian multipliers, as indicated by [143], [144]. The Lagrangian for this problem is as in equation (2.12)

$$L = \frac{1}{2}||w||^2 - \sum_{i=1}^{l} \alpha_i \, [y_i(x_i.w + b) -], \tag{2.12}$$

However, when the data is intrinsically nonlinear, it is hard to separate data points in the original input space. Fortunately, there is a simple technique that makes the linear SVM work well with the non-linear case. The idea relies on the kernel trick, which allows building a separating hyper-plane that converts data from the original dimension into a high dimensional space where the training set can be linearly separable[145],[146]. Figure (2.3) shows the linear reparability of kernel trick.

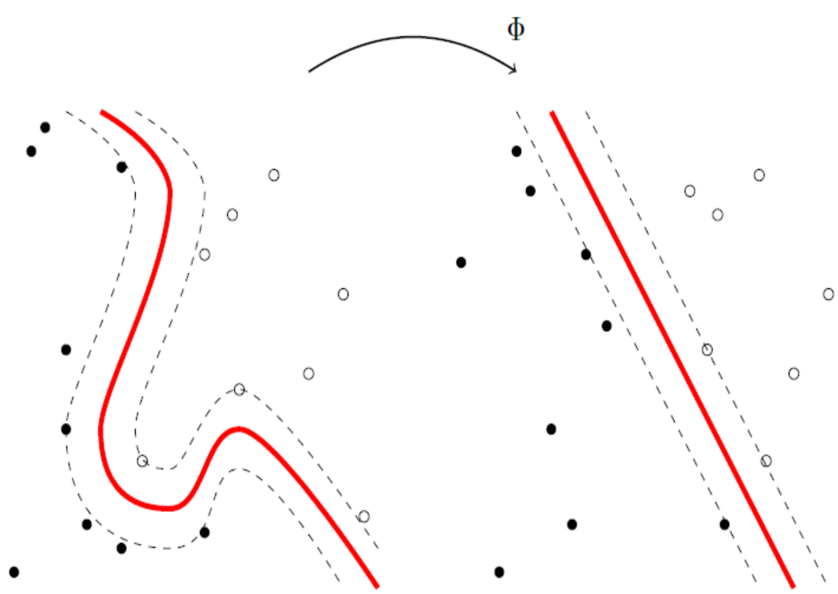
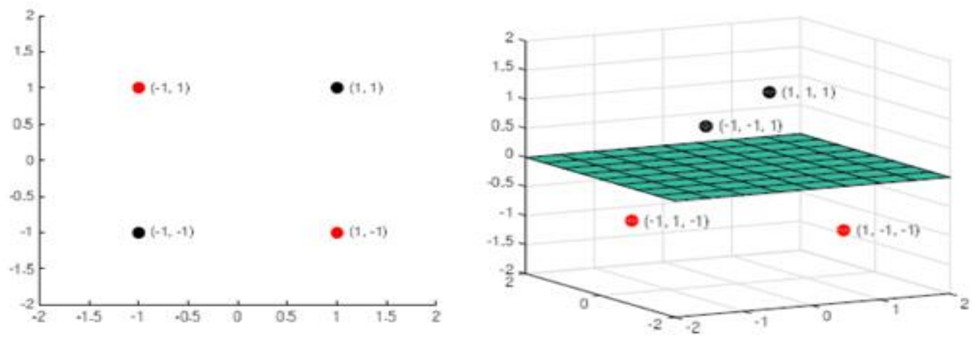


Figure 2.3: Kernel maps a non-linearly separable data into a high dimension linearly separable problem. The source code utilized to create this figure is edited from [147].

To see how the mapping to a high dimension feature space provides linear separation, considering the example in figure 2.4. The points of red class and black class in figure 4a can never be divided linearly in the original space R^2. Then after mapping to R^3, the two classes are separated by the *XY* hyper-plane as presented in figure 4b.



a) Input space R^2 b) feature space R^3

Figure 2.4: An example of the kernel trick, a) Input space R2 ,and b) feature space R3.

Various kernel functions are often used in practice such as linear, polynomial, Gaussian and sigmoid which defined as in the equations (2.13), (2.14), (2.15),and (2.16), respectively:

Linear kernel $$K(x_i, x_j) = x_i . x_j, \quad (2.13)$$

Polynomial kernel $$K(x_i, x_j) = (r + \gamma x_i . x_j)^p \quad (2.14)$$

Gaussian kernel $$K(x_i, x_j) = exp\left(-\gamma \left\|x_i - x_j\right\|^2\right), \quad (2.15)$$

Sigmoid kernel $$K(x_i, x_j) = tanh(r + \gamma x_i . x_j) \quad (2.16)$$

SVM is assumed to be a little sensitive to unbalanced class problems than other classifiers, as the class borders are determined concerning just a few supporting vectors and the size of class does not influence the boundary very much. However, many studies have investigated how to enhance the efficiency of SVM for unbalanced datasets. Research works in [148],[149] suggested using SVM in combination with SMOT to resolve the imbalanced problem. In [150] authors have investigated experiments on SVMs with two datasets of different imbalanced ratio. They found that as the training data become more unbalanced, the rate of support vector among the majority and minority classes also became more unbalanced. This is because the classifier can encounter a lot of major samples near the border.

2.1.3 Decision Tree (DT)

It is a widely used classifier since it has a good performance as opposed to other data mining algorithms[67]. It contains a variety of simplified decisions to create a tree form as connected cyclic graph. It is made up of three categories of nodes: root, inner and leaves. The root denotes a beginning that has no incoming but has outgoing edges. The inner nodes are represented by the attributes that have only one incoming branch and at least two outgoing branches for each available attributes. While the leaves are denoted by the class labels. These shapes are expressed as if-then rules that can be used to identify new samples [151].

In general, a DT model is constructed in two stages: building and pruning. In the building stage, the decision tree is constructed by iteratively dividing the training dataset based on the optimum criterion until all or most of data-records relate to each section have the same class label. After constructing the DT, a pruning stage is done to decrease the size of the tree. The tree which is too big is prone to over-fitting. To avoid over-fitting, pruning can take place to improve the power of the DT by pruning the branches of the original tree. The tree pruning method is based on a failure: starting at the base of the tree and testing each non-leaf sub-tree. If replacing this sub-tree by leave or by its most widely utilized branch will result in a smaller predicted failure ratio, then the tree pruned accordingly [40]. While constructing decision trees, the class sample marked as a leaf is detected by testing the training cases represented by the leaf and selecting the highest repeated class. In the existence of an imbalanced class, DT will have to run several checks to recognize minority from majority classes [40]. In some classification models, the splitting process is ended before finding the branches to predict the minority classes. In other classification approaches, the branches to predict the minority classes may be trimmed as prone to over-fitting. A correct way to predict a small number of instances in the minority class is almost ineffective in minimizing the error rate considerably, comparing to the error rate that increased by over-fitting. Although pruning is based on the prediction error, there is a strong likelihood that the branches to predict the minority class will be eliminated and the latest node will be marked as a majority class [23], [152].

2.2 UNDER-SAMPLING TECHNIQUE

Under-sampling is a traditional technique used to overcome the class imbalance problem. It tries to minimize the number of instances in the major class and produce a more balanced class distribution to help the classifiers differentiate between the minority and majority classes correctly.

2.2.1 Random Under-Sampling (RUS)

RUS is one of the simplest under-sampling techniques. It can efficiently balance the instances in various classes and provide a more balanced dataset. In RUS, the majority class instances are discarded arbitrarily until the optimal balance between classes is reached, thereby speeding up the training and prediction process. However, a main disadvantage of RUS is that possibly useful information may be eliminated when instances are selected to be removed and thus potentially deteriorates the performance of the prediction model [93].

2.2.2 Edited Nearest Neighbor (ENN)

The main philosophy of ENN is to exclude samples of the majority class depending on the nearest k neighbor belonging to the minority instances. When the number of neighbors predominant in each majority from minority samples, those occurrences of the majority class is excluded as overlapping cases [46], [153]. Figure 2.5 presents an example of ENN.

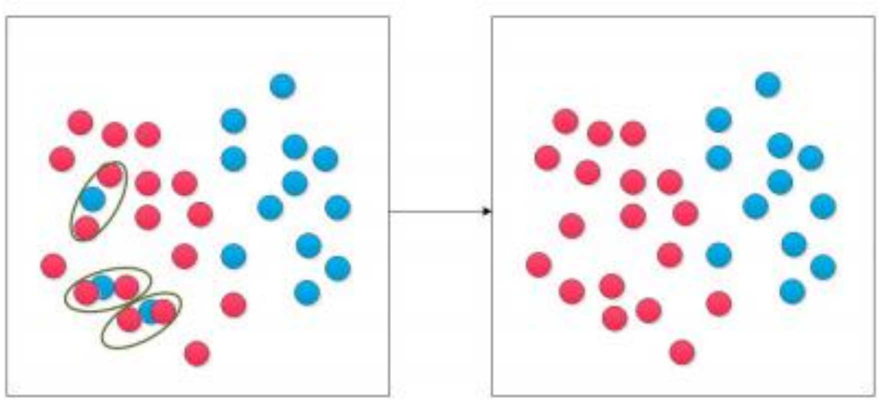


Figure 2.5: An example of ENN.

2.2.3 Tomek Link (TL)

TL is a down-sampling model designed by Tomek. It aims to remove the noise and boundary points from major class samples. TL is known to improve the Nearest neighbor rule by analyzing pairs of samples belonging to different classes but are the closest neighbor to each other and removing the major instance of the pair [103], [154]. The algorithm works as follow:

1- (a) is a sample of class 1, (b) is a sample of class 2
2- dis(a, b) is the distance measure between a & b.
3- (a, b) is a TL, if for any sample x, dis(a, b)< dis(a, x) or dis(a, b) < dis(b, x).

If any two samples are TL, then one of these samples is noise, or the two samples exist at the class boundary. Many studies use TL as a down-sampling technique such as [155]–[157].

2.3 EVALUATION METRICS

Usually, a classifier is evaluated by a confusion matrix containing four output values from classification process which indicate the True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). TP defined as the number of positive cases accurately accepted as positives. TN is the number of negative cases accurately assigned as negatives. FP is the number of negative cases inaccurately recognized as positives, and the FN is the number of positive cases inaccurately assigned as negative. A confusion matrix for two classes issue is presented in table (2.1). The positive refers to the minor class, and the negative refers to major class.

Table 2.1: Confusion matrix for two classes issue [102].

		Predicted	
		Positive	Negative
True	Positive	TP(True Positive)	FN(False Negative)
	Negative	FP(False Positive)	TN(True Negative)

Accuracy is the commonly used performance measurement in the classification process. However, it is not an acceptable criterion when assessing the unbalanced class performance, since the classifier has a significant bias against the dominant class and failed to identify the few instances of the minor class [102]. For this purpose, more effective measures have been employed depending on the confusion matrix which explained as follows:

1- Sensitivity or Recall is also indicated as True positive rate (TPR), and it is defined as the ratio of positive samples classified correctly, that tells how good the technique is in determining the minority class samples [158].

2- Specificity is also indicated as True negative rate (TNR), and it is defined as the ratio of negative samples classified correctly, that tells how good the technique is in determining the majority class samples [158].

3- Precision is also indicated as positive predictive value (PPV), and it is defined as the probability of the correctly classified positive samples [138].

4- F1-Measure (F1-M) which combines TPR and PPV, it is a harmonic mean between them. It reveals the discrimination capability of a classifier in the minority class. As a result, F1-M provides more understanding of the functionality of a classifier than the accuracy measure[151].

5- Balanced-ACcuracy (B-AC) [159]. The Balanced accuracy is the arithmetic average of TPR and TNR.

Unlike accuracy, both TPR and PPV are less sensitive to changes in data distributions. As an estimation of the accuracy for the positive class, PPV is somewhat sensitive to data distributions, while TPR is not. TPR does not offer much insight into how many instances are wrongly classified as positive. Similarly, PPV does not tell us how many positive instances are wrongly classified. However, if used correctly, PPV and TPR can efficiently assess performance in unbalanced class classification. The aforementioned metrics are given below by equations (2.17), (2.18), (2.19), (2.20) and (2.21):

$$sensitivity = \frac{TP}{TP + FN} \tag{2.17}$$

$$specificity = \frac{TN}{TN + FP} \tag{2.18}$$

$$precision = \frac{TP}{TP + FP} \tag{2.19}$$

$$F1 - M = \frac{2 * TPR * PPV}{TPR + PPV} \tag{2.20}$$

$$B - AC = \frac{TPR + TNR}{2} \tag{2.21}$$

3. HELLINGER DISTANCE UNDER-SAMPLING MODEL (HDUS)

This chapter aims to provide a technique that addresses the problem of the unbalanced class distribution that effects on classification efficiency of minor class instances. We proposed an under-sampling model that extracts samples from the majority class, whereas the minority class data would remain unchanged. It depends on the idea that it is best to maintain the samples of the minority class as actual as they are, in such a manner that no larger or smaller amount of samples is trained on them. The classifier would then be supplied with an accurate identification ability for the instances of the original minor class. The choice of instances in Under-sampling method relies on how to pick the major class samples in a method to maintain the compatibility amongst classes. For this reason, we used the Hellinger Distance metric to select samples from the majority class depending on their distance with the minority class samples.

3.1 HELLINGER DISTANCE (HD)

HD is an asymmetric and non-negative metric for measuring how far apart the two discrete distributions lying in the space of probability distributions mutual to them[160], [161]. It is, presented by Ernst Hellinger in 1909, derived from the Bhattacharyya coefficient [162].

Recently, HD has gained a lot of interest in the data mining field to identify fails in the performance of a classifier due to changes in the data distribution. It has presented very successful calculation in identifying match points in classifier due to changes in class probability [163]. Let a and b are probability functions which are continuous regarding to a third probability function x. HD between a and b can be defined in terms of integral as in equation (3.1):

$$HD\ (a,b) = \frac{1}{\sqrt{2}} \sqrt{\int \left(\sqrt{\frac{da}{d\lambda}} - \sqrt{\frac{db}{d\lambda}} \right)^2 d\lambda} \tag{3.1}$$

Here, $\frac{da}{d\lambda}$ and $\frac{db}{d\lambda}$ are two probability density functions. In machine learning, we generally equate the conditional probabilities resulting from countable data numbers instead of continued function. The available data can be identified in terms of a finite set of classes and attributes (which represent as (y/x), where y refers to a majority(j) or minority(n) class, and x refers to attributes. As we are concerned in assessing on a discrete instead of a continual space, we might transform the integral into summation as in equation (3.2) to sum all values and reformulate the context of the aforementioned equation to be as in equation (3.2)

$$HD\ (j,n) = \frac{1}{\sqrt{2}}\ \sqrt{\sum\left(\sqrt{j/xi} - \sqrt{n/xi}\,\right)^2} \tag{3.2}$$

Equation 3.2 represents the distance between classes along with the set of features. The $\frac{1}{\sqrt{2}}$ in the equation is for guaranteeing that HD (j, n) ≤ 1 with all distributions so that it bounds values in between 0 and 1. Hellinger distance can be utilized to calculate the degree of resemblance between two probability distributions; if the distance is 0, the two distributions are the same, and if the distance is 1, they are the furthest. However, there are several criteria to be considered [164]:

1- Symmetric property: HD (j,n) = HD (n,j)
2- Non-negative condition: HD (j,n) >= 0
3- Identity property: HD(j, n) =0 if and only if j= n
4- Triangular inequality

In addition to these criteria, previous studies revealed that the advantages of employing HD are: the measure depends only on the attributes of the concerned instances, which is not altered when an extended set of attributes is considered. And, the measure doesn't rely on the size of the instance on which the attributes are evaluated [165]. More than that, in [166]–[169] proved that Hellinger distance is robust in the existence of an intensive skewed class distribution since an excess of one class instances would only help to put the distribution of the instance nearer to the actual distribution. In this case, if a feature is a strong class identifier, it will not be influenced by the unbalanced class ratio due to its isometric contours. Accordingly, the aforementioned properties are the motivation to use HD in our suggested model.

3.2 THE PROPOSED HELLINGER DISTANCE UNDER-SAMPLING MODEL (HDUS)

In this study, we proposed (HDUS) to address the problem of imbalance class classification. It manipulates samples of the major class while leaving the samples of the minor class without change. The model employs Hellinger metric to measure the resemblance value between the majority and minority instances, and select a number of majority instances with the highest resemblance values until balanced with the number of the minor instances. Then, the original minority class instances will combine with the selected instances of the majority class; which consider to be consistent with the minority instances, to produce a balanced training dataset that give the purist identification of the minor class. Thus, HDUS works to reduce the number of majority class samples, with the goal of enhancing the predictive efficiency of the minor class, which is the class of the greatest interest in most real word datasets. Table 3.1, presents the HDUS Pseudo code algorithm.

Table 3.1: The HDUS pseudo code algorithm.

Input: Imbalanced Training dataset (ITrD)

Output: Balanced Training dataset (BTrD)

Group the ITrD according to the classes
C1= ITrD (class1) // C1 indicates the minor class which contains less number of instances
C2= ITrD (class2) // C2 indicates the major class which contains more number of instances
For i in rows of (C2)
    For j in rows of (C1)
        $Sim_{i,j}$ = calculate the similarity between C2(i) and C1(j) using Hellinger Distance
        append $Sim_{i,j}$ To HD(i)
    Next j
    select m top values from HD (i) // where m is a given number of neighbouring minority class
    HD_{sum}(i)= sum the selected m top values
Next i
C2HD=select w majority class instances according to the highest similarity value in HD_{sum}(i), // where w is a given number
return (BTrD= C2HD +C1)

3.3 DATASETS

In this study, four unbalanced medical datasets were used to estimate the performance of the (HDUS) model. The number of features, instances, majority instances, minority instances and the minority percentage (MP) of the instances are provided with each dataset. The datasets are summarized below.

3.3.1 Colorectal Cancer Dataset (CRCD)

The CRCD comes from the University Hospital of Southampton, which was used with the permission of the authorized surgeon (co-author), and all data are nameless. The details relates to patients with primary tumour at twelve colorectal (CRC) positions, who are undergo tumour resection operation. The data contain 1005 cases, each case represent a single patient with fourteen attributes, along with a class label. The 1005 cases consist 760 patients with primary colorectal tumour; represent the major instances, and 245 patients with primary colorectal tumour spreading to other part of the body named as CRC metastasis; represent the minor instances. The percentage of the minority instances is 24% of the data. Table 3.1 displays the attributes of CRCD.

3.3.2 PIMA Dataset (PIMAD)

This dataset has been obtained from the UCI repository [170]. It consists of 9 attributes, along with the class attribute. The class attribute specifies if the patient having or not having diabetes. This data consists of 768 cases, where 268 with diabetes (represent a minor instances) and 500 with no diabetes (represent a major instances). The minority percentage is (34%). Table 3.2 displays the attributes of PIMAD.

3.3.3 Thoracic Surgery Dataset (THSD)

This dataset has been obtained from the UCI [170]. It is for patients with primary lung tumour who then undergo tumour resection surgery. The data has 17 attributes, along with the class attribute. The data consists of 470 cases, with 70 are for patients died through the first year after operation (represent the minor instances) and 400 are alive (represent the major instances). The minority percentage is (14%). Table 3.3 displays the attributes of THSD.

3.3.4 Breast Cancer dataset (BCD)

This dataset has been obtained from the UCI [170]. It consists of ten attributes, along with the class attribute. The class attribute specifies if the breast tumour returned or not. This data consist 286 cases, 85 represent the minor class and 201 represent the major class. The minority percentage is (29%). Table 3.4 displays the attributes of BCD.

Table 3.2: The attributes of CRCD.

No.	Attribute Name	Data type
1-	Tumour site (in colorectal)	Categorical
2-	Surgery Type	Categorical
3-	Operation Type (on which part of colorectal the operation has done)	Categorical
4-	Differentiation	Categorical
5-	Dukes Stage of tumour	Categorical
6-	T stage 5th edition	Categorical
7-	N stage 5th edition	Categorical
8-	EMVI	Categorical
9-	Tumour Perforation	Categorical
10-	Resection Margin	Categorical
11-	Neoadjuvant therapy n-CRT	Categorical
12-	Chemotherapy	Categorical
13-	Radiotherapy	Categorical
14-	CRC Metastasis (class)	Categorical

Table 3.3: The attributes of PIMAD.

No.	Attribute Name	Data type
1-	Number of times pregnant	Numeric
2-	Plasma glucose concentration	Numeric
3-	Diastolic blood pressure	Numeric
4-	Triceps skinfold thickness	Numeric
5-	Amount of insulin	Numeric
6-	Body mass index	Numeric
7-	Diabetes pedigree function	Numeric
8-	Age	Numeric
9-	Class	Categorical

Table 3.4: The attributes of THSD.

No.	Attribute Name	Data type
1-	Diagnosis	Categorical
2-	Forced vital capacity	Numeric
3-	A volume that has been exhaled at the end of the first of forced expiration	Numeric
4-	Performance status	Categorical
5-	Pain before surgery	Categorical
6-	Hemoptysis before surgery	Categorical
7-	Dyspnoea before surgery	Categorical
8-	Cough before surgery	Categorical
9-	Weakness before surgery	Categorical
10-	Size of the original tumour	Categorical
11-	Type 2 diabetes mellitus	Categorical
12-	Myocardial Infarction up to six months	Categorical
13-	Peripheral arterial diseases	Categorical
14-	Smoking	Categorical
15-	Asthma	Categorical
16-	Age at surgery	Numeric
17-	One year survival period(class)	Categorical

Table 3.5: The attributes of BCD.

No.	Attribute Name	Data type
1-	tumor size	Categorical
2-	Inv nodes	Categorical
3-	Node caps	Categorical
4-	menopause	Categorical
5-	deg malig	Categorical
6-	Breast side	Categorical
7-	breast quad	Categorical
8-	irradiat	Categorical
9-	Age	Categorical
10-	Class (recurrence/ no-recurrence)	Categorical

3.4 FINDINGS AND RESULTS

To assess the performance of the suggested HDUS method, we employed four unbalanced medical datasets, named CRCD, PIMAD, THSD and BCD, using three different classifiers, included KNN, DT and SVM. We have compared the results of HDUS with a baseline model (not use any sampling model) and with state-of-arts under-sampling techniques (ENN, TL and RUS).

For all experiments in our study, the datasets are divided to 75% for training and 25% for testing. Then a 5-fold-cross-validation is used as an evaluation criterion, to confirm an accurate assessment of the methods. All the experiments were performed through a code written in Python programming language, which provides all the necessary functions. The findings of the data sets(CRCD, PIMAD, THSD and BCD) are displayed in tables 3.1, 3.2, 3.3 and 3.4, by using B-AC, recall, precision, specificity and F1-M.

As seen in tables 3.1-3.4, the 1st column shows that there is an unbalanced classification problem in all data sets. This provides a poor sensitivity rating for predicting the rare class; ranging from 7.9 per cent in THSD, to 39.78 per cent in PIMAD, whereas assigning high specificity ratings for predicting the majority class samples.

The second, third and fourth columns of the tables present the results of the state-of-arts under-sampling models: TL, RUS and ENN respectively. They show the enhancement obtained by utilizing these models, as stated by the sensitivities, which represent the capability of these methods to predict the minority class (the class of importance). Although the TL achieved lower performance in all data sets, it is higher than the baseline excepting for the THSD, which did less than others.

An additional enhancement is obtained in the fifth column by the suggested HDUS model, which shown in the metrics recall, F1-Measure and balanced accuracy. This stated that the efficiency of HDUS shows considerable improvement over the compared models. HDUS resulted in the top rating of sensitivity for the data sets(referring to the highest ability to identify the rare class). It ranks more than 80 per cent in both the CRCD and the PIMAD, nearly 70 per cent in BCD and nearly 60 per cent in THSD. It also results in the best F1-M and B-AC rate.

Table 3.6: The results for CRCD by three classifiers and five models.

CRCD		baseline	TL	RUS	ENN	HDUS
KNN	Sensitivity %	29.4	37.2	62.7	62.7	80.3
	Specificity %	85.3	75.3	60	60.3	50.2
	Precision %	44.1	37.2	38.1	38.5	33.9
	F1-M %	35.2	37.2	47.4	47.7	47.6
	B-AC %	57.4	56.3	61.3	61.5	65.2
SVM	Sensitivity %	5.8	29.4	62.7	47.1	76.4
	Specificity %	92.3	82.3	54.6	70.7	55.8
	Precision %	23.0	39.4	35.1	38.7	35.4
	F1-M %	9.3	33.7	45.0	42.4	48.3
	B-AC %	49.1	55.8	58.6	58.9	66.1
DT	Sensitivity %	35.2	45.1	62.7	66.6	81.0
	Specificity %	68.4	59.2	46.9	43.8	56.9
	Precision %	30.5	30.2	31.6	31.7	39.9
	F1-M %	32.7	36.2	42.1	43.04	53.4
	B-AC %	51.8	52.1	54.8	55.2	68.9
AVG	Sensitivity %	23.5	37.2	62.7	58.8	79.2
	Specificity %	82.1	72.3	53.8	58.3	54.3
	Precision %	32.5	35.6	34.9	36.3	36.4
	F1-M %	27.3	36.4	44.9	44.9	49.8
	B-AC %	52.7	54.7	58.3	58.5	66.7

Table 3.7: The results of PIMAD by three classifiers and five models.

PIMAD		baseline	TL	RUS	ENN	HDUS
KNN	Sensitivity %	60.1	64.5	70.7	70.5	83.8
	Specificity %	83.8	80	73.8	74.9	62.5
	Precision %	63.2	56.6	56.4	56.4	50.3
	F1-M %	61.5	60.3	62.7	62.7	62.9
	B-AC %	70.9	72.2	72.4	72.7	73.2
SVM	Sensitivity %	0	66.1	74.4	74	79.03
	Specificity %	100	83.8	73.8	76.9	71.5
	Precision %	0	58.1	58.5	58.04	57.0
	F1-M %	0	61.8	65.5	65.06	66.2

	B-AC %	50	74.9	74.6	75.8	75.2
DT	Sensitivity %	61.2	69.3	70.9	70.9	91.9
	Specificity %	79.2	66.9	58.4	68.4	66.7
	Precision %	58.4	50	44.9	50.7	56.9
	F1-M %	59.8	58.1	55	59.2	70.3
	B-AC %	70.2	68.1	64.7	69.7	79.3
AVG	Sensitivity %	39.7	66.6	72.1	71.8	84.9
	Specificity %	87.6	76.9	68.7	73.4	66.9
	Precision %	40.5	54.9	53.2	54.1	54.7
	F1-M %	40.1	60.2	61.2	61.7	66.5
	B-AC %	63.7	71.8	70.5	72.7	75.9

Table 3.8: The results for THSD by three classifiers and five models.

THSD		baseline	TL	RUS	ENN	HDUS
KNN	Sensitivity %	0.0	0.0	42.8	4.7	23.8
	Specificity %	100.0	98.9	60.8	91.7	75.2
	Precision %	0.00	0.0	19.1	11.1	19.2
	F1-M %	0.00	0.0	26.4	6.6	21.2
	B-AC %	50.0	49.4	51.8	48.2	49.5
SVM	Sensitivity %	0.0	0.0	66.6	4.7	71.4
	Specificity %	100.0	100.0	47.4	91.7	44.2
	Precision %	0.0	0.0	21.5	11.1	21.1
	F1-M %	0.0	0.0	32.5	6.6	32.6
	B-AC %	50.0	50.0	57.04	48.2	57.8
DT	Sensitivity %	23.8	14.2	42.8	38.1	80.9
	Specificity %	87.6	91.7	48.4	81.4	40.0
	Precision %	29.4	27.2	15.2	30.7	25.9
	F1-M %	26.3	18.7	22.5	34.0	39.3
	B-AC %	55.7	53.0	45.6	59.7	60.4
AVG	Sensitivity %	7.9	4.7	50.7	15.8	58.7
	Specificity %	95.8	96.9	52.2	88.3	53.1
	Precision %	9.8	9.1	18.6	17.6	22.1
	F1-M %	8.7	6.2	27.1	15.7	31.0
	B-AC %	51.9	50.8	51.5	52.0	55.9

Table 3.9: The results for BCD by three classifiers and five models.

BCD		baseline	TL	RUS	ENN	HDUS
KNN	Sensitivity %	33.3	44.4	57.1	50	61.1
	Specificity %	84.9	70.3	65.8	70.3	73.5
	Precision %	42.8	40	40.7	42.8	44.0
	F1-M %	37.5	42.1	47.5	46.1	51.2
	B-AC %	59.1	57.4	61.4	60.1	67.3
SVM	Sensitivity %	22.2	38.8	66.6	44.4	66.6
	Specificity %	94.3	88.6	64.1	83.0	69.8
	Precision %	57.1	53.8	38.7	47.0	42.8
	F1-M %	32	45.2	48.9	45.7	52.1
	B-AC %	58.2	63.7	65.4	63.7	68.2
DT	Sensitivity %	38.8	38.8	44.4	44.4	77.7
	Specificity %	66.04	64.1	62.2	69.8	66.0
	Precision %	28	26.9	28.5	33.3	43.7
	F1-M %	32.5	31.8	34.7	38.1	56.0
	B-AC %	52.4	51.5	53.3	57.1	71.9
AVG	Sensitivity %	31.4	40.7	56.1	46.2	68.5
	Specificity %	81.7	74.4	64.1	74.4	69.8
	Precision %	42.6	40.2	36.0	41.1	43.5
	F1-M %	36.2	40.5	43.8	43.5	53.1
	B-AC %	56.6	57.5	60.1	60.3	69.1

3.5 DISCUSSION

This section discusses the proposed HDUS model, addresses the imbalanced class distribution by employing the Hellinger distance similarity measure to enhance the predictive efficiency of the rare class instances.

It is important to handle the class inequality problem by using a suitable approaches, as we can see from the following observations. As presented in the results section, the baseline model displays a high TNR for predicting majority class cases, but a low TPR for predicting rare class cases, which is a class of importance in an unbalanced data sets. Using traditional down-sampling methods demonstrates good development, mostly by RUS. Nevertheless, using RUS tends not to have been efficient, as it excludes significant cases at random and may even induce over-fitting due to the spread of scars cases with no limitation. More than that, the results by ENN are worse than that by RUS, excepting for that in PIMAD, and TL worst in the study. On the other hand, HDUS is shown to exceed all models in the study due to the powerful measuring of Hellinger distance, that has a skew intensive feature help to not influence by the class inequality. To clarify the comparison between the various under-sampling models used in this chapter and to estimate their effectiveness, figure 3.1, 3.2 and 3.3 offer a graphical demonstration of the rate for (recall, F1-Measure and B-AC) resulting from five methods used in four unbalanced medical data sets. The graphics show that results differs when various under-sampling methods are used. Figure 3.1 shows that HDUS has advanced dramatically in estimating the sensitivity of minor class instances in all data sets. A similar situation can be seen in figure 3.2 for F1-Measure, that defines the trade-offs among sensitivity and precision, also figure 3.3 for Balanced accuracy, that defines the trade-offs among specificity and sensitivity.

Concerning the classification methods, it should be noted that the advantage of performing classification rises as the unbalanced class problem is properly tackled. In this chapter, various classifiers can improve identifying rare class when using HDUS. Particularly, Decision Tree shows the highest result with HDUS. It also achieved good outcomes with ENN and TL, while support vector machine more suitable with RUS model. Table 3.5 shows a performance of the used classification algorithms with under-sampling models, and four data sets have increased the predictability of the minority class samples by the metrics recall, F1-M and B-AC. Lastly, the findings of HDUS model can be regarded as an introductory study but an encouraging methodology for under-sampling the unbalanced data sets to increase the classification efficiency of minority class cases.

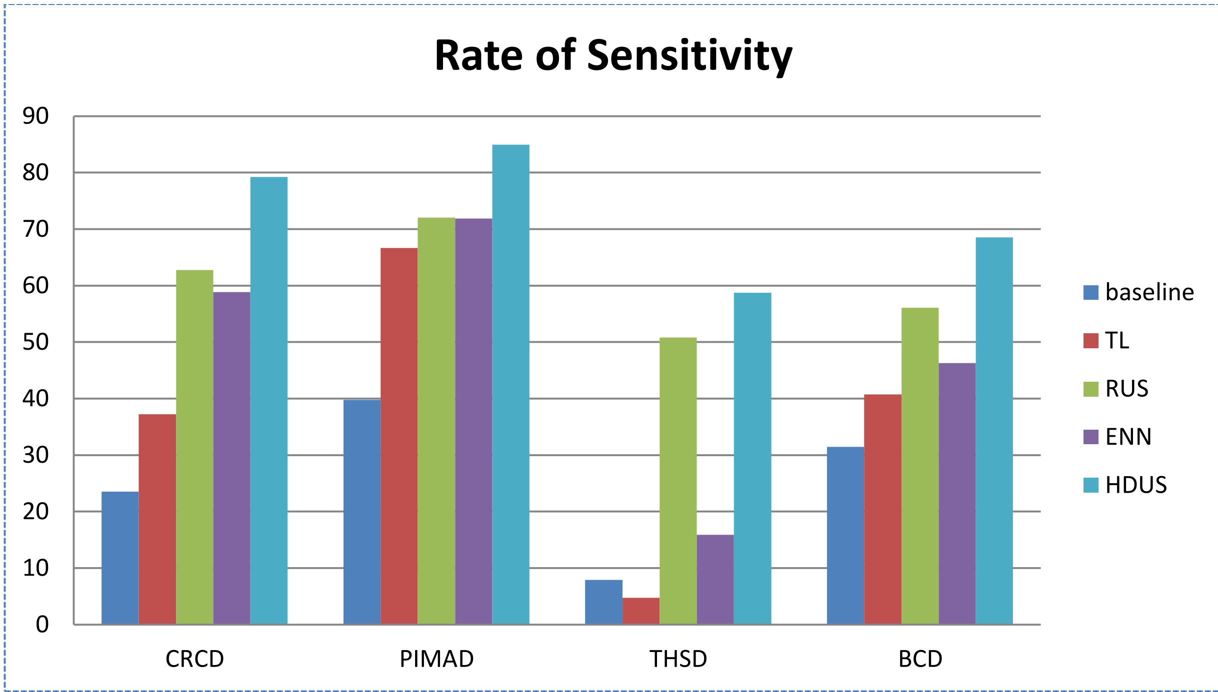


Figure 3.1: A rate of sensitivity for five methods with four unbalanced medical datasets.

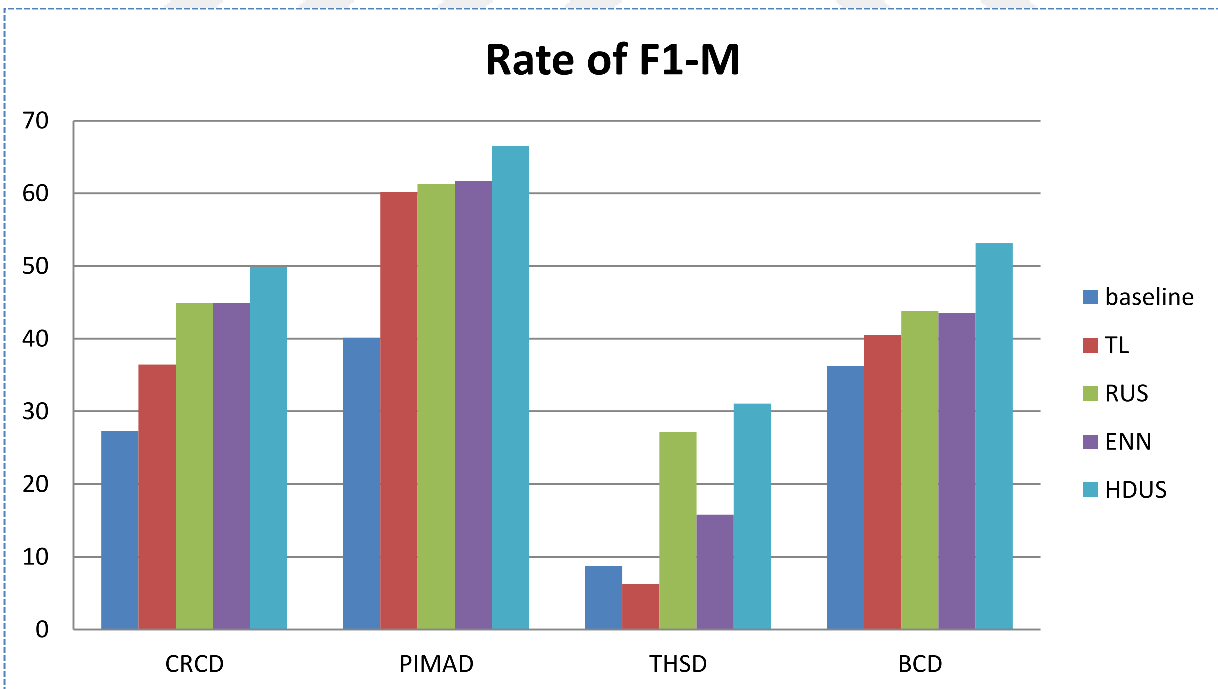


Figure 3.2: A rate of F1-Measure for five methods with four unbalanced medical datasets.

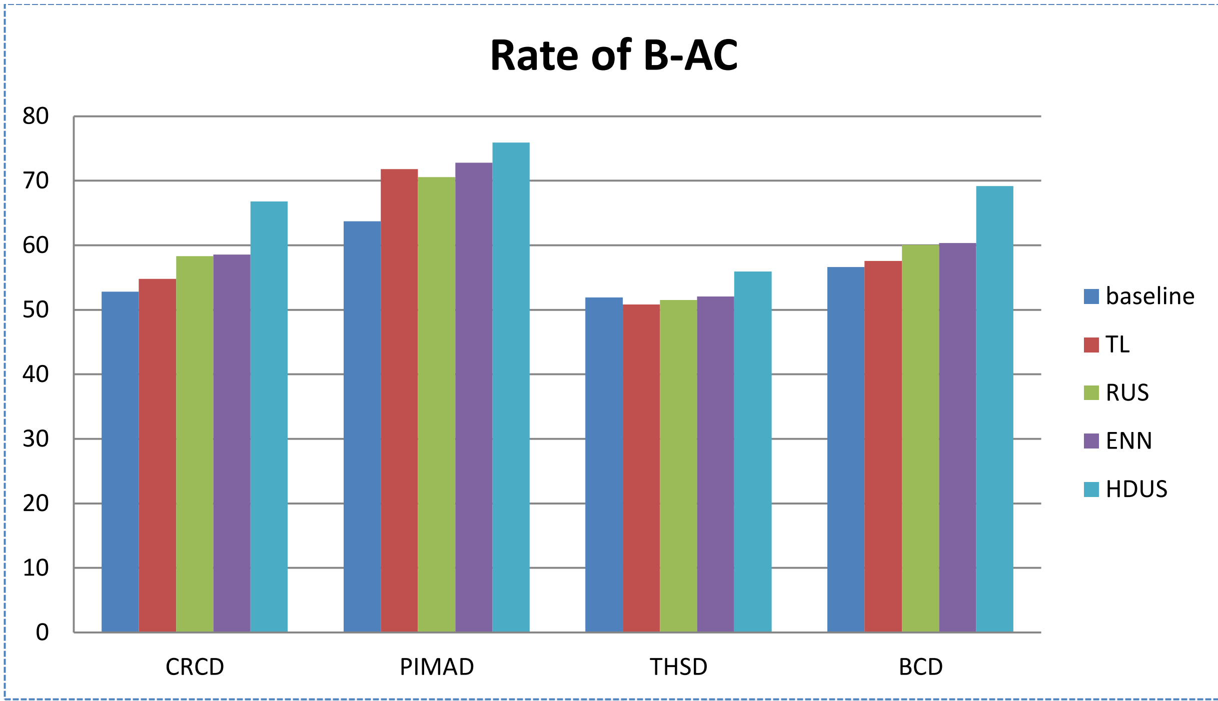


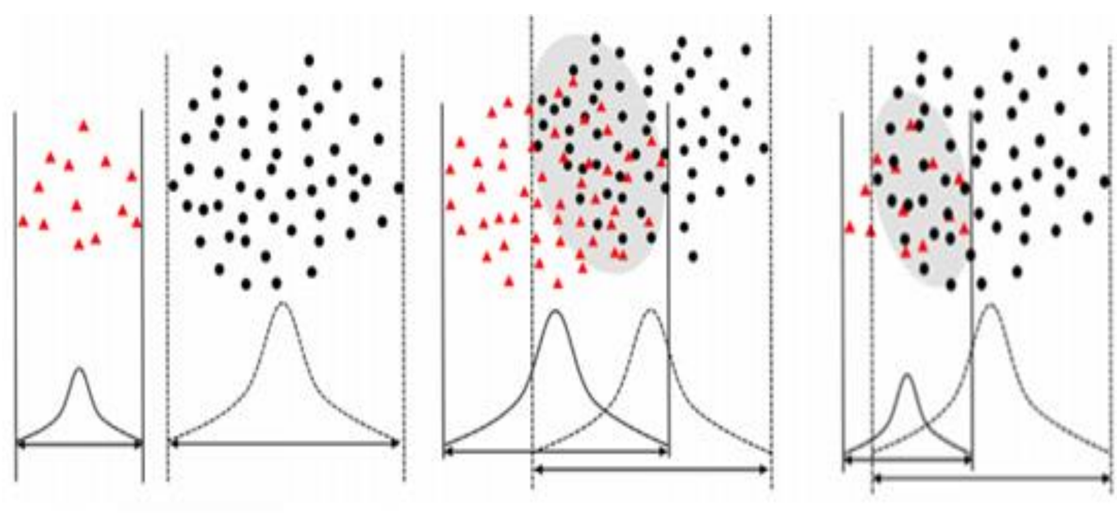
Figure 3.3: A rate of balanced accuracy for five methods with four unbalanced medical datasets.

Table 3.10: A mean rate for three classification algorithms with five methods on four unbalanced medical datasets.

		baseline	TL	RUS	ENN	HDUS
KNN	Sensitivity %	30.7	36.5	58.3	47.0	62.2
	Specificity %	88.5	81.1	65.1	74.3	65.3
	Precision %	37.5	33.4	38.6	37.2	36.8
	F1-M %	33.5	34.9	46.0	40.8	45.7
	B-AC %	59.3	58.8	61.7	60.6	63.8
SVM	Sensitivity %	7.03	33.6	67.6	42.5	73.3
	Specificity %	96.6	88.7	60.0	80.6	60.3
	Precision %	20.0	37.8	38.4	38.7	39.1
	F1-M %	10.3	35.1	48.0	39.9	49.8
	B-AC %	51.8	61.1	63.9	61.6	66.8
DT	Sensitivity %	39.8	41.9	55.2	55.0	82.9
	Specificity %	75.3	70.5	54.0	65.8	57.4
	Precision %	36.6	33.6	30.1	36.6	41.6
	F1-M %	37.8	36.2	38.6	43.5	54.7
	B-AC %	57.5	56.2	54.6	60.4	70.1

4. A HYBRID UNDER-SAMPLING AND FEATURE SELECTION MODEL

In the unbalanced class classification, the unequal class distribution and overlapping are detrimental problems that affect to obtain a correct prediction of the minority class [6],[171]. The unequal class distribution problem is referred to as the situation where the number of samples of the major class differs from the number of samples of the minor class. The overlapping problem is referred to the situation when samples of the majority and minority classes are distributed around heterogeneous area, and the features that associate with an overlapping border region are ignored [172], [173]. However, it is documented that a large number of attributes may weaken the detection of the boundary line between classes because either any of these features may be redundant or they do not have a strong correlation between them. Figure 4.1 shows an example for classification (a) unequal class distribution, (b) class overlap problem, and (c) both unequal class distribution and overlapping problems. The shaded area denotes to an overlapping problem. The solid lines and the dashed lines display the distributions for each class.



(a) (b) (c)

Figure 4.1: Example for classification (a) Unequal class distribution, (b) Class overlapping problem, and (c) Unequal class distribution and overlapping problems.

The study [2] found that borderline identification is better after elimination of noise features. Therefore, using feature selection can optimize the contrast between classes and reduce the impact of an overlapping problem[8]. But, the conventional feature selection models may not work well in an unbalanced dataset and the features may be affected by the instances of the major class thereby influence the predicting of rare class which we usually pay more attention [125]. Therefore, using feature selection alone may not address the overlapping issue, and it is important to conduct the resampling on training data instances before applying feature selection to avoid the bias of the classifiers against the majority class. In this chapter, we aim to provide a technique that addresses the problems of classifying caused by both the unequal class distribution and overlapping, to improve the performance of minority class samples.

Our contribution in this chapter is to introduce a method that develops the classification of rare class in the context of unbalanced data by combining both under-sampling and feature selection techniques. The object of Under-sampling technique is to balance the distribution of instances among the majority and minority classes to eliminate the noise and border instances that hinder the ability to classify imbalanced data. Using selecting features technique can reduce the borders of the problem by restricting the effect of those features which cause difficulties in discriminating the minority class. The proposed model use Hellinger Distance under-sampling model [174], building in chapter 3 to choose instances from the major class having highest similarity with the minor class instances, the result then processes with ANOVA f-test to select the relevant features which then evaluated by classification algorithms.

4.1 ANALYSIS OF VARIANCE (ANOVA) F-TEST FEATURE SELECTION

ANOVA f-test is a type of statistical measure used to analyze data for which reaction attributes are evaluated within different conditions defined by classification attributes. The standard aim of ANOVA is comparing the mean of reacting attributes with different combinations of classification attributes. ANOVA was utilized to determine if the attribute indicates a substantial variation between two or more classes [175]–[178]. The statistic for ANOVA is named f-test, that can be performed by the below equations[179]:

1- A variance among the classes is measured as in equation 4.1:

$$Variance\ among\ classes\ = \sum_{i=1}^{k} ni(\ YM_i - YM)^2/(k-1) \tag{4.1}$$

Where in equation 4.1, ni is the observation in ith class, YMi is the sample mean in the ith class, YM is the average of the data and k is classes.

2- The variance within the classes is measured as in equation 4.2:

$$Variance\ within\ the\ class = \sum_{i=1}^{k}\sum_{j=1}^{ni} (Y_{ij} - YM_i\ )^2/(N-k) \tag{4.2}$$

Where in equation 4.2, Yij is the jth observation in the ith out of k class, YMi is the sample mean in the ith class, N is the overall sample size and k is classes.

3- ANOVA f-test is measured as in equation 4.3 :

$$f - test = \frac{variance\ among\ classes}{variance\ within\ the\ class} \tag{4.3}$$

4.2 THE PROPOSED HDUS FEATURE SELECTION (HDUS+FS)MODEL

In this chapter, our contribution is to introduce a method that combines HDUS model and ANOVA feature selection in a cascade manner to identify the appropriate subset of instances and features which improve classifying the minority class in the context of an unbalanced data sets. The HDUS works to balance the instance distribution among the majority and minority classes by removing the instances from the major class that hinder the discrimination between classes. Then, the ANOVA f-test will select the most relevant features to reduce the boundaries of the problem by restricting the effect of those features, which cause difficulties in determining minority samples.

The proposed model implemented in three stages before the classification process. Below is a brief description about the proposed model.

Let D[N, M] is a training dataset; N is a number of rows (instances) and M is a number of columns (features). Each row and column in a dataset is considered as one vector. Our method procedure is as follows:

1- Split the data to groups according to number of classes. In our datasets, we have two groups of classes. Class1 denotes to the minor class rows, and class2 denotes to the major class rows.
2- HDUS model employs Hellinger distance metric to find the similarity value between the majority instances that are considered to be consistent with minority instances, and select a number of majority instances with the highest similarity values until balanced with the number of the minority instances. Then, forward them to feed the next step. The obtained data get rid of the redundant instances but have the same structure as the original (M) features.
3- The compressed data from previous stage is fed to ANOVA f-test to measure the weight of each feature with respect to target class feature and choose the features with highest f-test value which have strong correlation to class feature, thereby creating a new subset selected from M features while maintaining the important information as much as possible.

4.3 EXPERIMENTAL RESULTS

To investigate the performance of the HDUS-FS model, we performed experiments on four unbalanced medical data sets (CRCD, PIMAD, THSD and BCD). The main information of these datasets is summarized in section (3.3). We used three classification algorithms: DT, SVM and KNN, to obtain the performance measure over these datasets in terms of, sensitivity, specificity, precision, F1-M and B-AC. These metrics described in section (2.3).

The current experimental setup trains the classifiers at 75% data and considers the rest for testing. Then a 5-fold-cross-validation is used as an evaluation criterion, to confirm an accurate assessment of the methods. All the experiments were performed through a code written in Python programming language, which provides all the necessary functions. We compared HDUS+FS method against the baseline model (with original data), the model of ANOVA feature selection alone and the models that combined three state-of-arts, under-sampling models, with ANOVA feature selection; TL+FS, ENN+FS and RUS+FS.

The results are shown in tables 4.1, 4.2, 4.3, and 4.4. As presented in the first column of the four tables (the baseline model), the sensitivity rate for classifying the minor class is quite low; it ranges from 7.9% in THSD, which is the most unbalanced datasets having MP= 14%, to 41.4% in PIMAD, which is moderately unbalanced having MP= 34%, where the baseline classifier specifies all the cases to the major class.

When examining the second column (FS-alone), the ANOVA feature selection model, we can notice that the obtained outcomes are very close to the results of the first column with all initial features over most classifiers, with a little improvement in PIMAD and BCD, as expressed in term of sensitivity, F1-M and B-AC.

Furthermore, the third, fourth and fifth columns display the result of state-of-arts under-sampling methods combined with ANOVA feature selection; TL+FS, RUS+FS and ENN+FS. Comparing these hybrid models with the model of feature selection alone, we can notice an enhancement in sensitivity made by using these models, except in TL+FS score lower sensitivity than FS-alone in both THSD and BCD.

More enhancement is accomplished in the sixth column of all tables by HDUS+FS model over the baseline, FS-alone and the three hybrid state-of-arts-under-sampling with FS models. The proposed model results in the highest average rate of sensitivity, F1-M and B-AC for CRCD, THSD and BCD datasets, except for PIMAD, where ENN+FS exceeded the proposed model by just 2%.

Accordingly, the above observations show that the hybrid under-sampling and feature selection models, which simultaneously solve both imbalanced and overlapping problems, are generally better than utilizing FS-alone, particularly when the Minority Percentage is very low. These observations refer to the highest ability of the hybrid under-sampling and feature selection to identify the rare class, which is the class of interest in many datasets.

Table 4.1: Results for CRCD using three classifiers and six methods including the baseline, FS-alone, and hybrid under-sampling with FS models

CRCD		baseline	FS-alone	TL+FS	RUS+FS	ENN+FS	HDUS+FS
KNN	Sensitivity %	29.4	23.5	41.1	52.9	70.5	88.3
	Specificity %	85.3	85.3	69.2	66.1	56.1	55.8
	Precision %	44.1	38.7	34.4	38.0	38.7	35.9
	F1-M %	35.2	29.2	37.5	44.2	50.0	49.7
	B-AC %	57.4	54.4	55.2	59.5	63.3	66.1
SVM	Sensitivity %	5.8	9.8	23.5	76.4	74.5	76.4
	Specificity %	92.3	93.1	80.0	55.3	49.2	43.8
	Precision %	23.1	35.7	31.5	40.2	36.5	34.8
	F1-M %	9.3	15.3	26.9	52.7	49.0	47.8
	B-AC %	49.1	51.4	51.7	65.9	61.8	60.1
DT	Sensitivity %	37.2	33.3	56.8	68.6	68.6	82.3
	Specificity %	73.1	66.9	59.2	48.4	46.9	50.7
	Precision %	35.1	28.3	35.3	34.3	33.6	39.6
	F1-M %	36.1	30.6	43.6	45.7	45.1	53.5
	B-AC %	55.1	50.1	58.1	58.5	57.7	66.5
AVG	Sensitivity %	24.1	22.2	40.5	66.0	71.2	82.4
	Specificity %	83.5	81.7	69.4	56.6	50.7	50.1
	Precision %	34.1	34.2	33.7	37.5	36.3	36.8
	F1-M %	26.9	25.1	36.0	47.5	48.1	50.3
	B-AC %	53.8	52.0	55.0	61.3	61.0	66.2

Table 4.2: Results for PIMAD using three classifiers and six methods including the baseline, FS-alone, and hybrid under-sampling with FS models.

PIMAD		baseline	FS-alone	TL+FS	RUS+FS	ENN+FS	HDUS+FS
KNN	Sensitivity %	58.1	51.6	74.1	77.4	88.7	80.1
	Specificity %	83.8	81.5	80.7	72.3	70.7	60.7
	Precision %	63.1	57.1	64.7	57.1	59.1	52.6
	F1-M %	60.5	54.2	69.1	65.7	70.9	63.5
	B-AC %	70.9	66.5	77.4	74.8	79.7	70.4
SVM	Sensitivity %	0.0	32.2	62.9	77.4	83.8	88.7
	Specificity %	100.0	93.8	86.9	78.4	75.3	72.3
	Precision %	0.0	71.4	69.6	63.1	61.9	51.2
	F1-M %	0.0	44.4	66.1	69.5	71.2	64.9
	B-AC %	50.0	63.1	74.9	77.9	79.6	80.5
DT	Sensitivity %	66.1	53.2	67.7	64.5	74.1	72.5
	Specificity %	78.4	80.0	71.5	65.3	64.6	58.4
	Precision %	59.4	55.9	53.1	47.1	50.0	45.4

	F1-M %	62.6	54.5	59.5	54.4	59.7	55.9
	B-AC %	72.3	66.6	69.6	64.9	69.4	65.5
AVG	Sensitivity %	41.4	45.7	68.2	73.1	82.2	80.4
	Specificity %	87.4	85.1	79.7	72.1	70.2	64.5
	Precision %	40.8	61.5	62.5	55.7	57.0	49.7
	F1-M %	41.0	51.1	64.9	63.2	67.3	61.4
	B-AC %	64.4	65.4	74.0	72.5	76.2	72.1

Table 4.3: Results for THSD using three classifiers and six methods including the baseline, FS-alone, and hybrid under-sampling with FS models.

THSD		baseline	FS-alone	TL+FS	RUS+FS	ENN+FS	HDUS+FS
KNN	Sensitivity %	0.0	4.7	4.7	38.1	4.7	57.1
	Specificity %	100.0	97.9	98.9	73.2	90.7	55.3
	Precision %	0.0	33.3	50.0	23.5	10.0	18.4
	F1-M %	0.0	8.3	8.7	29.1	6.4	27.9
	B-AC %	50.0	51.3	51.8	55.6	47.7	56.2
SVM	Sensitivity %	0.0	0.0	0.0	66.6	0.0	76.1
	Specificity %	100.0	100.0	100.0	55.6	100.0	46.2
	Precision %	0.0	0.0	0.0	24.5	0.0	22.0
	F1-M %	0.0	0.0	0.0	35.9	0.0	34.1
	B-AC %	50.0	50.0	50.0	61.1	50.0	61.2
DT	Sensitivity %	23.8	14.2	14.2	52.3	38.1	66.1
	Specificity %	88.6	91.7	88.6	61.8	79.3	53.6
	Precision %	31.2	27.2	21.4	22.9	28.5	25.1
	F1-M %	27.0	18.7	17.1	31.8	32.6	36.3
	B-AC %	56.2	53.0	51.4	57.1	58.7	59.8
AVG	Sensitivity %	7.9	7.3	6.3	52.3	14.2	66.4
	Specificity %	96.2	96.5	95.8	63.5	90.1	51.7
	Precision %	10.4	20.2	23.8	23.6	12.8	21.8
	F1-M %	9.0	9.0	8.6	32.2	13.0	32.8
	B-AC %	52.0	52.2	51.1	57.9	52.1	59.1

Table 4.4: Results for BCD using three classifiers and six methods including the baseline, FS-alone, and hybrid under-sampling with FS models.

BCD		baseline	FS-alone	TL+FS	RUS+FS	ENN+FS	HDUS+FS
KNN	Sensitivity %	33.3	55.5	50.0	66.6	77.7	88.8
	Specificity %	84.9	69.8	81.1	60.3	54.7	49.1
	Precision %	42.8	38.4	47.3	36.3	36.8	36.0
	F1-M %	37.5	45.4	48.6	47.1	50.0	51.2
	B-AC %	59.1	62.6	65.5	63.5	66.2	68.9

SVM	Sensitivity %	22.2	44.4	50.0	66.6	72.2	66.6
	Specificity %	94.3	90.5	86.7	67.9	62.2	67.3
	Precision %	57.1	61.5	56.2	41.3	39.3	40.3
	F1-M %	32.0	51.6	52.9	51.1	50.9	50.2
	B-AC %	58.2	67.5	68.4	67.3	67.2	67.0
DT	Sensitivity %	38.8	55.5	38.8	66.6	61.1	66.6
	Specificity %	66.0	79.2	86.7	60.3	66.0	67.9
	Precision %	28.0	47.6	50.0	36.3	37.9	41.3
	F1-M %	32.5	51.2	43.7	47.1	46.8	51.1
	B-AC %	52.4	67.4	62.8	63.5	63.5	67.3
AVG	Sensitivity %	31.4	51.8	46.3	66.6	70.3	74.1
	Specificity %	81.7	79.8	84.9	62.8	61.0	61.4
	Precision %	42.6	49.2	51.2	38.0	38.1	39.2
	F1-M %	34.0	49.4	48.4	48.3	49.2	50.8
	B-AC %	56.6	65.8	65.6	64.7	65.6	67.7

4.4 DISCUSSION

In this chapter, we investigate about combining a preprocessing under-sampling method with feature selection technique in a cascade manner. For under-sampling, we used the HDUS model proposed in chapter 3, and for feature selection, we used ANOVA f-test. The proposed model, HDUS+FS, addresses the unequal class distribution and overlapping problems in unbalanced datasets to enhance the identification of the rare class instances. The model has been evaluated through three different classification algorithms (KNN, SVM, DT) and has been conducted on four imbalanced medical datasets with different percentages of original minority instances(CRCD (24%), PIMAD (34%), THSD (14%), BCD (29%)). We are compared HDUS+FS against the baseline model with original data, the model of ANOVA feature selection alone and the models that combined three state-of-arts, under-sampling models, with ANOVA feature selection; TL+FS, ENN+FS and RUS+FS. Observing the results of the classifiers from the six different models gives the readers the following insights into the improvement that HDUS+FS achieves:

1- The need for addressing the imbalanced class problem. Despite the progress made by sampling techniques, substantially fewer studies developed feature selection approaches especially to handle the unbalanced class problems. Indeed, a proper selection of relevant features is important for understanding and also for achieving better predictive results. However, if a feature selection is used alone with an imbalanced dataset, it may not produce successful classification results, as indicated by the findings displayed in (4.3). Considerably, when the FS is integrated with techniques to handle class inequality, such as under-sampling, the efficiency of classification can be significantly improved. The superiority of our hybrid model called HDUS+FS, which tackles both imbalanced and overlapping problems, has been displayed in the previous section using our proposed model (HDUS) and a widely used ANOVA f-test. The results confirm that the combination approach is noticeably superior over FS alone for all tested medical datasets (CRCD, PIMAD, THSD, BCD). Therefore, the implementation of proper technique to overcome unequal class problem before feature selection is a primary need to attain acceptable outcomes. More than that, using feature selection with imbalanced medical data is essential to discover the knowledge (e.g., to select important features that easily diagnosis disease).

2- As pointed out previously, implementing sampling technique with FS is more effective, to handle the imbalanced issues in the specific domain, than employing feature selection alone. All the hybrid models used in this work have shown to be valuable, and there is no combination permanently superior to the others. Although our proposed model HDUS+FS outperforms the others in three datasets, it does not seem to be convenient in PIMAD which has moderated data. On the other hand, the most noticeable improvement for our proposed model is with THSD, which has a low value of original minor samples. Therefore, adjusting the distribution of classes to be more balanced has indeed been a good choice if the original number of minor samples is quite poor.

To clarify the comparison between the various hybrid models used in this chapter and to estimate their efficiency, figure 4.2, 4.3 and 4.4 offer a graphical demonstration of the rate for(recall, F1-M and B-AC) resulting from the six methods used in four unbalanced medical data sets. The graphics show that results differ when various under-sampling methods are combined with feature selection. Figure 4.2 shows that HDUS+FS has advanced dramatically in estimating the sensitivity of minor class instances in all data sets. A similar situation can be seen in figure 4.3 for F1-M, that defines trade-offs among recall and precision, also figure 4.4 for B-AC, that defines the trade-offs among specificity and recall.

3- Concerning the classification methods, it should be noted that the advantage of performing classification rises as the unequal class distribution and the overlapping issue is properly tackled. In this chapter, various classifiers can benefit from using HDUS+FS model. For HDUS+FS, the KNN shows the highest results. For RUS+FS, SVM presents top results. While DT is more suitable with both TL+FS and ENN+FS. Table 4.5 shows the performance of the used classifiers with hybrid methods, and four unbalanced data sets have increased the reducibility of the rare class instances by the metrics recall, F1-M and B-AC.

Lastly, the findings of HDUS+FS model can be regarded as an introductory study but an encouraging methodology in addressing the skewed distribution and overlapping problem in the unbalanced data sets to increase the classification efficiency of minor class cases.

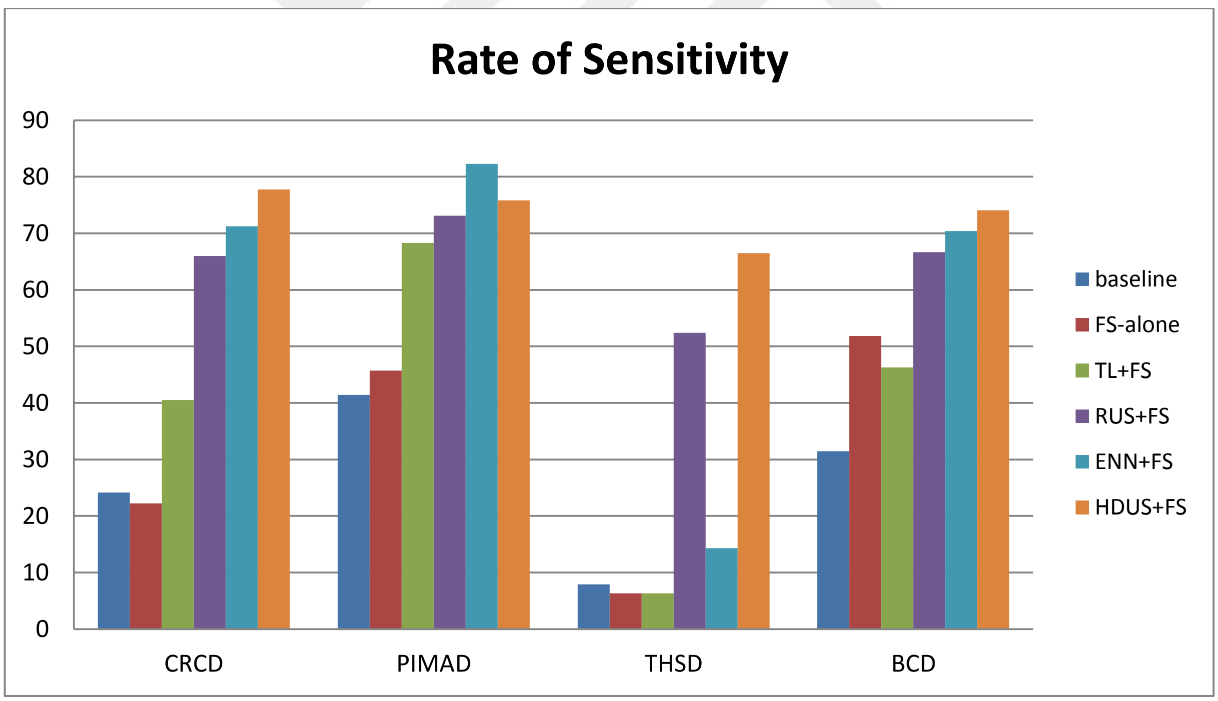


Figure 4.2: A mean rate of sensitivity for six methods (the baseline, FS-alone, and hybrid under-sampling with FS models) on four unbalanced medical datasets.

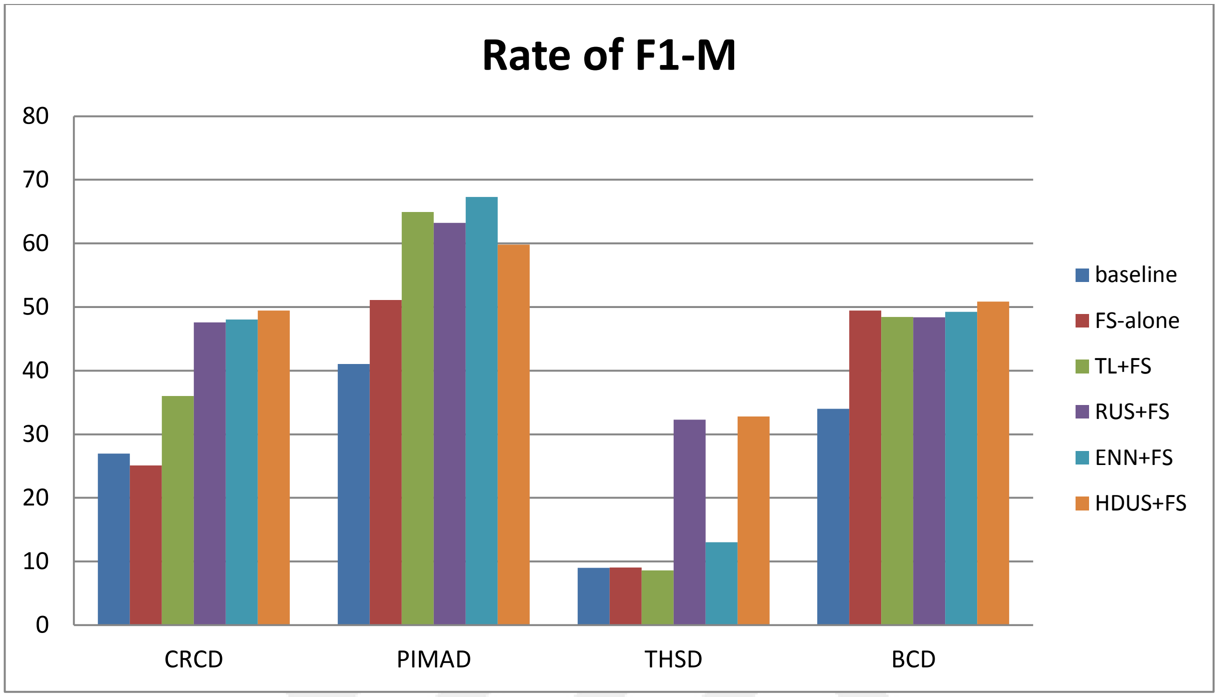


Figure 4.3: A rate of F1-Measure for six methods (the baseline, FS-alone, and hybrid under-sampling with FS models) on four unbalanced medical datasets.

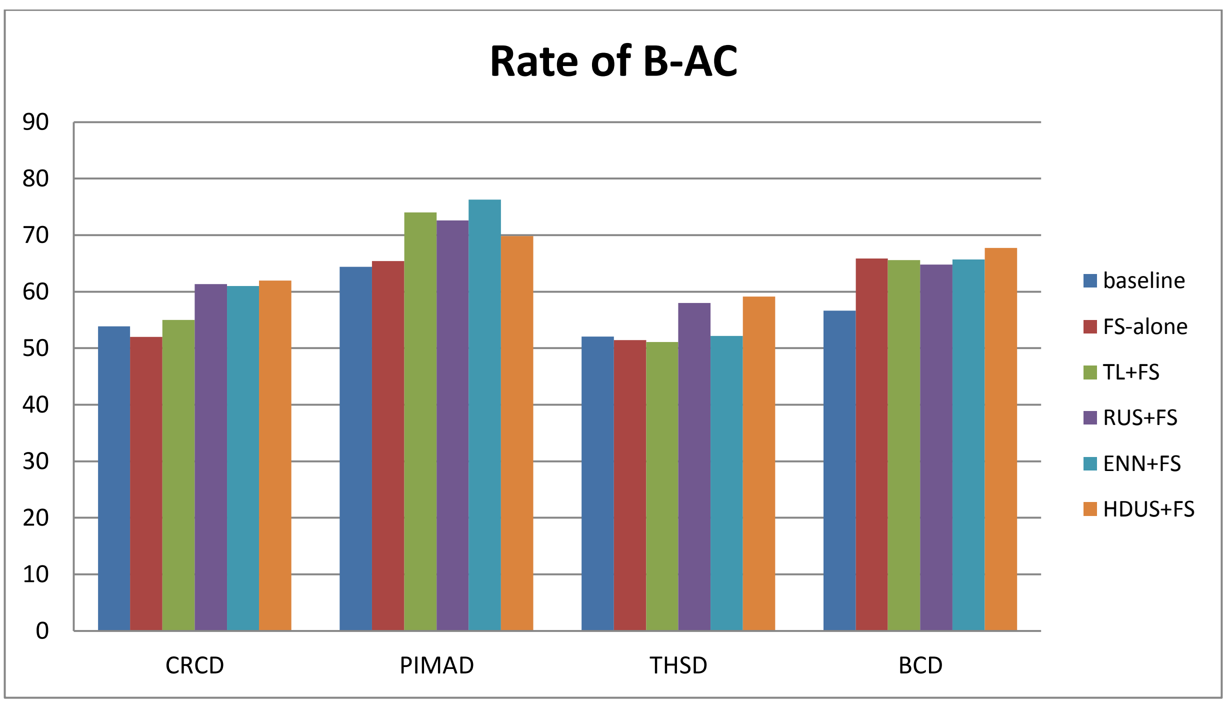


Figure 4.4: A rate of balanced accuracy for six methods (the baseline, FS-alone, and hybrid under-sampling with FS models) on four unbalanced medical datasets.

Table 4.5: A mean rate for three classification algorithms with six methods (the baseline, FS-alone, and hybrid under-sampling with FS models) on four unbalanced medical datasets.

		baseline	FS-alone	TL+FS	RUS+FS	ENN+FS	HDUS+FS
KNN	Sensitivity %	30.2	33.8	42.5	58.7	60.4	78.6
	Specificity %	88.5	83.6	82.5	68.0	68.1	55.2
	Precision %	37.5	41.9	49.1	38.7	36.1	35.7
	F1-M %	33.3	34.3	41.0	46.5	44.3	48.1
	B-AC %	59.3	58.7	62.5	63.3	64.2	66.9
SVM	Sensitivity %	7.0	21.6	34.1	71.8	57.6	77.0
	Specificity %	96.6	94.3	88.4	64.3	71.7	57.4
	Precision %	20.1	42.1	39.3	42.3	34.4	37.1
	F1-M %	10.3	27.8	36.5	52.3	42.8	49.3
	B-AC %	51.8	58.0	61.2	68.1	64.6	67.2
DT	Sensitivity %	41.5	39.1	44.4	63.1	60.5	71.9
	Specificity %	76.5	79.4	76.5	59.0	64.2	57.6
	Precision %	38.4	39.7	39.9	35.1	37.5	37.8
	F1-M %	39.5	38.8	41.0	44.7	46.1	49.2
	B-AC %	59.0	59.2	60.5	61.0	62.3	64.8

5. CONCLUSION AND FUTURE WORK

5.1 CONCLUSION

The main outcome of this study is the development of two under-sampling models to address the unbalanced distribution and overlapping problems in binary data classification. These models were tested on four imbalanced medical datasets (CRCD, PIMAD, THSD, and BCD) using three classifiers (KNN, SVM, and DT).

For the first objective of this research, a new under-sampling model named, HDUS, has been presented to address the problem of the unequal class distribution in the context of an imbalanced dataset that effects on classification efficiency of rare class instances. HDUS works to eliminate the major class samples by employing the Hellinger distance metric to estimate the similarity among each major class sample and minor class samples. Then chooses the samples from the major class that are most resemblance to the minor class samples. The motivation from this process is to keep the minor class samples without change and to select the major samples that are alike with the minor samples. By this simple pre-processing method, three positive effects are accomplished with regard to the rebalancing of the unbalanced data set: 1) The minority samples, which are uncommon occurrences, have been well maintained in number; 2) The process does not generate synthetic samples from any classes so that the total training dataset is controlled without any rise in samples; 3) As the major samples are filtered out other than those that seem identical to the minor samples, the classification efficiency of the minor class would be improved. HDUS performs by maintaining all available minor class instances, that are assumed to be valuable, and by lowering the excessively large number of major class instances, retaining only the major instances that have a close distance to minor instances. The experimental results show that HDUS can overcome the performance of the selective state-of-the-art under-sampling models. HDUS is found to be a useful pre-processing method to fix training datasets containing a few but significant cases of the minority class.

For the second objective of this research, we present a model that combined under-sampling and feature selection, named HDUS+FS. The presented model has been built from a double viewpoint: 1) employing an under-sampling method to adjust the distribution of instances between classes by removing the instances from the major class, that complicated the discrimination between classes; 2) employing a feature selection to eliminate the features that can indicate a high degree of overlap in the boundary region, thereby making difficulties in determining minority samples. The experiments demonstrate the robustness of our proposed HDUS+FS model which outperforms feature selection alone and the combination of some state-of-the-arts-under-sampling with Feature selection.

With the obtained results of the two models (HDUS and HDUS+FS), we can conclude that:

1. Adjusting the distribution of classes to be more balanced has indeed been a good choice if the original number of minor samples is quite poor.
2. It is important to handle the imbalanced class problems by using a suitable approaches before performing classification process.
3. Implementing sampling technique with FS is more effective, to handle the imbalanced issues in the specific domain, than employing feature selection alone
4. Concerning the classification methods, it should be noted that the advantage of performing classification rises as the unbalanced class problems are properly tackled.

5.2 FUTURE WORK

For a future project, several different paths relevant to this study may be continued. Several motivating problems are illustrated as follows:

1. Testing the performance of both models (HDUS and HDUS+FS) on higher complicated datasets with different characteristics, containing a larger number of features and/or instances such as high dimensional datasets.
2. Developing the HDUS model to handle the unbalanced class problem for the multiclass classification.
3. Investigation the algorithm level methods and integrated with the HDUS model.
4. Comparing HDUS with other sampling approaches using various classification methods.
5. In the unbalanced class issue, the ideal class distribution rate between the major class and minor class is still unknown. Additional investigation is required to determine the optimum class distribution rate.

6. Investigating and employing the combination of HDUS with over-sampling technique.
7. Hoping to continue the research study on unbalanced class classification to solve many other complicated issues.
8. Our experiment in the second model, the combination of under-sampling method with feature selection) confirms the importance of this topic for further investigation, we will experiment with another combination of sampling methods with feature selection techniques.

REFERENCES

[1] A. Mahani and A. Riad Baba Ali, "Classification Problem in Imbalanced Datasets," in *Recent Trends in Computational Intelligence*, IntechOpen, 2020, pp. 1–23.

[2] Y. Liu, Y. Wang, X. Ren, H. Zhou, and X. Diao, "A Classification Method Based on Feature Selection for Imbalanced Data," *IEEE Access*, vol. 7, pp. 81794–81807, 2019.

[3] M. A. U. H. Tahir, S. Asghar, A. Manzoor, and M. A. Noor, "A Classification Model For Class Imbalance Dataset Using Genetic Programming," *IEEE Access*, vol. 7, pp. 71013–71037, 2019.

[4] J. Gu, Y. Zhou, and X. Zuo, "Making class bias useful: A strategy of learning from imbalanced data," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4881 LNCS, pp. 287–295, 2007.

[5] F. Feng, K.-C. Li, J. Shen, Q. Zhou, and X. Yang, "Using Cost-Sensitive Learning and Feature Selection Algorithms to Improve the Performance of Imbalanced Classification," *IEEE Access*, vol. 8, pp. 69979–69996, 2020.

[6] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data : Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci. (Ny).*, vol. 250, pp. 113–141, 2013.

[7] G. Cuaya, A. Muñoz-Meléndez, and E. F. Morales, "A Minority Class Feature Selection Method," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7042 LNCS, 2011, pp. 417–424.

[8] M. Denil and T. Trappenberg, "Overlap versus Imbalance," pp. 220–231, 2010.

[9] Liu and Motoda, "Feature Selection for Knowledge Discovery and Data Mining," *SPRINGER Sci. MEDIA*, vol. 454, 1998.

[10] G. P. Wang, J. X. Yang, and R. Li, "Imbalanced SVM-based anomaly detection algorithm for imbalanced training datasets," *ETRI J.*, vol. 39, no. 5, pp. 621–631, 2017.

[11] M. Orooji and J. Chen, "Predicting Louisiana Public High School Dropout through Imbalanced Learning Techniques," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 2019, pp. 456–461.

[12] D. Veganzones and E. Séverin, "An investigation of bankruptcy prediction in imbalanced datasets," *Decis. Support Syst.*, vol. 112, pp. 111–124, Aug. 2018.

[13] C. Huang, Y. Li, C. L. Chen, and X. Tang, "Deep Imbalanced Learning for Face Recognition and Attribute Prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, Jun. 2019.

[14] L. Borrajo, R. Romero, E. L. Iglesias, and C. M. Redondo Marey, "Improving imbalanced scientific text classification using sampling strategies and dictionaries.," *J. Integr. Bioinform.*, vol. 8, no. 3, p. 176, Sep. 2011.

[15] N. F. Hordri, S. Sophiayati, N. Firdaus, and S. Mariyam, "Handling Class Imbalance in Credit Card Fraud using Resampling Methods," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 11, pp. 390–396, 2018.

[16] H. Han, M. Huang, Y. Zhang, and J. Liu, "Decision Support System for Medical Diagnosis Utilizing Imbalanced Clinical Data," *Appl. Sci.*, vol. 8, no. 9, p. 1597, Sep. 2018.

[17] S. Abdellatif, M. A. Ben Hassine, S. Ben Yahia, and A. Bouzeghoub, "ARCID: A New Approach to Deal with Imbalanced Datasets Classification," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10706 LNCS, 2018, pp. 569–580.

[18] K. K. Bejjanki, J. Gyani, and N. Gugulothu, "Class Imbalance Reduction (CIR): A Novel Approach to Software Defect Prediction in the Presence of Class Imbalance," *Symmetry (Basel).*, vol. 12, no. 3, p. 407, Mar. 2020.

[19] V. S. Akondi, V. Menon, J. Baudry, and J. Whittle, "Novel K-Means Clustering-based Undersampling and Feature Selection for Drug Discovery Applications," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2019, no. Mcl, pp.

2771–2778.

[20] K. Yoon and S. Kwek, “A data reduction approach for resolving the imbalanced data issue in functional genomics,” *Neural Comput. Appl.*, vol. 16, no. 3, pp. 295–306, May 2007.

[21] W. Feng, W. Huang, and J. Ren, “Class Imbalance Ensemble Learning Based on the Margin Theory,” *Appl. Sci.*, vol. 8, no. 5, p. 815, May 2018.

[22] A. Ghazikhani, H. S. Yazdi, and R. Monsefi, “Class Imbalance Handling Using Wrapper-based Random Oversampling,”, *In20th Iranian Conference on Electrical Engineering (ICEE2012)*, pp. 611-616,, 2012.

[23] N. Japkowicz and S. Stephen, “The class imbalance problem A systematic study,” , *Intelligent data analysis* ,vol. 6, pp. 429–449, 2002.

[24] R. C. Prati, G. E. A. P. A. Batista, and D. F. Silva, “Class imbalance revisited: a new experimental setup to assess the performance of treatment methods,” *Knowl. Inf. Syst.*, vol. 45, no. 1, pp. 247–270, Oct. 2015.

[25] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, “RUSBoost: A Hybrid Approach to Alleviating Class Imbalance,” *IEEE Trans. Syst. Man, Cybern. - Part A Syst. Humans*, vol. 40, no. 1, pp. 185–197, Jan. 2010.

[26] E. Byon, A. K. Shrivastava, and Y. Ding, “A classification procedure for highly imbalanced class sizes,” *IIE Trans. (Institute Ind. Eng.*, vol. 42, no. 4, pp. 288–303, 2010.

[27] N. Japkowicz, “The Class Imbalance Problem: Significance and Strategies,” *Proc. 2000 Int. Conf. Artif. Intell.*, pp. 111--117, 2000.

[28] N. Japkowicz, “Learning from imbalanced data sets: a comparison of various strategies,” *AAAI Work. Learn. from Imbalanced Data Sets*, pp. 0–5, 2000.

[29] M. Galar, A. Fern, E. Barrenechea, and H. Bustince, “A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches,” IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews vol.

42, no. 4, pp. 463–484, 2011.

[30] S. García, J. Luengo, and F. Herrera, "Data Preprocessing in Data Mining," Cham, Switzerland: Springer International Publishing vol. 72. 2015.

[31] M. Buda, A. Maki, and M. A. Mazurowski, “A systematic study of the class imbalance problem in convolutional neural networks,” *Neural Networks*, vol. 106, pp. 249–259, Oct. 2018.

[32] I. Brown and C. Mues, “An experimental comparison of classification algorithms for imbalanced credit scoring data sets,” *Expert Syst. Appl.*, vol. 39, no. 3, pp. 3446–3453, 2012.

[33] S. García and F. Herrera, “Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy,” *Evol. Comput.*, vol. 17, no. 3, pp. 275–306, 2009.

[34] P. Branco, L. Torgo, and R. P. Ribeiro, “A survey of predictive modeling on imbalanced domains,” *ACM Comput. Surv.*, vol. 49, no. 2, pp. 1–56, 2016.

[35] T. M. Khoshgoftaar and K. Gao, “Feature Selection with Imbalanced Data for Software Defect Prediction,” in *2009 International Conference on Machine Learning and Applications*, 2009, pp. 235–240.

[36] J. Van Hulse, T. M. Khoshgoftaar, A. Napolitano, and R. Wald, “Feature Selection with High-Dimensional Imbalanced Data,” in *2009 IEEE International Conference on Data Mining Workshops*, 2009, pp. 507–514.

[37] Y. Saeys, I. Inza, and P. Larrañaga, “A review of feature selection techniques in bioinformatics.,” *Bioinformatics*, vol. 23, no. 19, pp. 2507–17, Oct. 2007.

[38] S. Maldonado, R. Weber, and F. Famili, “Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines,” *Inf. Sci. (Ny).*, vol. 286, pp. 228–246, 2014.

[39] M. Wasikowski and X. Chen, “Combating the Small Sample Class Imbalance Problem Using Feature Selection,” *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1388–1400,

Oct. 2010.

[40] Y. SUN, A. K. C. WONG, and M. S. KAMEL, “CLASSIFICATION OF IMBALANCED DATA: A REVIEW,” *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 04, pp. 687–719, Jun. 2009.

[41] W.-C. Lin, C.-F. Tsai, Y.-H. Hu, and J.-S. Jhang, “Clustering-based undersampling in class-imbalanced data,” *Inf. Sci. (Ny).*, vol. 409–410, pp. 17–26, Oct. 2017.

[42] S. J. Yen and Y. S. Lee, “Cluster-based under-sampling approaches for imbalanced data distributions,” *Expert Syst. Appl.*, vol. 36, no. 3 PART 1, pp. 5718–5727, 2009.

[43] G. M. Weiss and F. Provost, “Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction,” *J. Artif. Intell. Res.*, vol. 19, pp. 315–354, Oct. 2003.

[44] S. Visa and A. Ralescu, “The effect of imbalanced data class distribution on fuzzy classifiers - Experimental study,” *IEEE Int. Conf. Fuzzy Syst.*, pp. 749–754, 2005.

[45] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, “Classification with class imbalance problem: A review,” *Int. J. Adv. Soft Comput. its Appl.*, vol. 5, no. 3, pp. 176–204, 2013.

[46] M. M. Nwe and K. T. Lynn, “KNN-Based Overlapping Samples Filter Approach for Classification of Imbalanced Data,” in *Studies in Computational Intelligence*, vol. 845, Springer International Publishing, pp. 55–73, 2020.

[47] E. Szmidt and M. Kukier, “Classification of Imbalanced and Overlapping Classes using Intuitionistic Fuzzy Sets,” *In2006 3rd International IEEE Conference Intelligent Systems,* pp. 722–727, September, 2006.

[48] A.Fernández, S.García, and F.Herrera, “Addressing the Classification with Imbalanced Data : Open Problems and New Challenges on Class Distribution,” *In International conference on hybrid artificial intelligence systems , Springer, Berlin, Heidelberg.* pp. 1–10, 2011.

[49] V. López, A. Fernández, J. G. Moreno-torres, and F. Herrera, "Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics Victoria," Expert Systems with Applications, vol. 39, pp. 6585–6608, 2012.

[50] P.Vuttipittayamongkol, E.Elyan,A.Petrovski and C.Jayne, "Overlap-Based Undersampling for Improving Imbalanced Data Classification," , In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 689-697). Springer, November, 2018.

[51] L. Chen, B. Fang, Z. Shang, and Y. Tang, "Tackling class overlap and imbalance problems in software defect prediction," *Softw. Qual. J.*, vol. 26, no. 1, pp. 97–125, 2018.

[52] C.-F. Tsai, W.-C. Lin, Y.-H. Hu, and G.-T. Yao, "Under-sampling class imbalanced datasets by combining clustering analysis and instance selection," *Inf. Sci. (Ny).*, vol. 477, pp. 47–54, Mar. 2019.

[53] P. Vorraboot, S. Rasmequan, K. Chinnasarn, and C. Lursinsap, "Improving classification rate constrained to imbalanced data between overlapped and non-overlapped regions by hybrid algorithms," *Neurocomputing*, vol. 152, pp. 429–443, Mar. 2015.

[54] T. Jo and N. Japkowicz, "Class imbalances versus small disjuncts," *ACM SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 40–49, Jun. 2004.

[55] C. Drummond and R. C. Holte, "C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling," *Work. Learn. from Imbalanced Datasets II*, pp. 1–8, 2003.

[56] M. M. Rahman and D. N. Davis, "Addressing the Class Imbalance Problem in Medical Datasets," *Int. J. Mach. Learn. Comput.*, vol. 3, no. 2, pp. 224–228, 2013.

[57] R. Guermazi, I. Chaabane, and M. Hammami, "AECID: Asymmetric entropy for classifying imbalanced data," *Inf. Sci. (Ny).*, vol. 467, no. August, pp. 373–397, 2018.

[58] J. K. B, W. Kowalczyk, S. Menzel, and B. Thomas, "Improving Imbalanced Classification by Anomaly Detection," *In International Conference on Parallel Problem*

Solving from Nature, Springer, Cham. vol. 1, no. 766186, pp. 512–523, 2020.

[59] B. S. Raghuwanshi and S. Shukla, "UnderBagging based reduced Kernelized weighted extreme learning machine for class imbalance learning," *Eng. Appl. Artif. Intell.*, vol. 74, no. June, pp. 252–270, 2018.

[60] Y. Liu, Y. Wang, X. Ren, H. Zhou, and X. Diao, "A Classification Method Based on Feature Selection for Imbalanced Data," *IEEE Access*, vol. 7, pp. 81794–81807, 2019.

[61] R. Anand, K. G. Mehrotra, C. K. Mohan, and S. Ranka, "An improved algorithm for neural network classification of imbalanced training sets," *IEEE Trans. Neural Networks*, vol. 4, no. 6, pp. 962–969, 1993.

[62] R. Guermazi, I. Chaabane, and M. Hammami, "AECID: Asymmetric entropy for classifying imbalanced data," *Inf. Sci. (Ny).*, vol. 467, no. March 2019, pp. 373–397, 2018.

[63] J. Mathew, M. Luo, C. K. Pang, and H. L. Chan, "Kernel-based SMOTE for SVM classification of imbalanced datasets," in *IECON 2015 - 41st Annual Conference of the IEEE Industrial Electronics Society*, 2015, pp. 001127–001132.

[64] Y. Tang, Y.-Q. Zhang, N. V Chawla, and S. Krasser, "SVMs modeling for highly imbalanced classification.," *IEEE Trans. Syst. Man. Cybern. B. Cybern.*, vol. 39, no. 1, pp. 281–8, Feb. 2009.

[65] X. Gu, T. Ni, and H. Wang, "New fuzzy support vector machine for the class imbalance problem in medical datasets classification.," *ScientificWorldJournal.*, vol. 2014, p. 536434, 2014.

[66] Xiangrong Zhang, Qiang Song, Yaoguo Zheng, Biao Hou, and Shuiping Gou, "Classification of imbalanced hyperspectral imagery data using support vector sampling," in *2014 IEEE Geoscience and Remote Sensing Symposium*, pp. 2870–2873, 2014.

[67] C. O. Truică and C. A. Leordeanu, "Classification of an imbalanced data set using decision tree algorithms," *UPB Sci. Bull. Ser. C Electr. Eng. Comput. Sci.*, vol. 79, no. 4, pp. 69–84, 2017.

[68] Y. Li and X. Zhang, “Improving k nearest neighbor with exemplar generalization for imbalanced classification,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6635 LNAI, no. PART 2, pp. 321–332, 2011.

[69] W. Liu and S. Chawla, “Class confidence weighted kNN algorithms for imbalanced data sets,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6635 LNAI, no. PART 2, pp. 345–356, 2011.

[70] S. Zhang, “KNN-CF Approach: Incorporating Certainty Factor to kNN Classification.,” *IEEE Intell. Informatics Bull.*, vol. 11, no. 1, pp. 24–33, 2010.

[71] S. Liu, P. Zhu, and S. Qin, “An Improved Weighted KNN Algorithm for Imbalanced Data Classification,” in *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, 2018, pp. 1814–1819.

[72] C. Zhang, J. Song, Z. Pei, and J. Jiang, “An Imbalanced Data Classification Algorithm of De-noising Auto-Encoder Neural Network Based on SMOTE,” *MATEC Web Conf.*, vol. 56, no. 1, pp. 95–99, 2016.

[73] C. Zhang, K. C. Tan, H. Li, and G. S. Hong, “A Cost-Sensitive Deep Belief Network for Imbalanced Classification,” *IEEE Trans. Neural Networks Learn. Syst.*, vol. 30, no. 1, pp. 109–122, Apr. 2018.

[74] O. Octavio Loyola-Gonzalez, J. F. C. O. Martinez-Trinidad, J. A. Carrasco-Ochoa, and M. Garcia-Borroto, “Cost-Sensitive Pattern-Based classification for Class Imbalance problems,” *IEEE Access*, vol. 7, pp. 60411–60427, 2019.

[75] X. Jiang, S. Pan, G. Long, F. Xiong, J. Jiang, and C. Zhang, “Cost-Sensitive Parallel Learning Framework for Insurance Intelligence Operation,” *IEEE Trans. Ind. Electron.*, vol. 66, no. 12, pp. 9713–9723, Dec. 2019.

[76] G. Weiss, K. McCarthy, and B. Zabar, “Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?,” *Dmin*, pp. 1–7, 2007.

[77] G. Weiss, “Mining with rarity: A unifying framework.,” *ACM Sigkdd Explor. Newsl.*, vol. 6, no. 1, pp. 7–19, 2004.

[78] H. He *et al.*, “Ensemble MultiBoost Based on RIPPER Classifier for Prediction of Imbalanced Software Defect Data,” *IEEE Access*, vol. 7, pp. 110333–110343, 2019.

[79] M. Naseriparsa, “Combination of PCA with SMOTE Resampling to Boost the Prediction Rate in Lung Cancer Dataset,” . arXiv preprint arXiv, vol. 77, no. 3, 1403.1949. 2014.

[80] J. Xiao, L. Xie, C. He, and X. Jiang, “Dynamic classifier ensemble model for customer classification with imbalanced class distribution,” *Expert Syst. Appl.*, vol. 39, no. 3, pp. 3668–3675, Feb. 2012.

[81] Q.-Y. Yin, J.-S. Zhang, C.-X. Zhang, and N.-N. Ji, “A Novel Selective Ensemble Algorithm for Imbalanced Data Classification Based on Exploratory Undersampling,” *Math. Probl. Eng.*, vol. 2014, no. ii, pp. 1–14, 2014.

[82] S. Huda, J. Yearwood, H. F. Jelinek, M. M. Hassan, G. Fortino, and M. Buckland, “A Hybrid Feature Selection With Ensemble Classification for Imbalanced Healthcare Data: A Case Study for Brain Tumor Diagnosis,” *IEEE Access*, vol. 4, pp. 9145–9154, 2016.

[83] N. Liu, X. Li, E. Qi, M. Xu, L. Li, and B. Gao, “A Novel Ensemble Learning Paradigm for Medical Diagnosis With Imbalanced Data,” *IEEE Access*, vol. 8, pp. 171263–171280, 2020.

[84] M. Zhu *et al.*, “Class Weights Random Forest Algorithm for Processing Class Imbalanced Medical Data,” *IEEE Access*, vol. 6, pp. 4641–4652, 2018.

[85] C. Y. Wang, L. L. Hu, M. Z. Guo, X. Y. Liu, and Q. Zou, “imDC: an ensemble learning method for imbalanced classification with miRNA data.,” *Genet. Mol. Res.*, vol. 14, no. 1, pp. 123–33, Jan. 2015.

[86] K. Polat, “Similarity-based attribute weighting methods via clustering algorithms in the classification of imbalanced medical datasets,” *Neural Comput. Appl.*, vol. 30, no. 3, pp. 987–1013, Aug. 2018.

[87] W. Jindaluang, V. Chouvatut, and S. Kantabutra, “Under-sampling by Algorithm with Performance Guaranteed for Class-imbalance Problem,” *Int. Comput. Sci. Eng. Conf.*, pp. 215–221, 2014.

[88] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, “Learning from class-imbalanced data: Review of methods and applications,” *Expert Syst. Appl.*, vol. 73, no. September 2017, pp. 220–239, 2017.

[89] J. Mathew, C. K. Pang, M. Luo, and W. H. Leong, “Classification of Imbalanced Data by Oversampling in Kernel Space of Support Vector Machines,” *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 9, pp. 4065–4076, 2018.

[90] E. Rendón, R. Alejo, C. Castorena, F. J. Isidro-Ortega, and E. E. Granda-Gutiérrez, “Data sampling methods to dealwith the big data multi-class imbalance problem,” *Appl. Sci.*, vol. 10, no. 4, 2020.

[91] F. Ren, P. Cao, W. Li, D. Zhao, and O. Zaiane, “Ensemble based adaptive over-sampling method for imbalanced data learning in computer aided detection of microaneurysm,” *Comput. Med. Imaging Graph.*, vol. 55, pp. 54–67, 2017.

[92] H. Guan, Y. Zhang, M. Xian, and H. D. C. Xianglong, “SMOTE-WENN : Solving class imbalance and small sample problems by oversampling and distance scaling,” Applied Intelligence, pp. 1-16, 2020.

[93] S. Cateni, V. Colla, and M. Vannucci, “A method for resampling imbalanced datasets in binary classification tasks for real-world problems,” *Neurocomputing*, vol. 135, pp. 32–41, Jul. 2014.

[94] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[95] R. Blagus and L. Lusa, “SMOTE for high-dimensional class-imbalanced data,” *BMC Bioinformatics*, vol. 14, no. 1, p. 106, Dec. 2013.

[96] R. Blagus and L. Lusa, “Evaluation of SMOTE for High-Dimensional Class-Imbalanced

Microarray Data," in *2012 11th International Conference on Machine Learning and Applications*, 2012, vol. 2, no. 1, pp. 89–94.

[97] A. Wosiak and S. Karbowiak, "Preprocessing compensation techniques for improved classification of imbalanced medical datasets," in *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, FedCSIS 2017*, 2017, vol. 11, pp. 203–211.

[98] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, p. 27, Dec. 2019.

[99] R. F. A. B. De Morais and G. C. Vasconcelos, "Under-Sampling the Minority Class to Improve the Performance of Over-Sampling Algorithms in Imbalanced Data Sets," *In Proceedings of International Joint Conference on Artificial Intelligence*, September, 2017.

[100] R. Batuwita and V. Palade, "Efficient resampling methods for training support vector machines with imbalanced datasets," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*, 2010, pp. 1–8.

[101] P. P. E. S. K. Dimitris Kanellopoulos, "Handling imbalanced datasets: A review," *Int. Trans. Comput. Sci. Eng.*, vol. 30, Dec. 2006.

[102] X. Guo, Y. Yin1 , C. Dong, G. Yang and G. Zhou, "On the Class Imbalance Problem," In *2008 Fourth international conference on natural computation*, IEEE, vol. 4, pp. 192-201., 2008.

[103] C. K. Aridas, S. Karlos, V. G. Kanas, N. Fazakis, and S. B. Kotsiantis, "Uncertainty Based Under-Sampling for Learning Naive Bayes Classifiers Under Imbalanced Data Sets," *IEEE Access*, vol. 8, pp. 2122–2133, 2020.

[104] S. Cateni, V. Colla, and M. Vannucci, "A method for resampling imbalanced datasets in binary classification tasks for real-world problems," *Neurocomputing*, vol. 135, pp. 32–41, 2014.

[105] J. Lu, C. Zhang, and F. Shi, “A classification method of imbalanced data base on PSO algorithm,” *Commun. Comput. Inf. Sci.*, vol. 624, pp. 121–134, 2016.

[106] A. Moreo, A. Esuli, and F. Sebastiani, “Distributional random oversampling for imbalanced text classification,” *SIGIR 2016 - Proc. 39th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 805–808, 2016.

[107] C. W. Yeh, D. C. Li, L. S. Lin, and T. I. Tsai, “A learning approach with under-and over-sampling for imbalanced data sets,” *Proc. - 2016 5th IIAI Int. Congr. Adv. Appl. Informatics, IIAI-AAI 2016*, pp. 725–729, 2016.

[108] J. Yun, J. Ha, and J. S. Lee, “Automatic determination of neighborhood size in SMOTE,” *ACM IMCOM 2016 Proc. 10th Int. Conf. Ubiquitous Inf. Manag. Commun.*, 2016.

[109] A. Al-Shahib, R. Breitling, and D. Gilbert, “Feature Selection and the Class Imbalance Problem in Predicting Protein Function from Sequence,” *Appl. Bioinformatics*, vol. 4, no. 3, pp. 195–203, 2005.

[110] N. Garcia-Pedrajas, J. Pérez-Rodríguez, and A. de Haro-García, “OligoIS: Scalable Instance Selection for Class-Imbalanced Data Sets,” *IEEE Trans. Cybern.*, vol. 43, no. 1, pp. 332–346, Feb. 2013.

[111] A. de Haro-García, G. Cerruela-García, and N. García-Pedrajas, “Instance selection based on boosting for instance-based learners,” *Pattern Recognit.*, vol. 96, p. 106959, Dec. 2019.

[112] M. Blachnik, “Instance Selection for Classifier Performance Estimation in Meta Learning,” *Entropy*, vol. 19, no. 11, p. 583, Nov. 2017.

[113] G. Yu, J. Tian, and M. Li, “Nearest neighbor-based instance selection for classification,” in *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 2016, pp. 75–80.

[114] M. Kubat and S. Matwin , “Addressing the Curse of Imbalanced Training Sets : One-

Sided Selection," *In Icml,*.vol. 97, pp. 179–186, 1997.

[115] J. Li, S. Fong, S. Hu, R. K. Wong, and S. Mohammed, "Similarity Majority Under-Sampling Technique for Easing Imbalanced Classification Problem," in *Communications in Computer and Information Science*, vol. 845, Springer Singapore, 2018, pp. 3–23.

[116] V. & Elyan, "Overlap-Based Undersampling Method for Classification of Imbalanced Medical Datasets," in *IFIP International Federation for Information Processing*, 2020, vol. 584, pp. 358–369.

[117] A. Kumari and U. Thakar, "Hellinger distance based oversampling method to solve multi-class imbalance problem," in *2017 7th International Conference on Communication Systems and Network Technologies (CSNT)*, 2017, pp. 137–141.

[118] G. Du, J. Zhang, Z. Luo, F. Ma, L. Ma, and S. Li, "Joint imbalanced classification and feature selection for hospital readmissions," *Knowledge-Based Syst.*, vol. 200, p. 106020, 2020.

[119] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: Special Issue on Learning from Imbalanced Data Sets," *ACM SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 1–6, 2004.

[120] H. Liu, M. Zhou, and Q. Liu, "An embedded feature selection method for imbalanced data classification," *IEEE/CAA J. Autom. Sin.*, vol. 6, no. 3, pp. 703–715, May 2019.

[121] Y. Wang, Y. Liu, L. Feng, and X. Zhu, "Novel feature selection method based on harmony search for email classification," *Knowledge-Based Syst.*, vol. 73, no. 1, pp. 311–323, 2015.

[122] I. Koprinska, M. Rana, and V. G. Agelidis, "Correlation and instance based feature selection for electricity load forecasting," *Knowledge-Based Syst.*, vol. 82, pp. 29–40, 2015.

[123] X. W. Chen and M. Wasikowski, "FAST: A roc-based feature selection metric for small samples and imbalanced data classification problems," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 124–132, 2008.

[124] Z. Liu, D. Tang, Y. Cai, R. Wang, and F. Chen, "A hybrid method based on ensemble WELM for handling multi class imbalance in cancer microarray data," *Neurocomputing*, vol. 266, pp. 641–650, Nov. 2017.

[125] L. Yin, Y. Ge, K. Xiao, X. Wang, and X. Quan, "Feature selection for high-dimensional imbalanced data," *Neurocomputing*, vol. 105, pp. 3–11, 2013.

[126] M. Alibeigi, S. Hashemi, and A. Hamzeh, "DBFS: An effective Density Based Feature Selection scheme for small sample size and high dimensional imbalanced data sets," *Data Knowl. Eng.*, vol. 81–82, pp. 67–103, 2012.

[127] Z. Liu, R. Wang, M. Tao, and X. Cai, "A class-oriented feature selection approach for multi-class imbalanced network traffic datasets based on local and global metrics fusion," *Neurocomputing*, vol. 168, pp. 365–381, 2015.

[128] L. A. S. Mark A. Hall, "Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper," in *FLAIRS conference*, 1999, pp. 235–239.

[129] N. Almugren and H. Alshamlan, "A survey on hybrid feature selection methods in microarray gene expression data for cancer classification," *IEEE Access*, vol. 7, pp. 78533–78548, 2019.

[130] S. Beniwal and J. Arora, "Classification and Feature Selection Techniques in Data Mining," *International journal of engineering research & technology (ijert)*, vol. 1, no. 6, pp. 1–6, 2012.

[131] S.-Y. Ohn, Y.-M. Park, C. W. Kim, T.-N. Vu, H.-N. Nguyen, and M. Y. Han, "Feature Elimination Approach Based on Random Forest for Cancer Diagnosis," *InMexican International Conference on Artificial Intelligence,* Springer, Berlin, Heidelberg, pp. 532–542, 2006.

[132] Y. B. Wah, N. Ibrahim, H. A. Hamid, S. Abdul-Rahman, and S. Fong, "Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy," *Pertanika J. Sci. Technol.*, vol. 26, no. 1, pp. 329–340, 2018.

[133] J. A. Olvera-López, J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad, and J. Kittler, "A review of instance selection methods," *Artif. Intell. Rev.*, vol. 34, no. 2, pp. 133–143, Aug. 2010.

[134] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics.," *Bioinformatics*, vol. 23, no. 19, pp. 2507–17, Oct. 2007.

[135] T. A. Abdallah and B. de La Iglesia, "Survey on Feature Selection," *Procedia Comput. Sci.*, vol. 91, no. Itqm, pp. 919–926, Oct. 2015.

[136] D. Jain and V. Singh, "Feature selection and classification systems for chronic disease prediction: A review," *Egypt. Informatics J.*, vol. 19, no. 3, pp. 179–189, 2018.

[137] M. Cover T and E. Hart P, "Nearest Neighbor Pattern Classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[138] M. B. Rodrigues *et al.*, "Health of Things Algorithms for Malignancy Level Classification of Lung Nodules," *IEEE Access*, vol. 6, pp. 18592–18601, 2018.

[139] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 20–29, Jun. 2004.

[140] and I. Z. Mani, Inderjeet, "kNN Approach to Unbalanced Data Distributions: A Case Study involving Information Extraction," in *Proceedings of workshop on learning from imbalanced datasets*, 2003.

[141] A. K. Shukla, "Feature selection inspired by human intelligence for improving classification accuracy of cancer types," *Comput. Intell.*, no. April, p. coin.12341, Jun. 2020.

[142] Y. Peng, 2013 "Tikz example – SVM trained with samples from two classes,". [Online]. Available: http://blog.pengyifan.com/tikz-example-svm-trained-with-samples-from-two-classes/

[143] CHRISTOPHER J.C. BURGES, "A Tutorial on Support Vector Machines for Pattern

Recognition," *Data Min. Knowl. Discov.*, vol. 2, pp. 121–167, 1998.

[144] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[145] M. Çınar, M. Engin, E. Z. Engin, and Y. Ziya Ateşçi, "Early prostate cancer diagnosis by using artificial neural networks and support vector machines," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 6357–6361, Apr. 2009.

[146] M.-W. Huang, C.-W. Chen, W.-C. Lin, S.-W. Ke, and C.-F. Tsai, "SVM and SVM Ensembles in Breast Cancer Prediction," *PLoS One*, vol. 12, no. 1, p. e0161501, Jan. 2017.

[147] Y. Peng, 2013 "Tikz example – Kernel trick," [Online]. Available: http://blog.pengyifan.com/%0Atikz-example-kernel-trick/.

[148] R. Akbani, S. Kwek, and N. Japkowicz, "Applying Support Vector Machines to Imbalanced Datasets," *Eur. Conf. Mach. Learn.*, pp. 39–50, 2004.

[149] F. Vilariño, P. Spyridonos, J. Vitrià, and P. Radeva, "Experiments with SVM and Stratified Sampling with an Imbalanced Problem: Detection of Intestinal Contractions," in *Lecture Notes in Computer Science*, vol. 3687, no. PART II, 2005, pp. 783–791.

[150] G. Wu and E. Y. Chang, "Class-boundary alignment for imbalanced dataset learning," *ICML Work. Learn. from Imbalanced Data Sets II*, pp. 49–56, 2003.

[151] E. Aličković and A. Subasi, "Breast cancer diagnosis using GA feature selection and Rotation Forest," *Neural Comput. Appl.*, vol. 28, no. 4, pp. 753–763, Apr. 2017.

[152] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986.

[153] Y. Huang, Y. Jin, Y. Li, and Z. Lin, "Towards Imbalanced Image Classification: A Generative Adversarial Network Ensemble Learning Method," *IEEE Access*, vol. 8, pp. 88399–88409, 2020.

[154] T. Elhassan, M. Aljurf M, and M. shoukri, "Classification of Imbalance Data using

Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method," *Glob. J. Technol. Optim.*, vol. 01, no. S1, 2016.

[155] M. H. Popel, K. M. Hasib, S. A. Habib, and F. M. Shah, "A Hybrid Under-Sampling Method (HUSBoost) to Classify Imbalanced Data," *2018 21st Int. Conf. Comput. Inf. Technol. ICCIT 2018*, no. December, pp. 1–7, 2018.

[156] R. M. Pereira, Y. M. G. Costa, and C. N. Silla, "MLTL: A multi-label approach for the Tomek Link undersampling algorithm: MLTL: The Multi-Label Tomek Link," *Neurocomputing*, vol. 383, pp. 95–105, 2020.

[157] D. Devi, S. kr Biswas, and B. Purkayastha, "Redundancy-driven modified Tomek-link based undersampling: A solution to class imbalance," *Pattern Recognit. Lett.*, vol. 93, pp. 1339–1351, 2017.

[158] M. Abdar, M. Zomorodi-Moghadam, R. Das, and I.-H. Ting, "Performance analysis of classification algorithms on early detection of liver disease," *Expert Syst. Appl.*, vol. 67, pp. 239–251, Jan. 2017.

[159] Q. Wei and R. L. Dunbrack, "The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics," *PLoS One*, vol. 8, no. 7, p. e67863, Jul. 2013.

[160] R. J. Lyon, J. M. Brooke, J. D. Knowles, and B. W. Stappers, "Hellinger Distance Trees for Imbalanced Streams," in *2014 22nd International Conference on Pattern Recognition*, 2014, pp. 1969–1974.

[161] P. Harsha, "Hellinger Distance," in *Wiley StatsRef: Statistics Reference Online*, vol. 2011, Chichester, UK: John Wiley & Sons, Ltd, 2014, pp. 1–8.

[162] A. Shemyakin, "Hellinger Distance and Non-informative Priors," *Bayesian Anal.*, vol. 9, no. 4, pp. 923–938, Dec. 2014.

[163] V. González-Castro, R. Alaiz-Rodríguez, and E. Alegre, "Class distribution estimation based on the Hellinger distance," *Inf. Sci. (Ny).*, vol. 218, pp. 146–164, Jan. 2013.

[164] A. Dal Pozzolo, R. Johnson, O. Caelen, S. Waterschoot, N. V. Chawla, and G. Bontempi,

"Using HDDT to avoid instances propagation in unbalanced and evolving data streams," in *2014 International Joint Conference on Neural Networks (IJCNN)*, 2014, pp. 588–594.

[165] C. Rao, "A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance," *Questiió: Quaderns d'Estadística, Sistemes, Informatica i Investigació Operativa*, vol. 19, no. 1. pp. 23–63, 1995.

[166] M. Schmid, T. Welchowski, M. N. Wright, and M. Berger, "Discrete-time survival forests with Hellinger distance decision trees," *Data Min. Knowl. Discov.*, vol. 34, no. 3, pp. 812–832, May 2020.

[167] R. Guermazi, I. Chaabane, and M. Hammami, "AECID: Asymmetric entropy for classifying imbalanced data," *Inf. Sci. (Ny).*, vol. 467, no. March 2019, pp. 373–397, Oct. 2018.

[168] D. A. Cieslak, T. R. Hoens, N. V. Chawla, and W. P. Kegelmeyer, "Hellinger distance decision trees are robust and skew-insensitive," *Data Min. Knowl. Discov.*, vol. 24, no. 1, pp. 136–158, Jan. 2012.

[169] G.-H. Fu, Y.-J. Wu, M.-J. Zong, and J. Pan, "Hellinger distance-based stable sparse feature selection for high-dimensional class-imbalanced data.," *BMC Bioinformatics*, vol. 21, no. 1, p. 121, Mar. 2020.

[170] "UCI machine learning repository." [Online]. Available: https://archive.ics.uci.edu/ml/datasets.php.

[171] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, 2016.

[172] S. Alshomrani, A. Bawakid, S. O. Shim, A. Fernández, and F. Herrera, "A proposal for evolutionary fuzzy systems using feature weighting: Dealing with overlapping in imbalanced datasets," *Knowledge-Based Syst.*, vol. 73, pp. 1–17, 2015.

[173] H. K. Lee and S. B. Kim, "An overlap-sensitive margin classifier for imbalanced and overlapping data," *Expert Syst. Appl.*, vol. 98, pp. 72–83, 2018.

[174] Z. Z. R. Al-Shamaa, S. Kurnaz, A. D. Duru, N. Peppa, A. H. Mirnezami, and Z. Z. R. Hamady, “The Use of Hellinger Distance Undersampling Model to Improve the Classification of Disease Class in Imbalanced Medical Datasets,” *Appl. Bionics Biomech.*, vol. 2020, pp. 1–10, Nov. 2020.

[175] M. Bejani, D. Gharavian, and N. M. Charkari, “Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks,” *Neural Comput. Appl.*, vol. 24, no. 2, pp. 399–412, 2014.

[176] R. G. Shaw and T. Mitchell-Olds, “ANOVA for unbalanced data: an overview,” *Ecology*, vol. 74, no. 6, pp. 1638–1645, 1993.

[177] M. Sheikhan, M. Bejani, and D. Gharavian, “Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method,” *Neural Comput. Appl.*, vol. 23, no. 1, pp. 215–227, 2013.

[178] C.-Y. Lee and Z.-J. Lee, “A novel algorithm applied to classify unbalanced data,” *Appl. Soft Comput.*, vol. 12, no. 8, pp. 2481–2485, Aug. 2012.

[179] M. Kumar, N. K. Rath, A. Swain, and S. K. Rath, “Feature Selection and Classification of Microarray Data using MapReduce based ANOVA and K-Nearest Neighbor,” *Procedia Comput. Sci.*, vol. 54, pp. 301–310, 2015.

CURRICULUM VITAE

Zina Al Shamaa: received the B.Sc. degree in Computer science from University of Technology in Iraq, 1991, and the M.Sc. degree in Computer Information System from Middle East University in Amman, 2012. She is currently pursuing the Ph.D. degree in Computer science at Altinbas University in Turkey. Since 2004, she has been a programmer in computer center, General Directorate of Educational Planning, Ministry of Education, Baghdad, Iraq. Her research interests include databases, machine learning and data mining, data science and big data and artificial intelligence.

PUBLICATION:

1- Published paper "The Use of Hellinger Distance Undersampling Model to Improve the Classification of Disease Class in Imbalanced Medical Datasets," *Appl. Bionics Biomech.*, vol. 2020, pp. 1–10, Nov. 2020.

2- Preparing second paper to publish "A hybrid under-sampling and feature selection model to maximize the performance of rare class in imbalanced datasets"