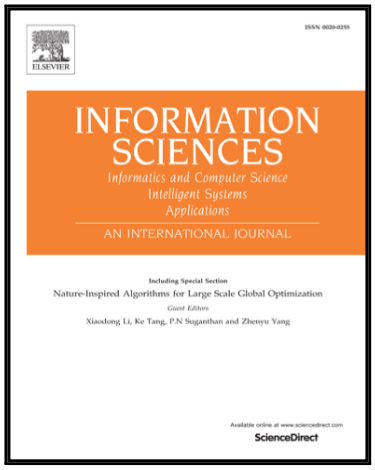
Journal Pre-proof



Data Imbalance in Classification: Experimental Evaluation

Fadi Thabtah , Suhel Hammoud, Firuz Kamalov,
Amanda H. Gonsalvesv

PII: S0020-0255(19)31049-7
DOI: https://doi.org/10.1016/j.ins.2019.11.004
Reference: INS 14996

To appear in: *Information Sciences*

Received date: 5 June 2019
Revised date: 4 November 2019
Accepted date: 8 November 2019

Please cite this article as: Fadi Thabtah , Suhel Hammoud, Firuz Kamalov, Amanda H. Gonsalvesv, Data Imbalance in Classification: Experimental Evaluation, *Information Sciences* (2019), doi: https://doi.org/10.1016/j.ins.2019.11.004

Data Imbalance in Classification: Experimental Evaluation

Fadi Thabtah[a,1,*], Suhel Hammoud[b], Firuz Kamalov[c], Amanda H. Gonsalvesv[d]

[a]*Manukau Institute of Technology*
[b]*University of Kalamoon*
[c]*Canadian University Dubai*
[d]*Manukau Institute of Technology*

Abstract

The advent of Big Data has ushered a new era of scientific breakthroughs. One of the common issues that affects raw data is class imbalance problem which refers to imbalanced distribution of values of the response variable. This issue is present in fraud detection, network intrusion detection, medical diagnostics, and a number of other fields where negatively labeled instances significantly outnumber positively labeled instances. Modern machine learning techniques struggle to deal with imbalanced data by focusing on minimizing the error rate for the majority class while ignoring the minority class. The goal of our paper is demonstrate the effects of class imbalance on classification models. Concretely, we study the impact of varying class imbalance ratios on classifier accuracy. By highlighting the precise nature of the relationship between the degree of class imbalance and the corresponding effects on classifier performance we hope to help researchers to better tackle the problem. To this end, we carry out extensive experiments using 10-fold cross validation on a large number of datasets. Our analysis confirms that class imbalance has a significant negative impact on the performance of a classifier. In particular, we determine that the relationship between the class imbalance ratio and the accuracy is convex. We also find that addressing the issue of class imbalance in the pre-training phase can have a substantial positive impact on the accuracy of a classifier.

Keywords: Classification, Class Imbalance, Data Analysis, Machine Learning, Statistical Analysis, Supervised Learning.

1. Introduction

In machine learning, classifiers are derived to minimize misclassification errors and thereby maximize predictive accuracy (Boyle, 2018). The underlying assumption in these classification methods is that the dataset under study has a roughly balanced number of instances per available

*I am corresponding author
Email addresses: Fadi.fayez@manukau.ac.nz (Fadi Thabtah), suhel.hammoud@uok.edu.sy (Suhel Hammoud), firuz@cud.ac.ae (Firuz Kamalov), gons16@manukaumail.com (Amanda H. Gonsalvesv)
[1]Auckland, New Zealand

class. In other words, it assumes that the prior probabilities of the target classes are similar (Guo, Yin, Dong, Yang,& Zhou, 2008). However, in many real-world domains, such as medical diagnostics, most of the classification data tends to be skewed towards negative class value. Data is said to be imbalanced if at least one of the target variable values has a significantly smaller number of instances when compared to the other values. Class imbalance is especially prevalent in models used to detect rare but important diseases such as autism spectrum disorder (Thabtah, 2018a; Thabtah, Kamalov & Rajab, 2018).

Longadge, Dongre & Malik (2013) define class imbalance problems in terms of skewness. The authors state that this problem occurs when a dataset is severely skewed which can result in a high rate of False Negatives (FN). Imbalanced class is an intrinsic issue that exists in a wide array of real-world applications such as fraud detection (Nazrul, 2018), text classification (Liu, Loh, Youcef-Toumi, & Tor, 2007), face and image recognition (Huang, Li, Loy, & Tang, 2018), and medical diagnosis (Belarouci & Chikh, 2017; Thabtah 2018b; Thabtah & Peebles 2019). Ouyang, Chen, & Wei (2017) highlight the effect of imbalanced classes on an oil spill detection system. This system uses synthetic aperture radar (SAR) to detect and prevent the environmental impact of oil spillages and deter illegal dumping. However, with only 10% of the spills originating from the sea beds, the database contains far less images of oil spills than those without oil spills. This creates further challenges associated with interpreting the results of data processing (Guo et al., 2008).

Classification performance of an algorithm is affected when the dataset under study is highly imbalanced, (Boyle, 2018; Buda, 2017; Anwar, 2012; Chawla, Bowyer, Hall & Kegelmeyer, 2002). Traditionally, classification algorithms are driven to increase the predictive accuracy of the derived classifiers. However, maximizing the overall accuracy may not be the best approach in case of an imbalanced dataset. While maximizing the overall accuracy a classifier focuses on the majority class as it has the higher weight in the data. As a result the classifier can achieve a high degree of accuracy on the majority class, and by extension on the overall dataset, all the while performing poorly on the minority set. It is worth noting that identifying the minority instances is often of greater importance as in the case of rare diseases. Thus, the accuracy metric would no longer be a proper evaluation measure and the derived classifiers may produce misleading information, especially with regards to the minority class (Yang & Wu, 2006; Chawla 2005).

The effects of class imbalance on the performance of convolutional neural networks was studied by Buda et al. (2018) who find that class imbalance leads to decreased performance of neural networks. The authors also conclude that the best method to combat class imbalance is oversampling and that the optimal oversampling rate is at 50%. Zhu et al (2017) investigated the problem of class imbalance in the context of customer churn prediction. Concretely, the authors use expected maximum profit criterion as one of the main performance measures and discover a lot of room for improvement in the profit based measure. The use of a particular evaluation metric is also found to have an impact on the solution. Prati et al. (2015) put forth an experimental design that aims to assess the performance of class imbalance treatment methods. In their study the authors show that low class imbalance ratios do not result in significant performance loss. However, the performance loss increases quickly for highly imbalanced data.

Our goal is to study the impact of changing class imbalance ratios on the accuracy of a classifier. It is certainly well known that class imbalance causes problems for classifiers. There are

many studies on the topic, such as: Ouyang et al. (2017); Ali, Shamsuddin, & Ralescu (2015); Hulse, Khoshgoftaar, & Napolitano (2007). However, only a limited number of experimental studies have considered the impact of varying class imbalance ratios on classifiers. The exact nature of the relationship between the degree of class imbalance and the classifier performance has not been satisfactorily studied. Our contribution is to highlight this relationship in more detail. To this end, we aim to perform a comprehensive study on the effects of modifying class imbalance ratios on classifier performance. Our study is based on the analysis of a number of different datasets using a range of performance metrics. In order to properly account for class skewness the evaluation metrics used in the paper include error rate, precision, recall, and area under ROC curve (AUC). The new information presented in the paper about the nature of the relationship between the degree of class imbalance and the corresponding classifier performance will help researchers to better understand the class imbalance problem.

A key aspect of our study is the investigation of class imbalance in the context of behavioural science. Although the issue of class imbalance is common in behavioural disorder diagnostics its effects have not been properly investigated. To fill this gap in knowledge we undertake the study of class imbalance in diagnosis of autism spectrum disorder (ASD). In particular, we analyse the impact of class imbalance on ASD detection systems by considering varying imbalance ratios and the corresponding classifier performance. Furthermore, we study various resampling techniques to combat class imbalance in ASD data. To summarize, our paper aims to accomplish the following research goals:

- Carry out a comprehensive study on the impact of varying class imbalance ratios on classifier performance using a wide range of evaluation metrics.
- Investigate the issue of class imbalance in the context of ASD and the effectiveness of various remedial techniques.

We used the probabilistic Nave Bayes as the base classifier in our experiments (Duda & Hart, 1973). All the experiments are performed using ten-fold cross validation. In order to study the effects of varying the imbalance ratio, the performance of classifiers on each dataset was recorded based on nine different ratios of the binary response variable namely 10%:90%, 20%:80%, 30%:70%, 40%:60%, 50%:50%, 60%:40%, 70%:30%, 80%:20%, 90%:10%. Section 4 contains further details about the experimental set up, methods used, and analysis of the results.

The paper is organized as follows: Section 2 outlines the problem and provides common examples from different applications. Section 3 reviews common techniques that handle the class imbalance problem such as under-sampling and oversampling. Section 3 is devoted to the description of the computational intelligence method. In Section 4, the data features, experiments, and results analysis are explained and Section 5 concludes the paper.

2. The Problem and Examples

To better understand the class imbalance problem consider an autism diagnosis dataset consisting of 970 instances not on the spectrum (controls) and 30 instances on the spectrum (cases). A classifier trained on this dataset will typically predict a test case to belong to No Autism since

the vast majority of training datasets are linked with controls. In fact, the nave strategy of guessing any new individual as not having autism would produce 97% accuracy. However, considering the nature of this behavioural application, it is more important to predict individuals who are on the spectrum so they can seek appropriate help. Early diagnosis could potentially slow down any further development of the disorder and speed up access to the necessary healthcare services. Since the dataset in our example is significantly imbalanced typical metrics such as error rate or predictive accuracy would not be suitable to evaluate the performance of classifiers. In the presence of an imbalanced class in a dataset, the classifier gives more importance to the majority class and the rare instances either go undiscovered, ignored, or assumed to be noisy data (Ali et al., 2015). This increases the misclassifications of the minority class, which is at times more crucial, particularly in medical informatics applications. For example, in the early diagnosis of coronary heart disease (CHD), misclassifying a non-CHD person as CHD will lead to additional medical tests. However, misclassifying a person with CHD as non-CHD will not only alter the course of the treatment, but can also lead to complications and prove fatal. As a result, it is important that the imbalanced class problem in classification is addressed early. Anomaly detection is closely related to imbalanced class as the members of the minority class can be treated as outliers and dealt accordingly (Abdelhamid, Ayesh & Thabtah, 2013).

Fraud detection is another important source of imbalanced data. The most common fraud analysis is associated with detecting credit card fraud. Fraud detection is performed by checking the company's transactions database. However, these datasets have many more legitimate cases than fraudulent ones making the prediction problem of fraudulent activities challenging (AlShboul Thabtah, Abdelhamid & Al-diabat, 2018). Another key element that affects the performance of the classifier is the sample size of the data (Chawla, 2005). A dataset with a low frequency class creates challenges in discovering patterns for the minority class which in turn hinders the classification of test data associated with the minority class (Sun, Wong, & Kamel, 2009). In a study by Japkowicz & Stephen (2002), the authors reported that the error rate of minority class instances can be reduced when adding more synthetic instances to that class. Another study by Chalwa (2005), revealed that creating new instances for the minority class using oversampling may stabilise the performance of classifiers.

3. Common Approaches for Addressing Data Imbalance

Various methods have been developed to handle the class imbalance problem that can be broadly divided into two main approaches: Data driven and algorithm driven. The former attempts to balance the class distribution (Buda, 2017). The latter approach adjusts the learning algorithm, or the classifier without amending the training set (Chawla, 2005). In the next sub-sections, we briefly review these approaches and reveal their advantage and disadvantages.

3.1. Data Driven Approaches

Data driven approaches adjust the class ratio in the input dataset to achieve balanced class distribution. This approach often employs sampling techniques like under-sampling, oversampling, or a combination of both (Ali et al, 2015).

3.1.1. Under-sampling

The primary under-sampling technique that arbitrarily eliminates examples of the majority class to balance the dataset is known as random under-sampling (Kotsiantis, Kanellopoulos, & Pintelas, 2006). Researchers such as Guo et al. (2008) reveal the primary drawback associated with this method to be the disposal of useful information that can prove to be crucial in the later classification stages. Many under-sampling approaches have been proposed such as Condensed Nearest Neighbour Rule (CNN), Wilsons Edited Nearest Neighbour Rule (ENN), Neighbourhood Cleaning Rule (NCL), Tomek Links, and One-sided selection (OSS) (Hart, 1968; Tomek, 1976; Wilson, 1972; Laurikkala, 2001).

Wilson (1972), proposes the CNN method in which if the class label of any example is different from the class of at least two of its nearest three neighbours, then the example is eliminated from the dataset. Tomek (1976) proposed the Tomek Links method, which is a modified CNN that considers only the data points close to the boundary to be important. In this method two samples, E_i and E_j, belonging to two different classes are considered together with the distance $D(E_i, E_j)$ between them. The Tomek Link method states that if there does not exist an example E_0 such that $D(E_i, E_0) < D(E_i, E_j)$ or $D(E_j, E_0) < D(E_i, E_j)$ then pair (E_i, E_j) is a Tomek Link. The formation of a Tomek Link implies the existence of a noisy example or else examples on the borderline (Anwar, 2012). The majority class examples in such Tomek Links are removed to address the class imbalance problem.

Laurikkala (2001) proposes the NCL method which utilizes three nearest neighbours for each example (E_i) in the training dataset. When E_i belongs to the majority class, and its three nearest neighbours have a different class, then E_i is removed. However, if E_i belongs to the minority class and its three nearest neighbours are different than those neighbours that belong to the majority class, then they are removed. The drawback of this method is that it can lead to an enormous number of computations when the input dataset is large and comprises a large number of majority class instances (Chawla, 2005). Anwar (2012) highlights another drawback in the ENN and NCL methods, which is class overlapping. The author states that when the degree of class overlapping in the training set is high, then the classifier built will be poor in terms of predictive accuracy, as the majority class examples situated close to the decision boundaries will be deleted. Tomek Links is similar to NCL in that it may lead to performance deterioration of the classifiers when the input dataset is large and with a high degree of overlapping among class values.

Johnson & Khoshgoftaar (2019) surveyed different approaches to class imbalance related to Artificial Neural Networks The authors studied unstructured datasets such as images and text. Their research addressed issues related to experimental design, implementation and evaluation of deep learning techniques in applications that are linked with data imbalance such as medical diagnosis and fraud detection, among others. Challenges related to complexity, output format, results and learning mechanisms have been discussed in the survey especially within the context of deep learning. The authors concluded that the class imbalance problem has not been sufficiently covered within deep learning and believed their approach could help improve the performance when the input data is imbalanced. Katip (2019) proposed an under-sampling technique based on consensus clustering to improve performance of imbalanced datasets. In using this approach, observations linked with the dominant class in the training dataset are under-sampled using the consensus clustering algorithm. Several different small imbalanced datasets and two large datasets were utilized

to test the new under-sampling technique. In the experiments, different classification algorithms including Nave Bayes and Regression among others have been used to derive classifiers and show the impact of the proposed under-sampling technique. The results reported by the classification algorithms when using the consensus clustering method in the pre-processing phase yielded better predictive accuracy on the datasets when compared to not using sampling techniques.

3.1.2. Oversampling

Oversampling is another common sampling approach used to deal with an imbalanced class problem. Various oversampling strategies are available including random oversampling, focused oversampling, and synthetic sampling (Estabrooks, Jo, & Japkowicz, 2004; Chawla et al., 2002; Kubat & Matwin, 1997). The method in which the instances of the minority class are randomly replicated until they have equal representation is known as random oversampling (Buda, 2017). Random oversampling has two major shortcomings: it increases the possibility of overfitting of the classifier on the training dataset, and if the original data already has high dimensionality, it mounts the computation cost thus increasing the training time of the classifier (Chawla, 2005). In focused oversampling, only those minority class values with samples occurring on the boundary between the majority and minority class values are resampled. However, Chawla et al., (2002) state that these methods of oversampling by replication lead to a more specific decision region of the minority class. To overcome this and broaden this decision region, they propose an advanced heuristic oversampling technique called the Synthetic Minority Oversampling Technique (SMOTE). SMOTE is a technique in which oversampling of the minority class is carried out by generating synthetic examples. These new synthetic minority class examples are the result of interpolation between closely located minority class samples. Chawla et al., (2002) describe the process of SMOTE as calculating the nearest same-class neighbours for every minority example and then based on the required oversampling rate, randomly choosing from these neighbouring examples. The synthetic examples are then generated at random points along the line segments joining the minority examples with these chosen neighbours. This process expands the decision region pertaining to the minority class. The authors tested SMOTE on a wide variety of datasets with different training sizes and degree of imbalance. It was observed that the larger and less specific decision regions of SMOTE resulted in this method outperforming oversampling by replacement.

3.2. Algorithm-Driven Approach

In this approach, the classification algorithm is adjusted to facilitate the learning task specifically with respect to the minority class. In this approach, no changes are made to the input data distribution. This approach includes cost sensitive learning, thresholding, and hybrid methods like ensemble learners among others.

3.2.1. Cost-Sensitive Learning

The basic assumption by any traditional classification algorithm is that there is a cost of misclassification is the same for all response variable values (Chawla, 2005; Fan, Stolfo, Zhang, & Chan, 1999). However, in real world applications the cost of misclassification may vary among target class values. For example, in the field of medical diagnosis predicting that a person suffers from a particular disease when he does not will lead to more medical examinations. On other hand,

misclassifying a person as healthy when he actually suffers from the disease can lead to excessive health costs and potential fatality. In this situation, the cost of misclassification is higher, so the classifier ought to be sensitive to the type of application and more importantly the cost associated with misclassification (Radwan, 2017; Buda, 2017). To address this issue, cost-sensitive learning was proposed, so if the underlying dataset has a class ratio of 1:3 favouring the majority class, then the misclassifying cost for the minority class will be three times that of the majority class (Japkowicz & Stephen, 2002). An important concept in cost-sensitive learning is the Cost Matrix (Buda, 2017) which comprises the average cost of misclassification for each available class. The goal of this method is to minimize misclassification cost by choosing a class with minimum conditional risk (Kotsiantis et al., 2006). The layout of the cost matrix is depicted in Table 1 below, where λ_{ij} is the cost of misclassifying a sample from True Class j to Class i.

Table 1: Cost Matrix

		Prediction	
		Class i	Class j
True	Class i	0	λ_{ij}
	Class j	λ_{ji}	0

The diagonal entries indicate correct classification, hence the cost associated with this is zero. Other common cost-sensitive learning models are MetaCost which is a wrapper method of making the classifier cost sensitive, and AdaCost which is a cost-sensitive version of AdaBoost (Radwan, 2017; Guo et al., 2008).

3.2.2. Thresholding Methods

Thresholding is a method used to address the problem of class imbalance in classification at the algorithmic level. Buda (2017) defines the set of methods used in a classifier to adjust the decision boundary as threshold moving or the post-scaling method. Certain algorithms, like Nave Bayes, generate probability estimates that are converted into predictions by cutting off at the threshold which is usually 0.5 (Radwan, 2017). The predicted class label depends on the associated probability lying above or below the threshold. However, when the class is imbalanced the performance of many classifiers is affected and thus the decision threshold of the classifiers needs to be adjusted such that the cost of misclassifications is reduced, and the prediction performance is improved.

Zou, Xie, Lin, Wu & Ju, (2016) obtain the decision threshold by greedily adjusting the F-score. The derived probabilities are evaluated to calculate the training set threshold with the best F-score using cross validation. Thereby, the threshold for the test data is determined using Equation 1

$$Thresholdtest = (\frac{Maxtrain - Thresholdtrain}{Maxtrain - Mintrain}) \times (Maxtest - Mintest) \quad (1)$$

3.2.3. Hybrid Methods

Ensemble learning approach, which is a combination of multiple classifiers derived from the training dataset samples, is another alternative to deal with data imbalance. Thereafter, the results of these classifiers are combined to get the final decision of the classification (Yan, Liu, Jin &

Hauptmann, 2003). The underlying success is that the base classifiers often have a diverse principle involved in its construction. Each component classifier is built on certain assumptions and has certain characteristics. However, when these components are combined, the individual variances are averaged out thus leading to a reduction in the variance of the ensemble and improvement in its generalization ability and performance (Sun et al., 2009). AdaBoost is one common ensemble learning method which achieves variance reduction (Freund & Schapire 1997). Sun et al., (2009) describes AdaBoost as a combination of up sampling and down sampling. At the onset, the weights for all the classes are set the same. Every time the classifier misclassifies a test instance, the weight associated with the misclassified class is increased. The weights assigned thus are representative of the importance of the specific class. This approach focuses the attention of the weak learner on important instances of the training dataset that often tend to get classified incorrectly, thereby increasing its classification accuracy. Friedman, Hastie, & Tibshirani (2000) show that AdaBoost is immune to overfitting and has the capability of reducing bias. These features of AdaBoost make it apt for dealing with the class imbalance problem. When the weights added are made cost sensitive, a new system called AdaCost is developed which produces lower cost of misclassification and thus performs better than AdaBoost (Fan et al., 1999).

SMOTEBoost is another combination of boosting learners and sampling technique. It focuses on increasing the classification accuracy of the minority class while keeping intact the accuracy of the underlying dataset (Buda, 2017). Unlike the oversampling methods which tend to overfit the data, this method adds new synthetic examples of the minority class using SMOTE at each boosting iteration (He & Garcia, 2009). The resulting final classifier has a broad decision region for the minority class and is consequently well-defined (He & Garcia, 2009). SMOTEBoost is shown to outperform boosting and the AdaCost when tested over several datasets (Chawla, 2005).

4. Experimental Analysis

This section discusses the setting of the experiments, datasets, and classification methods used to generate the results. The primary aim of this section is to investigate the effects of varying imbalance ratios on classifier performance. The task can be accomplished by conducting thorough experiments on multiple classification datasets using different class ratios. For each dataset, multiple samples with different class ratios is produced, then for each sample a classification method will be applied to derive a classifier. Finally, the performance of these classifiers is contrasted according to different evaluation metrics to seek the behaviour of these results in different class ratio scenarios.

4.1. Data and Experimental Setting

The experiments with different class ratios are implemented in Java and then integrated into the Waikato Environment for Knowledge Analysis Environment (WEKA 3.9) to facilitate analysis of the results. All classifier experiments are also conducted in WEKA 3.9 environment (Hall, et al., 2009). The experiments have been run on a computing machine with 2.5 Ghz processor and 8 GB RAM. We employ Nave Bayes as the base classifier for our experiments due to is efficiency and reliability (Duda & Hart, 1973). Nave Bayes is a probabilistic model that calculates the likelihood of a class label given a feature vector based on simplified Bayes theorem. Concretely, the Nave

Bayes method assumes conditional independence of feature values given the class label which allows for simplified probability calculations. Thus, there is no need to build a complete classification model. Nave Bayes is shown to perform competitively against more complex methods. The Nave Bayes algorithm implemented in this research assumes a multinomial distribution and it uses the default values of the WEKA platform.

We measure the performance of the classifiers derived by Nave Bayes on multiple versions of each dataset considered and according to various class ratios. The outcome of these performance measurements will provide us with a clear picture on the overall impact of class imbalance on the classification dataset. Every experiment is carried out using 10-fold cross validation (Kohavi, 1995). Cross validation consists of splitting the data into 10 equal parts with 9 parts used in the training phase and the remaining part employed in testing. The same process is repeated 10 times to produce the average error rate of the 10 runs.

We employ five datasets, available from the University of California Irvine (UCI) repository, to carry out our numerical experiments. The datasets are chosen to represent a diverse array of applications. Table 2 depicts the selected datasets characteristics including number of attributes, number of examples, any missing values, any continuous value, and class imbalance ratio. To implement our study datasets with different degrees of imbalance with respect to the target class are selected. We perform 100 runs per class ratio for each dataset. The results are averaged using different samples of 10-fold cross validation. In other words, for a single class ratio of a dataset 100 different runs were conducted using stratified 10-fold cross validation. We report the average error rate, predictive accuracy, recall, and precision results among others, derived by the Nave Bayes classification algorithm. The same process is repeated across different class ratios for all datasets. The class ratios considered for are 10%:90%, 20%:80%, 30%:70%, 40%:60%, 50%:50%, 60%:40%, 70%:30%, 80%:20%, and 90%:10%. These ratios are generated based on experimental shuffling of different random seeds per given dataset with stratification.

Table 2: The Datasets Description

	Dataset	# of ex	# of attr	# of labels	Missing	Cont	Ratio (%)
0	Cleve	303	11	2	Yes	No	54.45:45.55
1	Colic	368	22	2	Yes	Yes	63.04:36.06
2	Credit-German	1000	20	2	No	Yes	70:30
3	Diabetes	768	8	2	No	Yes	65.10:34.90
4	Hepatitis	155	19	2	Yes	Yes	79.35:20.65

We measure the classifier performance based on a set of evaluation metrics derived from the binary confusion matrix presented in Table 3. Concretely, we employ True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) counts as basis for computing the error rate, accuracy, recall and precision of the classifier. TP, FP, FN, and TN are explained briefly below:

- TP: Number of examples that are predicted to be positive which are actually positive.
- FP: Number of examples that are predicted to be positive which are actually negative.

- FN: Number of examples that are predicted to be negative which are actually positive.
- TN: Number of examples that are predicted to be negative which are actually negative.

Table 3: Binary confusion matrix

	Predicted class	
True class	Positive	Negative
Positive	TP	FN
Negative	FP	TN

The evaluation measures adopted are

- *Error rate*: The fraction of incorrectly classified instances:

$$Error\ rate = 1 - \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \tag{2}$$

- *Predictive Accuracy*: The fraction of correctly classified instances:

$$Predictive\ accuracy = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \tag{3}$$

- *Recall*: The proportion of total relevant outcomes correctly predicted by the learning algorithm:

$$Recall = \frac{t_p}{t_p + f_n} \tag{4}$$

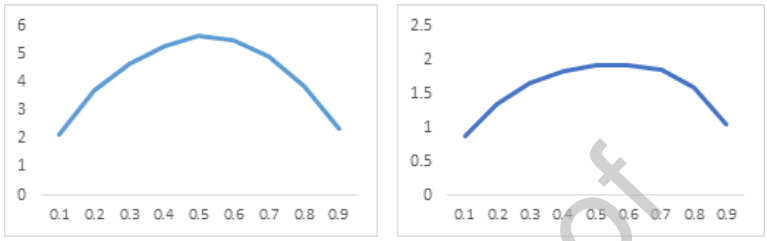
- *Precision*: The proportion of the outcomes that are relevant:

$$Precision = \frac{t_p}{t_p + f_p} \tag{5}$$

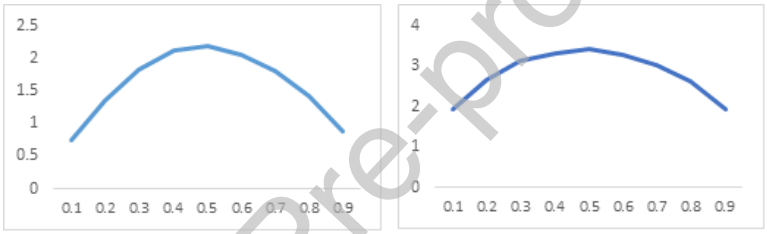
4.1.1. Results Analysis

The impact of class imbalance on the error rate for the surveyed datasets is shown in Figures 1a-1f. The x-axis in the figures indicates the fraction of the first class label in the sampled dataset. The y-axis indicates the mean error rate of the Nave Bayes classifier. The mean error rate is calculated based on 100 runs using 10-fold stratified cross validation. As shown in Figures 1a-1f, the results reveal the high impact of imbalance of the response variable on the error rate the classifier. It is apparent in the results that when the dataset is unbalanced with respect to the class label, the error rate of the derived classifiers is lower than when the same dataset is balanced (50%:50%). In Figures 1a-1c, the error rate is at maximum when the class ratio is 50%:50%. Furthermore, the error rates are at the minimum when the class ratio is 10%:90% and then start to increase exponentially until the ratio reaches 50%:50%. Afterwards, the error rate starts to decrease until it again reaches the minimum when the class ratio becomes 90%:10%. However, Figure 1c depicts

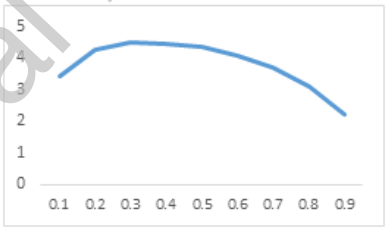
that there is slight change in the behaviour of the error rate as it reaches its maximum value when the class ratio is 30%:70%; it then starts to decrease until it reaches its minimum value at the class ratio of 90%:10%. In fact, the more the dataset is imbalanced, the better the performance. These results are likely due to the high fraction of the majority class which has a positive impact on the overall error rate while not necessarily accurate on the minority class. These results, if limited, reveal that class imbalance impacts the performance of the classifiers derived in terms of error rate.



(a) Credit-German and Cleve Datasets



(b) Pima and Hepatitis Datasets

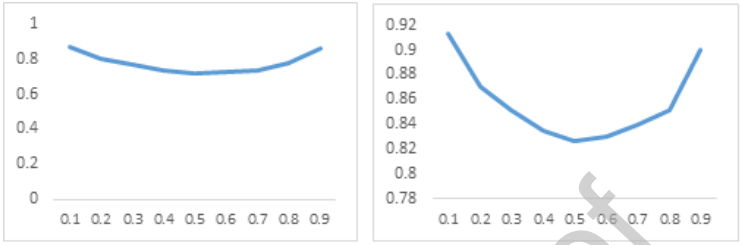


(c) Horse Colic Dataset

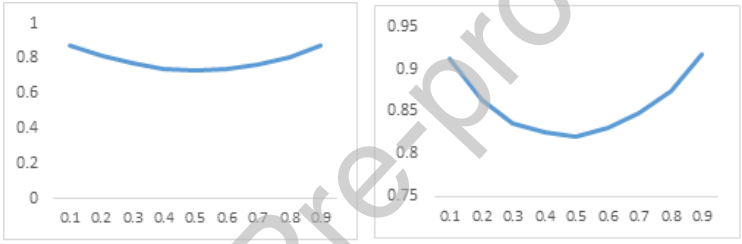
Figure 1: Error Rate Results for Different Class Percentages

Figures 2a-2c depict precision results obtained from the surveyed datasets using different class ratios. The results demonstrate that the highest precision rate is obtained when either the class ratios in the dataset is 90%:10% or 10%:90%, and the lower precision rate is achieved when the dataset is totally balanced, i.e. 50%:50%. The results are consistent across all the datasets. For example, in the Cleve dataset the precision drops from 91.32% to 86.99% to 85.09% as the class ratio increases from 10%:90% to 20%:80% to 30%:70%. The decrease of precision continues until the dataset becomes balanced at the class ratio 50%:50%. At this ratio, the precision is at its minimum, i.e. 82.27%. Afterwards, the precision rate starts to increase when the first class fraction becomes 60%. The increasing pattern continues until the first class fraction becomes 90%. At that

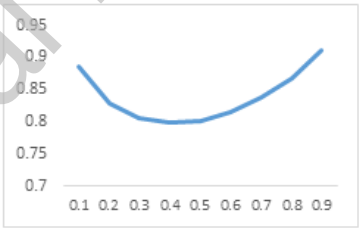
class ratio, the precision derived is 89.97%. The same pattern in terms of relationship between precision and class ratios is observed in the remaining datasets with slight differences in the degree of decrement or increment. To be exact, in the Pima and Credit-German datasets, the increase or decrease of precision rate with respect to class ratios is not highly significant when compared to the remaining datasets.



(a) Credit-German and Cleve Datasets



(b) Pima and Hepatitis Datasets

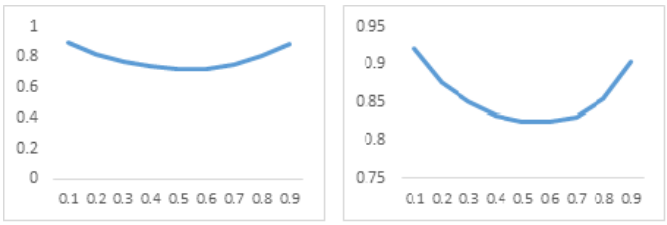


(c) Horse Colic Dataset

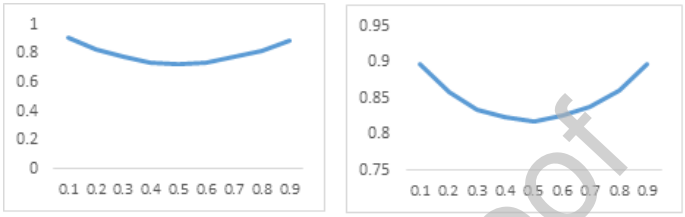
Figure 2: Precision Results for Different Class Percentages

Figures 3a-3c show the recall results obtained from the surveyed datasets. The reported results are consistent with error rate and precision results obtained earlier and clearly reveal that class ratio affects the performance of the derived classifiers. For example, in the Hepatitis dataset, the recall rate is around 90% when the response variable is highly imbalanced, i.e. when class ratios are 10%:90% and 90%:10%. On the other hand, the recall rate is the lowest at 81.74% when the data sample is balanced at 50%:50%. The same pattern can be seen in all considered datasets with the exception for the Horse-colic dataset in which the lowest recall obtained is at the class ratio of 30%:70%.

Figure 4 shows the weighted AUC results obtained on two of the surveyed datasets - Credit-



(a) Credit-German and Cleve Datasets



(b) Pima and Hepatitis Datasets

(c) Horse Colic Dataset

Figure 3: Recall Results for Different Class Percentages

German and Hepatitis - over a range of class ratios. The weighted AUC results obtained from the Hepatitis dataset are above 0.9. The results are consistent across different class ratios which points to high accuracy. On the other hand, the weighted AUC results obtained by the Nave Bayes classifier on the German-Credit dataset are only acceptable as they are around 80%.

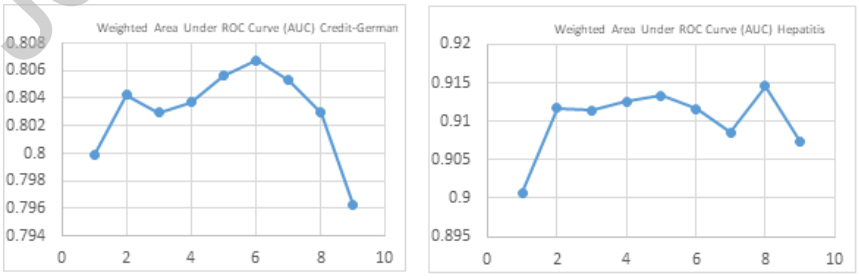


Figure 4: Weighted AUC Results for the Credit- German and Hepatitis Datasets

4.1.2. Autism Detection Application Results with Resampling Techniques

In this section, we consider an imbalanced medical dataset related to Autism Spectrum Disorder (ASD), which refers to a range of neurodevelopmental disorders that are characterised by slowed or restricted development (Thabtah, 2017b). The core symptoms of ASD are typically associated with the ability to communicate and use other methods of social interaction with others. Screening tests can be used to identify individuals that may have autistic traits. Typically, a screening test is in a form of questionnaire whereby the individual (or someone acting on their behalf) answers questions before receiving a score derived from their answers. The data used in our study is taken from the ASDTest (Thabtah, 2017a). The app has four questionnaires based on the age range of the individual taking the test. The chosen dataset is related to the Adult ASD test, for individuals 17 years and older.

Table 4 presents all the features of the autism dataset under consideration. A1 A10 represent the ten questions in the screening test related to Autism Quotient (AQ-10) method (Allison, et al., 2012). Each question has four possible answers: Definitely Agree, Slightly Agree, Slightly Disagree and Definitely Disagree. Depending on the answer each question is scored as 1 or 0 (1 being associated with autistic traits). A final score is derived from the 10 questions with the score of six or more indicating the presence of autistic traits. The original dataset contains 23 attributes, 6 of which are omitted in our experimentation as they are less significant and have no direct impact on the learning process of the classifier (Thabtah, et al., 2018). The total score is also omitted as it is directly correlated with the class prediction, and its inclusion may result in an overfitted model. The dataset includes 1118 instances, 53.3% of which are male and 46.7% are females. The median age of an individual in the dataset is 28. Nearly half (47%) of the individuals identify as White European ethnicity, followed by middle eastern and Asian, which make up 19% and 15% respectively. There are 358 instances associated with class YES (exhibiting ASD symptoms) which represents 32% of the dataset. The dataset is, therefore, deemed imbalanced with respect to the class label.

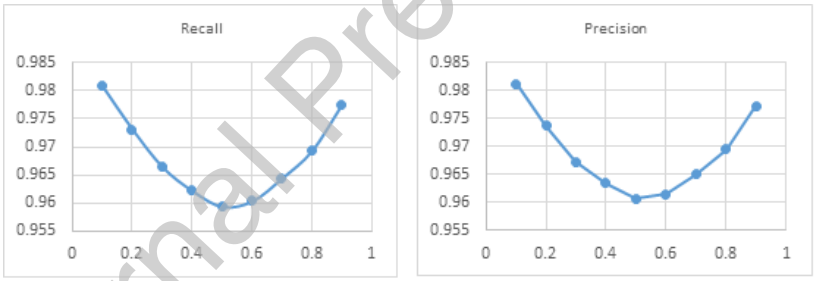
Figure 5a depicts recall and precision rates over a range of different class ratios derived by Nave Bayes classification algorithm from the Adult Autism dataset. The figures clearly demonstrate the largest recall and precision when the dataset is highly imbalanced particularly at 10%:90% and 90%:10% ratios. The performance of the derived classifier maintains high recall and precision rates of 97% even when the dataset is completely balanced at 50%:50%. The difference in results on recall and precision rates between the highly imbalanced (10%:90%) and balanced (50%:50%) sets is around 2%. Comparing these findings with the previous results derived by the same algorithm from the five classification datasets in Section 4.2.2, we discover the Adult Autism dataset to be less sensitive to data imbalance. Nevertheless, we observe similar patterns in terms of sensitivity and precision as before.

The weighted AUC is plotted in Figure 5b over a range of different class ratios for the Adult Autism dataset. The results of weighted AUC clearly show high accuracy in the ASD classification. In particular, the AUC is around 0.99 regardless of the class ratio used for data processing. We attribute these findings to having enough cases and controls in the Adult Autism dataset which enables the Nave Bayes algorithm to develop a more generalizable model that is less sensitive to data imbalance.

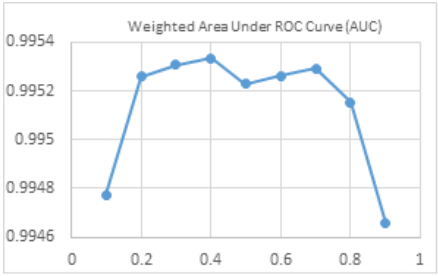
We conduct an experiment to measure the effects of two resampling techniques on the perfor-

Table 4: Confusion Matrix for the Binary Classification Problem

Attribute	Description	Type
A1	I often notice small sounds when others do not	Binary
A2	I usually concentrate more on the whole picture, rather than the small details	Binary
A3	I find it easy to do more than one thing at once	Binary
A4	If there is an interruption, I can switch back to what I was doing very quickly	Binary
A5	I find it easy to read between the lines when someone is talking to me	Binary
A6	I know how to tell if someone listening to me is getting bored	Binary
A7	When I'm reading a story, I find it difficult to work out the characters' intentions	Binary
A8	I like to collect information about categories of things	Binary
A9	I find it easy to work out what someone is thinking by their face	Binary
A10	I find it difficult to work out people's intentions	Binary
Age	Age of user	Numeric
Sex	Male or Female	Character
Ethnicity	List of common ethnicities - text format	String
Jaundice	Whether the case was born with jaundice	Boolean
Family_ASD	Whether anyone in immediate family has ASD	Boolean
Score	Score by AQ-10-Adult. Score greater than 3 indicates autistic traits	Numeric



(a) Recall and precision results for the adult autism datasets



(b) Weighted area under curve (AUC) desults for the Adult Autism dataset

Figure 5: Recall results for different class percentages

mance of the Naive Bayes classifier applied to the Adult Autism dataset. Random Under-sampling (RUS) is a resampling technique whereby the majority class in the dataset is reduced to match the size of the minority class instances (Fernndez et al., 2017). This reduction occurs by random deletion of instances in the majority class. Oversampling techniques such as Synthetic Minority Oversampling technique (SMOTE) involve increasing the number of minority class observations in the training set to balance the class ratio (Chawla et al., 2002). Concretely, SMOTE creates the new instances of the minority class by randomly interpolating between the existing instances of the minority class.

Our findings are presented in Table 5. The results obtained after applying RUS demonstrate a slight increase in the recall of the models derived by the Nave Bayes algorithm. More importantly, the results derived after applying SMOTE show an increase in both the precision and recall of the classification models. Note that the F1-score is also increased after applying SMOTE.

Table 5: Confusion Matrix for the Binary Classification Problem

	Precision	Recall	F1	AUC-ROC
No Resampling	0.94	0.93	0.94	0.99
Random Under-sampling	0.94	0.94	0.94	0.99
SMOTE	0.96	0.96	0.96	0.99

Recall is an important measure when dealing with medical datasets and classification problems. A higher recall indicates smaller chance of false negatives which could see an individual, who is on the autism spectrum, be misdiagnosed and thus not followed up with any treatment or further assessment by a medical professional.

AUC represents the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative sample. In our case it translates to how well the model can identify individuals that are on the ASD versus those who are not. The models results with respect to the AUC metric are high particularly when resampling techniques are used to balance the Adult Autism dataset. The resulting modelis expected to rank a randomly chosen positive sample above a negative one 99% of the time.

5. Conclusion

Class imbalance is one of the key issues in classification tasks. Machine learning algorithms focus on maximizing the total accuracy over the entire dataset leading to more attention being paid to the majority class samples. As a result the minority class samples are poorly predicted by the learning model. We study the class imbalance problem in classification by experimentally applying a probabilistic classification algorithm over a range of different class ratios. In particular, we carried out a large number of experiments on various datasets using Naive Bayes classifier. We employed a range of evaluation metrics implemented in Java and embedded within the WEKA platform. The bases for the comparison are average error rate, average recall, and average precision which are calculated over 100 runs for each class ratio and dataset to ensure consistency and statistical significance. The reported results clearly reveal that for the majority of datasets the

evaluation metrics are at their minimum values when the datasets are balanced (50%:50%). The highest evaluation metric values are derived when the dataset is imbalanced, i.e. 10%:90% or 90%:10%, which can be due to higher certainty in the available variables with respect to the response variable. Nevertheless, these results could be biased as the instances within the dataset are linked with a majority class, and hence treating instances of both class labels in terms of classification cost in the evaluation phase could be unfair and cause a performance boost for the majority class.

The main novelty and contribution of the paper is the study of the precise nature of the relationship between the degree of class imbalance and the corresponding classifier performance. This is a little explored topic. Although many are aware that class imbalance causes problems there are not any in depth studies about its precise effects. We conduct extensive experiments to highlight the effects of changing class imbalance on classifier performance. We hope that by learning more about the exact effects of class imbalance on classifier performance researchers can better tackle the problem. We are also first to measure the effects of class imbalance ratio in behavioural applications such as autism spectrum disorder screening. In addition to measuring effects of changing class ratio on medical applications we have experimentally contrasted different data resampling methods to seek which one can alleviate classification performance for different classification applications including ASD screening. We hope our work will open the door into enhancing the design of ASD screening systems to reduce false negative predictions.

In the near future, we will measure the impact of data sparsity on the classification performance for datasets with class imbalance. This is necessary because when data is sparse there will be a large number of attribute values that are zero. Reducing the dimensionality of these datasets before constructing classifiers by machine learning algorithms becomes essential especially for datasets which suffer from data imbalance. The study will be of considerable interest for data scientists, researchers in data science and information theory among others.

References

[1] F. Bryant and A. Satorra, "Principles and practice of scaled difference chi-square testing," *Structural Equation Modeling: A Multidisciplinary Journal* **19**(3)(2012) 372-398.

[2] Abdelhamid, N., Ayesh, A., & Thabtah, F. (2013). Phishing detection using associative classification data mining. ICAI'13 - The 2013 International Conference on Artificial Intelligence, pp. (491-499). USA.

[3] Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2015). Classification with class imbalance problem: A review. International Journal of Advances in Soft Computing and its Applications, 7(3), 176-204. Retrieved from https://pdfs.semanticscholar.org/1e48/70524f8de44d4f18c8f9f80eb797dfd25c89.pdf

[4] Allison, C., Auyeung, B., and Baron-Cohen, S. (2012). Toward brief red flags for autism screening: the short autism spectrum quotient and the short quantitative checklist for autism in toddlers in 1,000 cases and 3,000 controls [corrected]. Journal of American Academy of Child and Adolescent Psychiatry, 202-212.

[5] AlShboul, R., Thabtah, F., Abdelhamid, N. & Al-diabat, M. (2018) A visualization cybersecurity method based on features dissimilarity. Computers & Security 77, 289-303.

[6] Anwar, M. N. (2012). Complexity measurement for dealing with class imbalance problems in classification modelling. Thesis for Doctor of Philosophy, Massey University, Institute of Fundamental Sciences.

[7] Belarouci, S., & Chikh, M. A. (2017). Medical imbalanced data classification. Advances in Science, Technology and Engineering Systems Journal, 2(3), 116-124.

[8] Boyle, T. (2018). Dealing with imbalanced data: A guide to effectively handling imbalanced datasets in Python.

[9] Buda, M. (2017). A systematic study of the class imbalance problem in convolutional neural networks. Kth Royal Institute of Technology, School of Computer Science and Communication, Sweden.
[10] Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, 106, 249-259.
[11] Chawla, N. V. (2005). Data mining for imbalanced datasets: An overview. In O. Maimon, & L. Rokach, Data Mining and Knowledge Discovery Handbook (pp. 853-867). Boston: Springer.
[12] Chawla, N. V., Bowyer, K., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16(1), 321-357.
[13] Duda, R. O. & Hart P. E. (1973). Pattern classification and scene analysis. New York: John Wiley & Sons.
[14] Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. Computational Intelligence, 20(1), 18-36.
[15] Fan, W., Stolfo, S. J., Zhang, J., & Chan, P. K. (1999). AdaCost: Misclassification cost-sensitive boosting. Proceedings of the Sixteenth International Conference on Machine Learning (pp. 97-105). San Francisco: Morgan Kaufmann.
[16] Fernndez, A., Ro, S. d., Chawla, N. V., & Herrera, F. (2017). An insight into imbalanced Big Data classification: outcomes and challenges. Complex & Intelligent Systems.
[17] Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. The Annals of Statistics, 28(2), 337-407.
[18] Freund, Y. & Schapire, R. E. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences. 55: 119139.
[19] Guo, X., Yin, Y., Dong, C., Yang, G., & Zhou, G. (2008). On the class imbalance problem. 2008 Fourth International Conference on Natural Computation (pp. 192-201). Jinan, China: IEEE.
[20] Hall, M., Frank, E., Holmes G., Pfahringer B., Reutemann P. & Witten I. (2009) The WEKA data mining software: An update; SIGKDD explorations, 11(1).
[21] Hart, P. E. (1968). The condensed nearest neighbor rule. IEEE transactions on information theory, 14(5) 15-5 16.
[22] He, H. & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21, pp. 1263 -1284. IEEE.
[23] Huang, C., Li, Y., Loy, C. C., & Tang, X. (2018). Deep imbalanced learning for face recognition and attribute prediction. Arxiv.
[24] Hulse, J., Khoshgoftaar, T., & Napolitano, A. (2007). Experimental perspectives on earning from imbalanced data. Proceedings of the 24th International Conference on Machine Learning (pp. 935-942). Corvalis, Oregon, USA: Oregon State University.
[25] Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent Data Analysis, 6(5), 429-449.
[26] Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data. 2019;6(1):27.
[27] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann. 2 (12): 11371143.
[28] Kubat, M. & Matwin, S. (1997). Addressing the curse of imbalanced training sets: One-sided selection. In proceedings of the Fourteenth Intemational Conference on Machine Learning, pages 179-186, Nashville, Tennesse. Morgan Kaufmann.
[29] Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering, 30.
[30] Laurikkala, J. (2001). Improving identification of difficult small classes by balancing class distribution. Conference on Artificial Intelligence in Medicine in Europe (pp. 63-66). Portugal: Springer.
[31] Liu, Y., Loh, H. T., Youcef-Toumi, K., & Tor, S. B. (2007). Handling of imbalanced data in text classification: Category-based term weights. In A. Kao, & S. R. Poteet, Natural Language Processing and Text Mining (pp. 171-192). London: Springer.
[32] Longadge, R., Dongre, S. S., & Malik, L. (2013). Class imbalance problem in data mining: Review. International Journal of Computer Science and Network (IJCSN), 2(1).

[33] Nazrul, S. S. (2018). Fraud detection under extreme Ccass imbalance. Retrieved from https://towardsdatascience.com/fraud-detection-under-extreme-class-imbalance-c241854e60ch
[34] Onan A. (2019) Consensus Clustering-Based Under-sampling Approach to Imbalanced Learning. Scientific Programming. Volume 2019, Article ID 5901087.
[35] Ouyang, X. Q., Chen, Y. P., & Wei, B. H. (2017). Experimental study on class imbalance problem using an oil spill training data set. British Journal of Mathematics & Computer Science, 21(5), 1-9.
[36] Prati, R. C., Batista, G. E., & Silva, D. F. (2015). Class imbalance revisited: a new experimental setup to assess the performance of treatment methods. Knowledge and Information Systems, 45(1), 247-270.
[37] Radwan, A. M. (2017). Enhancing Prediction on Imbalance Data by Thresholding Technique with Noise Filtering. 8th International Conference on Information Technology (ICIT) (pp. 399 - 404). IEEE. Enhancing prediction on imbalance data by thresholding technique with noise
[38] Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence, 23(4), 687719.
[39] Thabtah, F. (2018a). An accessible and efficient autism screening method for behavioural data and predictive analyses. Health Informatics Journal.
[40] Thabtah F. (2018b). Machine learning in autistic spectrum disorder behavioral research: A review and ways forward Informatics for Health and Social Care 43,(2), 1-20.
[41] Thabtah, F., Kamalov, F., & Rajab, K. (2018). A new computational intelligence approach to detect autistic features for autism screening. International Journal of Medical Infromatics, 117, pp. 112-124.
[42] Thabtah, F., & Peebles, D. (2019). A new machine learning model based on induction of rules for autism detection. Health Informatics Journal, doi:146045821882471l. Thabtah, F. (2017a). ASDTest.
[43] Thabtah, F. (2017b). Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. INFORMATICS FOR HEALTH & SOCIAL CARE.
[44] Tomek, I. (1976). Two modifications of CNN. IEEE Transactions on Systems, Man, and Cybernetics, 6(11), 769 - 772.
[45] Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics, 2(3), 408 - 421.
[46] Yan, R., Liu, Y., Jin, R., & Hauptmann, A. (2003). On predicting rare class with SVM ensemble in scene classification. 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 3. Hong Kong, China: IEEE.
[47] Yang, Q., & Wu, X. (2006). 10 Challenge problems in data mining research. International Journal of Information Technology and Decision Making, 5(4), 597-604.
[48] Zou, Q., Xie, S., Lin, Z., Wu, M., & Ju, Y. (2016). Finding the best classification threshold in imbalanced classification. Big Data Research.
[49] Zhu, B., Baesens, B., & vanden Broucke, S. K. (2017). An empirical comparison of techniques for the class imbalance problem in churn prediction. Information sciences, 408, 84-99.

We declare that

a. There is no conflict of interest

b. This manuscript is the authors' original work and has not been published nor has it been submitted simultaneously elsewhere.

b. All authors have checked the manuscript and have agreed to the submission.