Computing and Informatics, Vol. 41, 2022, 981–1001, doi: 10.31577/cai_2022_4_981

NEW HYBRID DATA PREPROCESSING TECHNIQUE FOR HIGHLY IMBALANCED DATASET

Esraa Faisal Malik, Khai Wah Khaw

School of Management
Universiti Sains Malaysia
11800 Gelugor, Penang, Malaysia
e-mail: e.faisal.m31@gmail.com, khaiwah@usm.my

XinYing Chew

School of Computer Science
Universiti Sains Malaysia
11800 Gelugor, Penang, Malaysia
e-mail: xinying@usm.my

Abstract. One of the most challenging problems in the real-world dataset is the rising numbers of imbalanced data. The fact that the ratio of the majorities is higher than the minorities will lead to misleading results as conventional machine learning algorithms were designed on the assumption of equal class distribution. The purpose of this study is to build a hybrid data preprocessing approach to deal with the class imbalance issue by applying resampling approaches and CSL for fraud detection using a real-world dataset. The proposed hybrid approach consists of two steps in which the first step is to compare several resampling approaches to find the optimum technique with the highest performance in the validation set. While the second method used CSL with optimal weight ratio on the resampled data from the first step. The hybrid technique was found to have a positive impact of 0.987, 0.974, 0.847, 0.853 F2-measure for RF, DT, XGBOOST and LGBM, respectively. Additionally, relative to the conventional methods, it obtained the highest performance for prediction.

Keywords: Cost-sensitive learning, hybrid, imbalance dataset, resampling techniques

1 INTRODUCTION

Many real-world datasets such as data in bioinformatics [1], natural science [2, 3, 4], manufacturing industries [5], medicine and health [6, 7, 8] and finance [9, 10, 11] face the problem of imbalance dataset where the class distribution is heavily skewed towards the majority class as it has much more samples than the class of interest (minority). Due to the assumption of equal class distribution made by these algorithms, such a problem cannot be solved by conventional machine learning algorithms [12, 13]. This assumption will lead the algorithm to achieve a very high accuracy rate when evaluating the model because it will be biased towards the majority class [14]. Therefore, the accuracy rate is not a suitable measurement for the classification performance in the case of an imbalanced dataset [15]. Another good alternative that was found in the literature is the use of Area Under the Curve (AUC) [16] and the Geometric Mean (G-mean) [17]. Regardless of the severity of the imbalance, both are regarded as effective methods for evaluating an algorithm's performance. Most researchers tend to adopt a criterion named imbalance ratio (IR) to represent the severity of imbalance problem in an existing dataset in which it is known as the ratio between the instances minority and the majority class.

Another problem of the imbalanced dataset is that the misclassification of the minority class will be much higher than the misclassification of the majority class [18, 19]. Especially in the case of credit card fraud domain, where the misclassification of incorrectly classifying a fraud class as non-fraud is much higher and costly than if wrongly classify non-fraud as fraud. Despite the fact that numerous studies on resampling approaches in credit card fraud detection have been conducted at both the data and algorithm levels to solve the class imbalance dilemma. This study is one of the pioneers to integrate CSL with a variety of data-level resampling techniques to address the problem of a skewed imbalance dataset in fraud detection, to the best of our knowledge. Our proposed technique initially uses resampling approaches to alter the class distribution of our dataset based on the optimal performance for each classifier. Next, CSL will be applied to the resampled data from the previous step using the optimal CSL ratio. The initial experiment was performed to find the optimal classifier and create several feature subsets using feature selection technique such as correlation, variance, Infogain, wrapper, autoencoder and Principal Component Analysis (PCA) for the fraud datasets. Subsequently, to discover the optimal resampling approach and to obtain the optimal CSL weight ratio, two separate additional experiments were conducted. In the final experiment, the hybrid technique is compared with the existing resampling approach.

The rest of this study is structured as follows. In Section 2 we discuss the literature review. Section 3 summarizes the CRISP-DM cycle (cross industry-standard process for data mining) used in this study. Moreover, as the main contribution of this research, it also describes the proposed hybrid approach for fraud imbalanced data issue. Section 4 presents the results and discussion. Additionally, the conclusion as well as the current challenges and future prospects will be discussed in Section 5.

2 LITERATURE REVIEW

There are three common approaches for solving the imbalance issue [14]. First, the algorithms level approach in which it modifies the existing algorithm to become suitable for class imbalance such as K-Nearest Neighbours (KNN) [20] and Support Vector Machine (SVM) [21]. The second is the data level approach where the class distribution of an imbalanced dataset can be resampled. This approach is considered to be the most common and widely used approach for imbalanced classification since it helps avoiding the adjustments of machine learning algorithm [12]. There are three main subcategories in this approach: oversampling, undersampling and hybrid approach [12, 20, 22, 23]. To balance the dataset, in oversampling approach the examples from the minorities will get duplicated without adding any new information to the model. On the contrary, in undersampling approach, examples from the majorities will be eliminated from the training dataset. Furthermore, the hybrid method is a combination of the two approaches.

The author in [24] offered an overview of the well-known solutions to the class-imbalance problem at both data and algorithmic levels. Sampling is the most often used strategy for dealing with uneven data at the data level. For local classifiers, oversampling outperforms undersampling, however, some undersampling algorithms beat oversampling when using algorithms with global learning. On the other hand, the authors demonstrated that hybrid sampling strategies outperform oversampling and undersampling. Moreover, they also indicated that resampling is preferable to a CSL. The paper also highlighted that decision trees (DT), when combined with sampling strategies, have shown to be an effective solution to resolve the problem of unbalanced data.

To evaluate which strategy yields the best overall classifier, the authors in [25] developed three techniques for dealing with skewed class distribution and nonuniform misclassification costs were evaluated. The first technique integrates misclassification costs into the learning algorithm, whereas the other two use oversampling or undersampling to balance the training data. According to their findings, there is no clear victor amongst CSL, oversampling, and undersampling based on the results from all of the data sets. Given this, the next issue is whether we can identify the conditions in which each technique works best. The authors in [26] proposed a new undersampling approach for selecting instances from the majority class, their focus was on the instances that are more likely close to the decision boundary. Their proposed method is then combined with a weighted-SVM, with different weights used for misclassification the two classes. Their results reveal that the proposed method has outperform the sensitivity compared to individual weighted-SVM as well as the result from other studies for the same used dataset. In addition, the authors in [27] investigated the ability of the classification models to distinguish fraud and non-fraud transactions, as well as whether different resampling approaches may improve the models' accuracy. For resampling, they used random undersampling (RUS), random oversampling (ROS) and synthetic minority oversampling technique (SMOTE). Their result indicated that ROS gives more convincing results compared

to SMOTE in imbalance credit card datasets and Random Forest (RF) showed a robust performance in three resampling approaches. Similarly, the authors in [28] have shown similar results for the previous research. On the other hand, they have indicated that the algorithms results perform well when applied to the entire dataset rather than to the under-sampled dataset, owing to the under-sampling approach's weakness when applied to a large dataset, where removing the number of majority class members even equal to minority class members has a significant effect on the results [28]. Cross-validation, on the other hand, boosted the effectiveness of certain procedures while decreased the effectiveness of others [29].

Researchers have proved that hybrid sampling techniques which combine both undersampling and oversampling approaches can perform better than just oversampling or undersampling. The authors in [30] investigated seven class balancing techniques for the credit card fraud detection dataset. Their result indicated that Synthetic Minority Over-Sampling Technique with Edited Nearest Neighbors (SMOTE-ENN) is the best resampling approach in the credit card fraud detection domain. Furthermore, to address the imbalanced classification problem a proposed method called random hybrid sampling boosting (RHSBoost) was built by the authors in [31]. RHSBoost employs a hybrid sampling scheme that incorporates undersampling on the majority class and Random Over-Sampling Examples (ROSE) sampling on all data sequentially to address the imbalanced classification problem. Based on their results, RHSBoost tends to be an appealing classification model for the imbalanced dataset. More recently, a novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors used the oversampling technique named synthetic minority oversampling approach with natural neighbors (NaNSMOTE) to solve the most challenging problem, the selection of the parameter k and the determination of the neighbor number of each sample. The proposed model achieved a promising result. However, among the proposed six models, NaNSMOTE-IPF has achieved the greatest results [14].

The third common approach to solve the imbalance problem is the CSL approach in which it takes a combination between data level and algorithm level, as it incorporates the costs of prediction errors (as well as potentially other costs) into training data and adapts the learning process to accept costs. The CSL operates by presuming higher misclassification costs for the minority class, thus, it will be more biased toward the minority class. Furthermore, it also seeks to reduce the total cost of errors of both minority and majority class [12].

The authors in [32] developed three cost-sensitive boosting algorithms into the learning architecture of Adaboost. The cost items are used to signify the unequal identification importance between classes so that boosting techniques can purposefully bias learning towards the class associated with higher identification importance and finally increase the performance. Likewise, for the unbalanced classification, an efficient ensemble of cost-sensitive decision trees was presented by [33]. When cost-sensitive decision trees are combined with random subspace-based feature space partitioning, a pool of individual classifiers capable of enhanced recognition of the minority class is created. An evolutionary method is used to choose complementary

classifiers from the classifier pool, while the assignment of classifier weights, which is employed in the fusion stage, is handled as an optimization issue, and is contained in the evolutionary method. As a result, simultaneous selection and weighted fusion are used to utilize the individual capabilities of the classifiers available. The derivation of cost-matrices is a key challenge in cost-sensitive categorization. Based on Receiver operating characteristic (ROC) analysis, they addressed this in their methodology and shown that there is a clear association between the dataset imbalance ratio and the optimal cost.

The authors in [34] used meta cost procedure and Artificial Neural Network (ANN) to enhance credit card fraud detection and minimize the risk of loss either financially or reputationally. The authors noted that the cost of false negative (FN) should be higher than false positive (FB) as in such cases as fraud detection, where an organization tends to care more about FN to avoid cost resulting from missing a fraudulent activity which is usually greater than falsely alleging fraud. More recently, [35] addressed the imbalanced classification problem by introducing a Cost-sensitive Feature selection General Vector Machine (CFGVM) technique based on the General Vector Machine (GVM) and Binary Ant Lion Optimizer (BALO) methods, which assigns different cost weights to distinct classes of data. To increase classification performance, the BALO algorithm estimates cost weights and extracts more relevant features. Experiments on eleven unbalanced datasets revealed that CFGVM technique enhances the classification performance of minority class samples considerably. When compared to similar algorithms and the state-of-the-art algorithms, the newly suggested approach is greatly outperforming even the superior results – in terms of the performance and yields.

On the contrary, only a limited number of studies tried a combination of the two approaches. The authors in [34] proposed two scientific approaches for dealing with a class imbalance that make use of both the CSL and resampling technique. The first approach employs SVM to integrate and compare various sampling methods with CSL. The second approach suggests using CSL by locally maximizing the expense ratio (cost matrix). Their findings indicate that the first approach can lower misclassification costs while the second method can increase the classifier accuracy. Similarly, the authors in [36] used a Korean Bankruptcy dataset to develop a hybrid method for bankruptcy prediction utilizing an oversampling technique and CSL (HAOC). In the first phase, they used an oversampling method with the best balancing ratio. In the second phase, they employed a CSL model called the cluster-based boosting (CBoost) algorithm, to forecast the bankruptcy. The findings revealed that HAOC gave the optimal results compared to other resampling methods used for the same dataset value for bankruptcy prediction compared with the existing approaches. However, these results were based upon only one dataset and individual classifier and the effect of the HAOC using different types of datasets (or feature subsets), resampling techniques and CSL is unclear. Despite this interest and to the best of our knowledge, no studies were made on the impact of using both the resampling technique and CSL to address the problem of a skewed dataset in credit card fraud detection.

3 MATERIALS AND METHOD

This study was organized based on the CRISP-DM cycle (cross industry-standard process for data mining) [37]. As illustrated in Figure 1, the scheme of this flowchart includes data collection, data cleaning and preprocessing, data mining, and model evaluation.

3.1 Data Collection

The dataset was provided by Vesta Corporation. It was released by researchers from the IEEE Computational Intelligence Society (IEEE-CIS) Kaggle community as a benchmark dataset [38]. The dataset is broken into four excel sheet files; namely – test identity, test transaction, train identity and train transaction, which are joined by TransactionID. The test dataset includes around 506 K records, the training dataset contains 590 K instances and 434 features including the target class in which 31 are categorical and 403 are numerical attributes. Some attributes names were masked due to privacy concern and contract agreement. The training dataset contains around 569 K legitimate transactions and 20 K fraudulent transactions, which has a ratio of fraudulent cases equal to 0.035 and the ratio of non-fraudulent cases equal to 0.965. This ratio is extremely imbalanced and highly skewed towards the legitimate cases (presented by 1) which will lead to ambiguous results. Thus, some techniques need to be developed to overcome the issue of ambiguous performance results due to extremely imbalance datasets. However, to enhance the training time of the model we herein apply POC (proof of concept) with 59 054 transactions with the same fraud ratio of the original dataset for the training dataset and 25 335 for the test dataset.

3.2 Data Preprocessing

Usually, real-world data is inconsistent, incomplete, and most likely contains countless errors. Consequently, data preprocessing is a must in which it transforms the raw data into a more appropriate format for the learning purpose. It has four major phases: data cleaning, data normalization, categorical encoding, and dimensionality reduction. In the first phase, columns containing a minimum of 75 percent missing values were removed owing to the fact that the amount of information stored in that specific features is inadequate to build a prediction model [39]. While the rest of the missing values were replaced by mode and median for categorical and numerical values, respectively. Moreover, as machine learning algorithms function the best when the features in the dataset are on the same scale, the rescaling approach 'Min-Maxscaler' was used to rescale the data between $(0, 1)$ in the second phase [40]. The ordinal attributes were mapped as integers according to their sequences in the third phase; particularly, attributes with false and true were mapped into 0 and 1, respectively. Otherwise, for nominal attributes, we used a famous method called one-hot

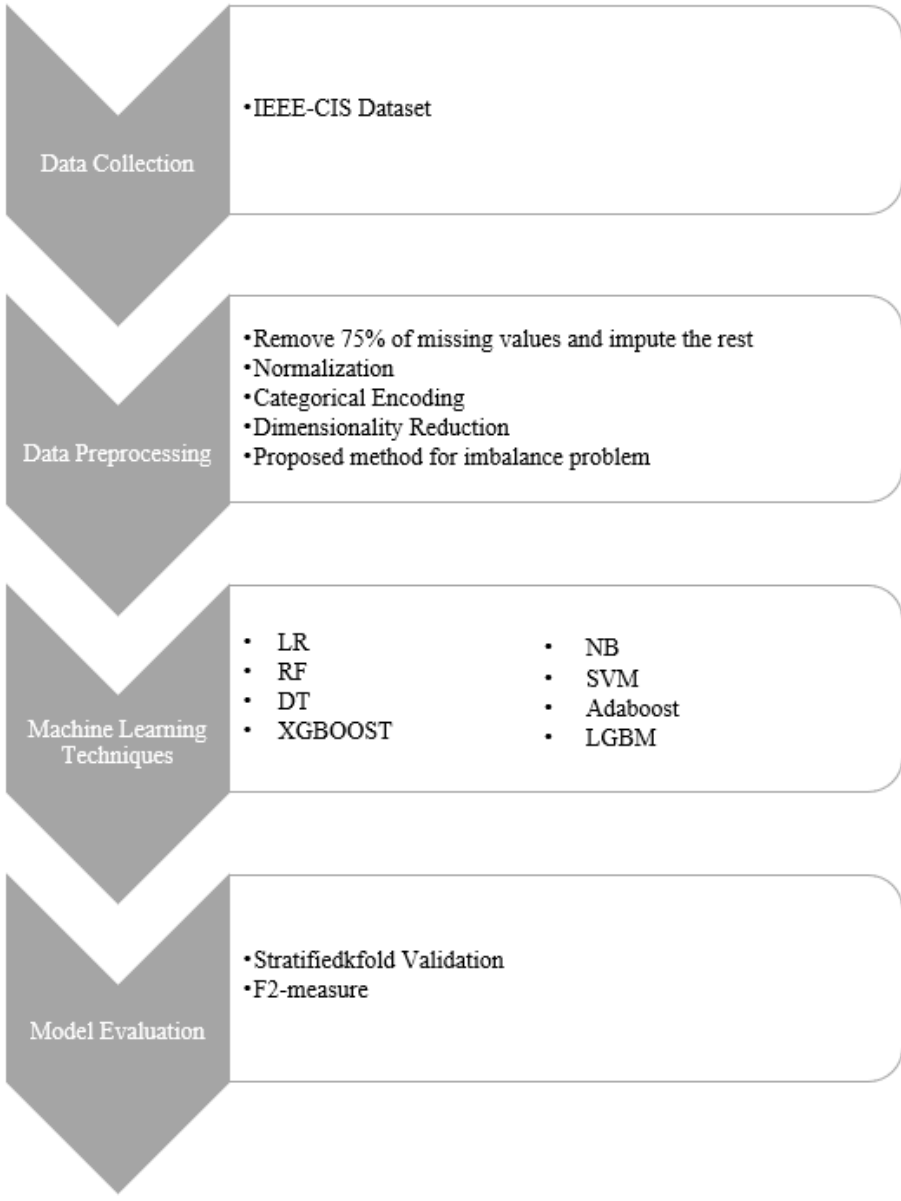


Figure 1. CRISP-DM cycle

encoding. Finally, dimensionality reduction techniques were utilized to assist in reducing the curse of dimensionality, training time and enhance the performance of the algorithm by eliminating irrelevant and unnecessary variables. In this study, we have implemented six feature selection; correlation, variance, Infogain, wrapper, autoencoder and PCA to choose the best subset from the 274 features.

3.2.1 Data Resampling

The data level approach or resampling approaches are the most common approaches in solving the imbalance issue as it helps ease the negative effect caused by class imbalance by balancing the class distribution on a data level. As previously mentioned, the properties of oversampling, undersampling and hybrid approach can be used to benefit in balancing the class distribution for the predictive performance enhancement. However, the absolute benefit of one form of resampling over another is not present. In addition, the application of these methods depends on the dataset itself. In the meantime, the drawback of the undersampling approach is that it is possible to delete potentially valuable samples of data that may be essential for the prediction phase. When the number of minorities is way less than the majorities, undersampling approaches become ineffective. Furthermore, the most significant disadvantage of the oversampling approach is that duplicating exact copies of existing instances increases the model's likelihood of overfitting and increases processing time. Therefore, to overcome these limitations of the above-mentioned methods and maximize the advantages, the hybrid approach will be the best fit for our problem. Several hybrid resampling approaches such as SMOTE-ENN [36], SMOTE-Tomek [41], SMOTE-NCR [42], SMOTE-OSS [43], Random-SMOTE [44], ROS and RUS [13] were applied to enhance the prediction performance for imbalanced dataset. At first, SMOTE algorithm was introduced in 2002 by Chawla et al. [45]. It is the most popular oversampling approach as it can prevent the problem of overfitting by not just duplicating the existing instances, yet to synthesize new instances from the minority class. In the beginning, SMOTE chooses a random minority example "a", then it locates its k nearest minority neighbours. The synthetic attributes are then generated by selecting one of the k nearest neighbours b at random and linking a and b to create a line in the feature space, thus the new sample will be located at a point along this line. This approach can be used extensively for the creation of new synthetic examples for the minority class. The resampling hybrid approaches are briefly described as follows:

SMOTE-ENN: At first, this approach uses SMOTE to perform oversampling, then Edited Nearest Neighbours (ENN) will eliminate the undesirable overlapping instances from both classes. Subsequently, any sample that is misclassified or different from two samples in the three nearest neighbours is removed from the training set [46].

SMOTE-Tomek: This approach was initially applied to enhance the classification of instances to solve the problem of issue of protein annotation in Bioinformat-

ics [47]. It works by first oversampling the original data by SMOTE, then, as an alternative of eliminating examples only from the majority, Tomek links eliminate instances from both classes.

SMOTE-OSS: This method was firstly used by [43, 48] to balance the dataset in which SMOTE will boost the instance number of minority classes, whereas, for the majority class, One-Sided Selection (OSS) will remove the borderline and noise samples to decrease the risk of misclassification as well as loss of essential information the data.

Random-SMOTE: The researchers in [45] mentioned that the combination of SMOTE and undersampling performs better than simple undersampling. It performs by randomly removing some majority examples from the training data until it becomes somewhat equal to the minorities. Therefore, a higher undersampling ratio will lead to a greater presence of minorities in the training dataset. Again, it does not matter the order in which these procedures are implemented since they are applied to various subsets of the training dataset.

ROS and RUS: ROS consists of randomly duplicating the instances in the minority class, while RUS is randomly removing instances in the majority class. Thus, this method is considered the simplest hybrid methods, although they are often ineffective with several limitations when used separately, they can be effective when utilized together. Since these two transformations are carried in different groups, it does not matter the order in which they are employed in the training dataset [13].

3.2.2 Cost-Sensitive Learning

On the other hand, cost-sensitive learning approach or CSL is used to enhance the algorithms' performance in an imbalance dataset. It aims to learn more about the minorities by lowering cost errors, this is achieved by considering the higher cost for misclassification for the positive class in respect to the negative class [49]. The reason behind the assumption is based on countless real-world datasets and the cost of errors is often unequal such as cancer diagnosis, spam email detection, and identifying a fraud. In such cases, the frequency of FN is higher and costlier than FB instances. For example, in fraud detection, the cost of incorrectly identifying a fraud transaction as legitimate is significantly higher than falsely identifying a legitimate transaction as fraud. As this is a binary classification problem, the positive and negative examples were denoted as (1) and (0), respectively. Table 1 illustrates the cost matrix.

Labels	Actual Negative	Actual Positive
Predicted Negative	$C(0,0)$	$C(0,1)$
Predicted Positive	$C(1,0)$	$C(1,1)$

Table 1. Cost matrix

There is no cost when correctly classifying the data. However, the total cost of misclassification of data is defined as a cost-weighted sum of the FN and FB, as shown in Equation (1), where the purpose of CSL is to minimize this cost [50].

$$\text{TotalCost} = C(0,1) * \text{False Negative} + C(1,0) * \text{False Positive}. \tag{1}$$

The efficiency of algorithms is highly dependent on the total cost; as a result, it must be defined with caution. However, in many domains, such as the one in this study, this could be a difficult task because it must be specified by domain experts. However, as we do not have prior information on the cost matrix or expertise in the domain, there is an alternative way by assigning the costs based on the inverse class distribution [13]. For illustration, our dataset contains a ratio of (1 : 28) of minorities examples in respect to the majorities examples. The ratio can be reversed and employed as the cost of misclassification errors. Despite the fact that this method assumes that the class distribution found in the training dataset is indicative of the larger problem and is appropriate for the chosen cost-sensitive ratio, it is an effective heuristic approach for cost setting in general. As a result, using this approach as a reference point is a great idea; then, test with a variety of related ratios to validate its applicability. There are specific machine learning algorithms that are exclusively designed to fit the cost matrix by using various cost metrics that can identify the cost of misclassification of any specific data example; from this group, the following classifiers were used; RF, DT, XGBOOST and LGBM.

3.2.3 The Proposed Approach

Several approaches were found in literature to solve the problem of imbalance classification. However, the problem is still existing and was not yet solved successfully especially for very small ratios such as the one existing in this dataset [36]. As a result, this study has proposed the development of a hybrid data preprocessing technique for credit card fraud detection that combines resampling approaches and CSL to improve the prediction overall performance. Figure 2 depicts the proposed flowchart for the hybrid method. After finishing the essential cleaning and preprocessing steps before feature selection and resampling, a stratified cross-validation method with five folds was used for the training dataset to validate the results in the first experiment and to determine the optimum classifiers for the dataset in this study. Employing the classifier from the first experiment, we experimented various dimensionality reduction methods to find the best possible features for IEEE-CIS dataset that will be introduced in the second experiment. Then, the training set will be re-balanced by ROS-RUS, SMOTE-Random, SMOTE-Tomek, SMOTE-ENN and SMOTE-OSS using four feature sets. Following that, CSL with the best balancing ratio will be applied to the best possible resampled set from the previous experiment for fraud detection.

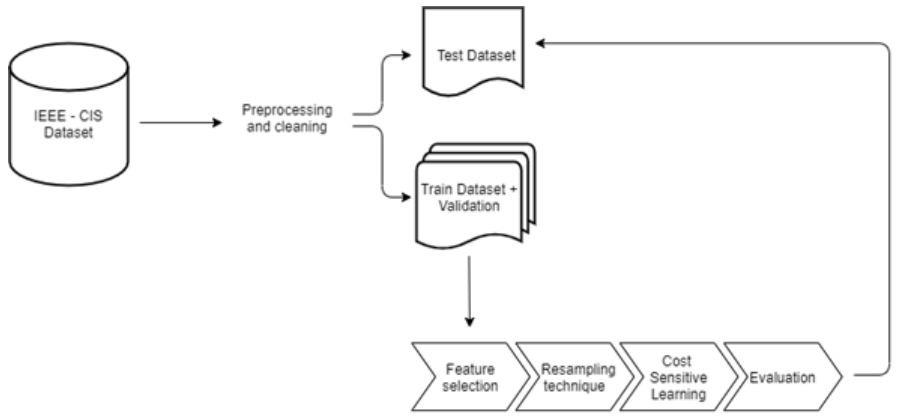


Figure 2. Flowchart of the proposed model

3.3 Machine Learning Techniques

Different classification algorithms have been applied to detect fraudulent transactions as discussed earlier. Yet, there is no optimal algorithm for a specific problem. Therefore, eight different linear and nonlinear algorithms – SVM, LR, RF, DT, XGBOOST, NB, SVM, Adaboost and LGBM – were selected from the literature as they indicated a promising performance in the context of fraud detection. Then, four of these algorithms are considered in this paper by setting a threshold of 0.3 F2-measure.

3.4 Model Evaluation

There are enough instances in the training dataset. Therefore, to evaluate the model, we used a validation method named stratifiedKfold [51], where $k = 5$, it divides the dataset into five portions, four of which are used for training and the remaining one for testing. This validation method ensures that each class in this study is represented roughly equally across all folds.

Additionally, the accuracy will be a misleading evaluation metric in the case of an imbalanced dataset, therefore, in this study, we have employed F2-measure to evaluate and compare the classifiers performance among the experiments. F2-measure is derived from F-measure, it enables the combination of recall and precisions into an individual measure that captures all properties. However, it gives more attention to recall because in some cases such as fraud features, increasing the number of FB is important but it is more important to diminish the FN cases. As shown in Equation 2, F2-measure is an abstraction of F-measure where a coefficient called β governs the balance of precision and recall in the measurement of the harmonic

mean.

$$F\beta = ((1 + \beta^2) \times \text{precision} * \text{recall})/((\beta^2) \times \text{precision} + \text{recall}). \quad (2)$$

To determine the name of $F\beta$ measure, β parameter will be utilized. To illustrate, when $\beta = 1$, the $F\beta$ measure will be named F1-measure. Similarly, when $\beta = 2$, the $F\beta$ measure will be called as F2-measure. It has three popular values which are:

- ($\beta = 0.5$): less weight on recall and extra on precision,
- ($\beta = 1$): equal weight on precision and recall,
- ($\beta = 2$): extra weight on recall and less on precision.

Anaconda Navigator, specifically the Jupyter Notebook environment using Python 3 language, was used to analyze the data in this study, and it was run on a laptop equipped with an Intel Core i7 – 10750H processor (2.60 GHz 6 cores), 16 GB RAM, and Windows 10. Additionally, we used Python available packages called imbalanced-learn package and Scikit-learn package in which they provide the mandatory resampling techniques and machine learning algorithms.

4 RESULTS AND DISCUSSION

After completing the preprocessing tasks, a baseline model with all the remaining features (274 features) for the training set was implemented using eight different algorithms. Setting a threshold of 0.3, the performance results in Table 2 show that RF, DT, XGBOOST and LGBM were the most suitable algorithms for IEEE-CIS dataset using F2-measure.

Algorithm	Baseline Result Using All Features
LR	0.126
RF	**0.330**
DT	**0.367**
XGBOOST	**0.400**
NB	0.157
SVM	0.0257
Adaboost	0.236
LGBM	**0.372**

Table 2. Baseline models results

In the second experiment, we applied dimensionality reduction techniques and used the four selected algorithms from the first experiment, this will assist in creating several feature sets to validate our proposed model. Table 3 illustrates the number of chosen attributes using various methods; filter, wrapper, PCA and autoencoder with its effect on the performance compared to the baseline of the dataset. The results

show that the ultimate performance of feature selection algorithms varies. Generally, the wrapper technique outperformed the three classifiers (RF, XGBOOST, and LGBM) with different selected features. Surprisingly, when applying DT classifier, the filter approach named Infogain led to the highest performance when ($K = 5$) with an increase of F2-measure around 0.012. An explanation for this might be that Infogain was regarded the core to create decision trees from a dataset, and each variable was assessed by Infogain to see which variable best maximised its value. On other hand, this will help to perfectly split the dataset into different groups by minimizing the entropy.

The best number of feature sets using DT is 220 features. For LGBM and XGBOOST the best feature set consists of 60 features, whereas RF contains only 10 feature sets. On the contrary, PCA, autoencoder and variance did not show any promising results, as they were lower than the baseline model for each classifier.

	Baseline	Correlation	Infogain			Autoencoder	PCA	Wrapper (RF)			Variance
		$< 0/8$	3	5	10						$= 0.01$
# Features	All	148	213	215	220	11	61	10	30	60	80
RF	0.304	0.292	0.337	0.349	0.347	0.030	0.126	**0.353**	0.352	0.335	0.113
DT	0.362	0.368	0.368	0.371	**0.374**	0.112	0.224	0.343	0.363	0.337	0.248
XGBOOST	0.364	0.385	0.391	0.388	0.391	0.023	0.208	0.324	0.389	**0.413**	0.143
LGBM	0.351	0.363	0.372	0.366	0.372	0.022	0.176	0.339	0.352	**0.383**	0.107

Table 3. Feature selection techniques

Using the results from the first experiment, now we have four different feature sets that will be utilized in the second experiment in each classifier to validate the effectiveness of our model. In the second experiment, we have applied five different resampling approaches in the data level using the selected features in the first experiment for each classifier. As shown in Table 4, the results indicated that among the utilized approaches, the two with the highest performance are SMOTE-ENN and ROS-RUS. SMOTE-ENN achieved the highest result for both LGBM and XGBOOST which are 0.500 and 0.522, respectively, while applying RF and DT classifiers, SMOTE-ENN gave the greatest performance among the five resampling approaches with a score equal to 0.454, 0.378 of F2-measure, respectively.

Classifier	SMOTE-ENN	SMOTE-Tomek	SMOTE-OSS	Random-SMOTE	RPS-RUS
RF	**0.454**	0.440	0.438	0.442	0.401
DT	**0.378**	0.359	0.362	0.358	0.337
XGBOOST	0.474	0.446	0.445	0.456	**0.523**
LGBM	0.472	0.434	0.427	0.431	**0.500**

Table 4. Resampling techniques and the proposed method

Table 5 demonstrates the results from the third experiment, where GridSearch was used to find the optimum CSL ratio using several values besides the heuristic value (28) that were mentioned previously (see Section 3.2.2). The results indicate that the best performance was achieved by values other than the heuristic. These

values were 1 and 10; in which for DT and RF the optimal CSL ratio was 1. While for LGBM and XGBOOST the value was 10. Conversely, the lowest performance was obtained when the CSL ratio was equal to 1 000. In our proposed method, we have utilized SMOTE-ENN and ROS-RUS with CSL ratios equal to 1 and 10, respectively.

CSL Ratio	1	10	28	50	75	100	1 000
RF	**0.362**	0.338	0.337	0.334	0.324	0.330	0.323
DT	**0.384**	0.370	0.361	0.347	0.355	0.349	0.336
XGBOOST	0.420	**0.543**	0.529	0.498	0.464	0.450	0.340
LGBM	0.393	**0.530**	0.498	0.448	0.409	0.376	0.269

Table 5. CSL ratios

Table 6 illustrates the results and compare the proposed method to previously used techniques. It is clear that our hybrid method has significantly boosted the performance of all the used algorithms. It shows a result of 0.987, 0.974, 0.847, 0.853 for RF, DT, XGBOOST and LGBM, respectively. The highest improvement was achieved by DT as it significantly increased with a 0.596 from the best resampling method in the second experiment. The second improvement was performed by RF which there was a rise of 0.533. Nevertheless, XGBOOST and LGBM have shown the lowest performance as they only increased by 0.324 and 0.353, respectively.

Classifier	Baseline	Best Resampling Result	Best CSL Result	Proposed Method
RF	0.304	0.454	0.362	**0.987**
DT	0.362	0.378	0.384	**0.974**
XGBOOST	0.364	0.523	0.543	**0.847**
LGBM	0.351	0.500	0.530	**0.853**

Table 6. Comparison table

5 CONCLUSION AND RECOMMENDATION

This research was carried out to develop a new hybrid approach to address the imbalanced dataset problem. The research has shown that our proposed method has a significant impact on the model. The results indicated that our proposed approach outperforms the conventional hybrid resampling techniques in the data level and CSL. Additionally, it was found that DT and RF have the best performances and improvement among the other classifiers. However, the most obvious finding which came out from this study is that despite the use of feature selection technique in an imbalanced dataset, the proposed method has boosted the performance. Nevertheless, this study suggests using Wrapper as a feature selection approach for RF, XGBOOST and LGBM. Whereas for DT, the optimum choice is to use Infogain. For

future studies, different fraud detection datasets should be utilized. Furthermore, a combination between resampling techniques on algorithms level and CSL could be performed.

Acknowledgement

This work is funded by Ministry of Higher Education Malaysia, Fundamental Research Grant Scheme (Grant No. FRGS/1/2022/STG06/USM/02/4), for the Project entitled "Efficient Joint Process Monitoring using a New Robust Variable Sample Size and Sampling Interval Run Sum Scheme".

A ABBREVIATION DICTIONARY

Adaboost – Adaptive Boosting
ANN – Artificial Neural Network
AUC – Area Under the Curve
BALO – Binary Ant Lion Optimizer
CBOOST – Cluster-Based Boosting
CFGVM – Cost-sensitive Feature selection General Vector Machine
CRISP-DM – Cross-Industry Standard Process for Data Mining
CSL – Cost-Sensitive Learning
DT – Decision Tree
ENN – Edited Nearest Neighbours
FB – False Positive
FN – False Negative
G-mean – Geometric Mean
GVM – General Vector Machine
HAOC – oversampling technique and CSL
IR – Imbalance Ratio
KNN – K-Nearest Neighbours
LGBM – Light Gradient Boosting
LR – Logistic Regression
NaNSMOTE – Minority Oversampling Technique with Natural Neighbours
NB – Naive Bayes
OSS – One-Sided Selection
PCA – Principal Component Analysis

POC – Proof of Concept

RF – Random Forest

RHSBOOST – Random Hybrid Sampling Boosting

ROC – Receiver Operating Characteristic

ROS – Random Oversampling

ROSE – Random Oversampling Examples

RUS – Random Undersampling

SMOTE – Synthetic Minority Oversampling Technique

SMOTE-ENN – Synthetic Minority Over-Sampling Technique with EditeNearest Neighbour

SVM – Support Vector Machine

XGBOOST – Extreme Gradient Boosting

REFERENCES

[1] Hossain, M. A.—Islam, S. M. S.—Quinn, J. M.—Huq, F.—Moni, M. A.: Machine Learning and Bioinformatics Models to Identify Gene Expression Patterns of Ovarian Cancer Associated with Disease Progression and Mortality. Journal of Biomedical Informatics, Vol. 100, 2019, Art. No. 103313, doi: 10.1016/j.jbi.2019.103313.

[2] Angermueller, C.—Pärnamaa, T.—Parts, L.—Stegle, O.: Deep Learning for Computational Biology. Molecular Systems Biology, Vol. 12, 2016, No. 7, Art. No. 878, doi: 10.15252/msb.20156651.

[3] Mater, A. C.—Coote, M. L.: Deep Learning in Chemistry. Journal of Chemical Information and Modeling, Vol. 59, 2019, No. 6, pp. 2545–2559, doi: 10.1021/acs.jcim.9b00266.

[4] Narla, L. M.—Rao, S. V.: Identification of Metals and Alloys Using Color CCD Images of Laser-Induced Breakdown Emissions Coupled with Machine Learning. Applied Physics B, Vol. 126, 2020, Art. No. 113, doi: 10.1007/s00340-020-07469-6.

[5] Abdelrahman, O.—Keikhosrokiani, P.: Assembly Line Anomaly Detection and Root Cause Analysis Using Machine Learning. IEEE Access, Vol. 8, 2020, pp. 189661–189672, doi: 10.1109/ACCESS.2020.3029826.

[6] Cruz, J. A.—Wishart, D. S.: Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Informatics, Vol. 2, 2006, pp. 59–77, doi: 10.1177/117693510600200030.

[7] Khan, M. A.—Ashraf, I.—Alhaisoni, M.—Damaševičius, R.—Scherer, R.—Rehman, A.—Bukhari, S. A. C.: Multimodal Brain Tumor Classification Using Deep Learning and Robust Feature Selection: A Machine Learning Application for Radiologists. Diagnostics, Vol. 10, 2020, No. 8, Art. No. 565, doi: 10.3390/diagnostics10080565.

[8] Lalmuanawma, S.—Hussain, J.—Chhakchhuak, L.: Applications of Machine Learning and Artificial Intelligence for Covid-19 (SARS-Cov-2) Pandemic: A Review. Chaos, Solitons and Fractals, Vol. 139, 2020, Art. No. 110059, doi: 10.1016/j.chaos.2020.110059.

[9] Huang, J.—Chai, J.—Cho, S.: Deep Learning in Finance and Banking: A Literature Review and Classification. Frontiers of Business Research in China, Vol. 14, 2020, Art. No. 13, doi: 10.1186/s11782-020-00082-6.

[10] Renault, T.: Sentiment Analysis and Machine Learning in Finance: A Comparison of Methods and Models on One Million Messages. Digital Finance, Vol. 2, 2020, No. 1, pp. 1–13, doi: 10.1007/s42521-019-00014-x.

[11] Zhong, X.—Enke, D.: Predicting the Daily Return Direction of the Stock Market Using Hybrid Machine Learning Algorithms. Financial Innovation, Vol. 5, 2019, No. 1, Art. No. 24, doi: 10.1186/s40854-019-0138-0.

[12] Del Rio, S.—Benítez, J. M.—Herrera, F.: Analysis of Data Preprocessing Increasing the Oversampling Ratio for Extremely Imbalanced Big Data Classification. 2015 IEEE Trustcom/BigDataSE/ISPA, Vol. 2, 2015, pp. 180–185, doi: 10.1109/Trustcom.2015.579.

[13] Brownlee, J.: Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning. Machine Learning Mastery, 2020.

[14] Li, J.—Zhu, Q.—Wu, Q.—Fan, Z.: A Novel Oversampling Technique for Class-Imbalanced Learning Based on SMOTE and Natural Neighbors. Information Sciences, Vol. 565, 2021, pp. 438–455, doi: 10.1016/j.ins.2021.03.041.

[15] Sahin, Y.—Bulkan, S.—Duman, E.: A Cost-Sensitive Decision Tree Approach for Fraud Detection. Expert Systems with Applications, Vol. 40, 2013, No. 15, pp. 5916–5923, doi: 10.1016/j.eswa.2013.05.021.

[16] Huang, J.—Ling, C. X.: Using AUC and Accuracy in Evaluating Learning Algorithms. IEEE Transactions on Knowledge and Data Engineering, Vol. 17, 2005, No. 3, pp. 299–310, doi: 10.1109/TKDE.2005.50.

[17] Kubat, M.—Holte, R.—Matwin, S.: Learning When Negative Examples Abound. In: van Someren, M., Widmer, G. (Eds.): Machine Learning: ECML-97. Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 1224, 1997, pp. 146–153, doi: 10.1007/3-540-62858-4_79.

[18] Tsai, C. F.—Lin, W. C.: Feature Selection and Ensemble Learning Techniques in One-Class Classifiers: An Empirical Study of Two-Class Imbalanced Datasets. IEEE Access, Vol. 9, 2021, pp. 13717–13726, doi: 10.1109/ACCESS.2021.3051969.

[19] Vuttipittayamongkol, P.—Elyan, E.—Petrovski, A.: On the Class Overlap Problem in Imbalanced Data Classification. Knowledge-Based Systems, Vol. 212, 2021, Art. No. 106631, doi: 10.1016/j.knosys.2020.106631.

[20] Hu, J.—Li, Y.—Yan, W. X.—Yang, J. Y.—Shen, H. B.—Yu, D. J.: KNN-Based Dynamic Query-Driven Sample Rescaling Strategy for Class Imbalance Learning. Neurocomputing, Vol. 191, 2016, pp. 363–373, doi: 10.1016/j.neucom.2016.01.043.

[21] Ma, H.—Wang, L.—Shen, B.: A New Fuzzy Support Vector Machines for Class Imbalance Learning. 2011 International Conference on Electrical and Control Engi-

neering, 2011, pp. 3781–3784, doi: 10.1109/ICECENG.2011.6056838.

[22] Roy, A.—Cruz, R. M.—Sabourin, R.—Cavalcanti, G. D.: A Study on Combining Dynamic Selection and Data Preprocessing for Imbalance Learning. Neurocomputing, Vol. 286, 2018, pp. 179–192, doi: 10.1016/j.neucom.2018.01.060.

[23] Ghosh, S.—Roy, S.—Bandyopadhyay, S. K.: A Tutorial Review on Text Mining Algorithms. International Journal of Advanced Research in Computer and Communication Engineering, Vol. 1, 2012, No. 4, pp. 223–233.

[24] Ganganwar, V.: An Overview of Classification Algorithms for Imbalanced Datasets. International Journal of Emerging Technology and Advanced Engineering, Vol. 2, 2012, No. 4, pp. 42–47.

[25] Weiss, G. M.—McCarthy, K.—Zabar, B.: Cost-Sensitive Learning Vs. Sampling: Which Is Best for Handling Unbalanced Classes with Unequal Error Costs? In: Stahlbock, R., Crone, S. F., Lessmann, S. (Eds.): Proceedings of the 2007 International Conference on Data Mining (DMIN 2007). CSREA Press, 2007, pp. 35–41.

[26] Anand, A.—Pugalenthi, G.—Fogel, G. B.—Suganthan, P.: An Approach for Classification of Highly Imbalanced Data Using Weighting and Undersampling. Amino Acids, Vol. 39, 2010, No. 5, pp. 1385–1391, doi: 10.1007/s00726-010-0595-2.

[27] Hordri, N. F.—Yuhaniz, S. S.—Azmi, N. F. M.—Shamsuddin, S. M.: Handling Class Imbalance in Credit Card Fraud Using Resampling Methods. International Journal of Advanced Computer Science and Applications, Vol. 9, 2018, No. 11, pp. 390–396, doi: 10.14569/IJACSA.2018.091155.

[28] Mohammed, R.—Rawashdeh, J.—Abdullah, M.: Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. 2020 11th International Conference on Information and Communication Systems (ICICS), 2020, pp. 243–248, doi: 10.1109/ICICS49469.2020.239556.

[29] Ata, O.—Hazim, L.: Comparative Analysis of Different Distributions Dataset by Using Data Mining Techniques on Credit Card Fraud Detection. Tehnički Vjesnik (Technical Gazette), Vol. 27, 2020, No. 2, pp. 618–626, doi: 10.17559/TV-20180427091048.

[30] Sisodia, D. S.—Reddy, N. K.—Bhandari, S.: Performance Evaluation of Class Balancing Techniques for Credit Card Fraud Detection. 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), 2017, pp. 2747–2752, doi: 10.1109/ICPCSI.2017.8392219.

[31] Gong, J.—Kim, H.: RHSBoost: Improving Classification Performance in Imbalance Data. Computational Statistics and Data Analysis, Vol. 111, 2017, pp. 1–13, doi: 10.1016/j.csda.2017.01.005.

[32] Sun, Y.—Kamel, M. S.—Wong, A. K.—Wang, Y.: Cost-Sensitive Boosting for Classification of Imbalanced Data. Pattern Recognition, Vol. 40, 2007, No. 12, pp. 3358–3378, doi: 10.1016/j.patcog.2007.04.009.

[33] Krawczyk, B.—Woźniak, M.—Schaefer, G.: Cost-Sensitive Decision Tree Ensembles for Effective Imbalanced Classification. Applied Soft Computing, Vol. 14, 2014, pp. 554–562, doi: 10.1016/j.asoc.2013.08.014.

[34] Ghobadi, F.—Rohani, M.: Cost Sensitive Modeling of Credit Card Fraud Using Neural Network Strategy. 2016 2nd International Conference of Signal Processing and

Intelligent Systems (ICSPIS), 2016, pp. 1–5, doi: 10.1109/ICSPIS.2016.7869880.

[35] Feng, F.—Li, K. C.—Shen, J.—Zhou, Q.—Yang, X.: Using Cost-Sensitive Learning and Feature Selection Algorithms to Improve the Performance of Imbalanced Classification. IEEE Access, Vol. 8, 2020, pp. 69979–69996, doi: 10.1109/ACCESS.2020.2987364.

[36] Le, T.—Vo, M. T.—Vo, B.—Lee, M. Y.—Baik, S. W.: A Hybrid Approach Using Oversampling Technique and Cost-Sensitive Learning for Bankruptcy Prediction. Complexity, Vol. 2019, 2019, Art. No. 8460934, doi: 10.1155/2019/8460934.

[37] Wirth, R.—Hipp, J.: CRISP-DM: Towards a Standard Process Model for Data Mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining (PADD 2000), Vol. 1, 2000, pp. 29–40.

[38] IEEE Computational Intelligence Society: IEEE-CIS Fraud Detection: Can You Detect Fraud from Customer Transactions? 2019, `https://www.kaggle.com/c/ieee-fraud-detection/`.

[39] Kelleher, J. D.—Namee, B. M.—D'Arcy, A.: Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. The MIT Press, 2015.

[40] Bisong, E.: Introduction to Scikit-Learn. Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners, Apress, Berkeley, CA, 2019, pp. 215–229, doi: 10.1007/978-1-4842-4470-8_18.

[41] Sanguanmak, Y.—Hanskunatai, A.: DBSM: The Combination of DBSCAN and SMOTE for Imbalanced Data Classification. 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2016, pp. 1–5, doi: 10.1109/JCSSE.2016.7748928.

[42] Al Abdouli, N. O.—Aung, Z.—Woon, W. L.—Svetinovic, D.: Tackling Class Imbalance Problem in Binary Classification Using Augmented Neighborhood Cleaning Algorithm. In: Kim, K. J. (Ed.): Information Science and Applications. Springer, Berlin, Heidelberg, Lecture Notes in Electrical Engineering, Vol. 339, 2015, pp. 827–834, doi: 10.1007/978-3-662-46578-3_98.

[43] Pristyanto, Y.—Setiawan, N. A.—Ardiyanto, I.: Hybrid Resampling to Handle Imbalanced Class on Classification of Student Performance in Classroom. 2017 1st International Conference on Informatics and Computational Sciences (ICICoS), IEEE, 2017, pp. 207–212, doi: 10.1109/ICICOS.2017.8276363.

[44] Dong, Y.—Wang, X.: A New Over-Sampling Approach: Random-SMOTE for Learning from Imbalanced Data Sets. In: Xiong, H., Lee, W. B. (Eds.): Knowledge Science, Engineering and Management (KSEM 2011). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 7091, 2011, pp. 343–352, doi: 10.1007/978-3-642-25975-3_30.

[45] Chawla, N. V.—Bowyer, K. W.—Hall, L. O.—Kegelmeyer, W. P.: SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research, Vol. 16, 2002, No. 1, pp. 321–357, doi: 10.1613/jair.953.

[46] Batista, G. E.—Prati, R. C.—Monard, M. C.: A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. ACM SIGKDD Explorations Newsletter, Vol. 6, 2004, No. 1, pp. 20–29, doi: 10.1145/1007730.1007735.

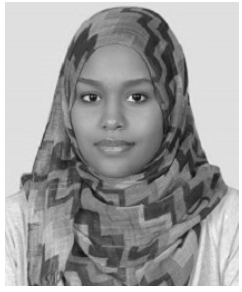
[47] Batista, G. E.—Bazzan, A. L.—Monard, M. C.: Balancing Training Data for Automated Annotation of Keywords: A Case Study. In: Lifschitz, S., Almeida Jr., N. F., Pappas Jr., G. J., Linden, R. (Eds.): II Brazilian Workshop on Bioinformatics (WOB). 2003, pp. 10–18.

[48] Pristyanto, Y.—Pratama, I.—Nugraha, A. F.: Data Level Approach for Imbalanced Class Handling on Educational Data Mining Multiclass Classification. 2018 International Conference on Information and Communications Technology (ICOIACT), IEEE, 2018, pp. 310–314, doi: 10.1109/ICOIACT.2018.8350792.

[49] Thai-Nghe, N.—Gantner, Z.—Schmidt-Thieme, L.: Cost-Sensitive Learning Methods for Imbalanced Data. The 2010 International Joint Conference on Neural Networks (IJCNN), IEEE, 2010, pp. 1–8, doi: 10.1109/IJCNN.2010.5596486.

[50] Elkan, C.: The Foundations of Cost-Sensitive Learning. Proceedings of the 17$^{\text{th}}$ International Joint Conference on Artificial Intelligence (IJCAI'01), Vol. 2, 2001, pp. 973–978.

[51] Bauder, R.—Khoshgoftaar, T.: Medicare Fraud Detection Using Random Forest with Class Imbalanced Big Data. 2018 IEEE International Conference on Information Reuse and Integration (IRI), 2018, pp. 80–87, doi: 10.1109/IRI.2018.00019.



Esraa Faisal Malik received her B.Sc. degree in management information system from the United Arab Emirates University (UAEU), UAE, in 2014, and her M.Sc. degree in data science and analytics from the School of Computer Science, Universiti Sains Malaysia (USM), Malaysia, in 2019. She is currently pursuing her Ph.D. in predictive analytics in business at the School of Management, Universiti Sains Malaysia (USM), Malaysia. Her research interests include machine learning, artificial intelligence, and data mining.

Khai Wah Khaw received his Ph.D. degree from the School of Mathematical Sciences, Universiti Sains Malaysia (USM). He is currently Senior Lecturer with the School of Management, USM. He is Professional Technologist with the Malaysia Board of Technologist (MBOT). His research interests include statistical process control and advanced analytics.

XinYing Chew received her Ph.D. degree in statistical quality control from the Universiti Sains Malaysia. She is currently Senior Lecturer with the School of Computer Sciences, Universiti Sains Malaysia. She is also Certified Trainer with the Human Resources Development Fund (HRDF). She is also Adjunct Research Fellow of the Swinburne University of Technology Sarawak Campus and Professional Technologist with the Malaysia Board of Technologist (MBOT). Her research interests include advanced analytics and statistical quality/process control.