A Machine Learning based Approach to Reduce Behavioral Noise Problem in an Imbalanced Data: Application to a fraud detection

Salma El Hajjami*, Jamal Malki†, Alain Bouju†, Mohammed Berrada*
*IASSE Laboratory, ENSA, USMBA
Fez, Morocco
Email : salma.elhajjami@usmba.ac.ma, mohammed.berrada@gmail.com
†L3i Laboratory
La Rochelle University
La Rochelle, France
Email : jamal.malki@univ-lr.fr, alain.bouju@univ-lr.fr

***Abstract*—The question of class imbalance has become more pronounced with the application of learning algorithms in real applications. It has received significant attention in the machine learning and data mining community. This problem is present in fraud detection, medical diagnostics, and a number of other areas where training data contains significantly more representatives of one class (called the majority class) than the other class (called the minority class). 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, which is the most interesting from a learning point of view and also involves a high cost when it is not well classified. However, the imbalance ratio is not the only cause of poor performance when learning from imbalanced data. Another critical factor that accompanies imbalanced data in the real world is the presence of a number of instances of the two classes being overlapped in feature space. This problem is commonly referred to as class overlap and we have called it "behavioral noise". In this paper, we propose One Side Behavioral Noise Reduction (OSBNR) approach to deal with the problem of class imbalance in the presence of a behavioral noise level. OSBNR is based on two stages. Firstly, a clustering is applied to groups similar instances of the minority class in multiple behavior clusters. Secondly, we select and eliminate instances of the majority class, considered as behavioral noise, which overlap with the behavior clusters of the minority class. The results of experiments conducted on a representative public dataset confirm that the proposed approach is effective for class imbalance problem in the presence of behavioral noise.**

***Index Terms*—Class imbalance, Class overlapping, Machine learning, Data mining, Fraud detection.**

I. Introduction

The advent of Big Data has ushered in a new era of scientific breakthroughs. One of the common issues that plague anomaly detection applications is the class imbalance problem which refers to an imbalanced distribution of instances of different classes. It is known as learning from imbalanced data [1], [2], [3]. Most machine learning systems assume that the dataset used for model construction is balanced. However, when it comes to real world applications, this balance is not always true. This problem is present in fraud detection, network intrusion detection, medical diagnostics and a number of other areas where the class of interest, such as rare diseases or fraudulent transactions, is less represented. The over-represented class is called the majority class and the interest class is called the minority class [4], [1].

Class imbalance has been widely recognized as a complicating factor for most learning algorithms, which assume a relatively uniform distribution of classes. However, in some scenarios, the data imbalance ratio of up to 1 in 100, 1 in 1000, 1 in 10,000 and often even more [5]. As a result, minority instances tend to be misclassified [6], which are usually the most important from a data mining perspective, so such misclassification of minority instances often has serious consequences in the application domain [7]. For example, in the credit card fraud detection system, undetected fraudulent transactions are much more serious and costly than the detection of normal behavior such as fraud.

The correlation between class imbalance and learning algorithms is not yet clear, some studies claim that the imbalance ratio is not the only cause of performance degradation when learning from imbalanced data [3], [8], [9], [10]. Another critical complicating factor for imbalanced data in the real world is the presence of a number of instances of the two classes being, to some degree, mixed in characteristic space. This problem is commonly referred as class overlapping or class separability and we have called it "behavioral noise". The behavioral noise or class overlap problem occurs when a region of the data space contains a similar amount of training data from each class [10], [3]. This situation leads to developing an inference with almost the same a priori probabilities in this area of overlap, which makes the distinction between the two classes very difficult, if not impossible. Most of the existing approaches have been proposed to deal with the problems of class imbalance and overlap separately. However, in many real applications (eg, credit card fraud detection), data sets frequently exhibit both class imbalance and overlap.

In this work, we consider the credit card fraud detection problem. This is a good example of a very imbalanced and overlapped data classification problem [6]. We present a new

PLAIBDE Project. Programme Opérationnel FEDER-FSE de la région Nouvelle Aquitaine - France, 2017-2020

978-1-7281-8376-3/20/$31.00 ©2020 IEEE

approach called: One Side Behavioral Noise Reduction (OSBNR) to deal with class imbalance problem, with a particular emphasis on the presence of behavioral noise. OSBNR mainly contains two steps. First, a cluster analysis is applied to groups similar instances of the minority class dataset in multiple behavior clusters. Second, we select and eliminate instances of the majority class, considered as behavioral noise, which overlap the behavioral clusters of the minority class.

This article is organized as follows: In Section II, common approaches to address the data imbalance are presented. Section III presents in detail our proposed approach to improve the classification of imbalanced data. While section IV reports the experiments and shows the results obtained. The final section concludes the paper and provides our insights into future works.

II. Common approaches for addressing data imbalance

Different approaches have been proposed to deal with imbalanced data problem and improve the performance of prediction [11], [12], [13]. Often, these approaches are based on either the data level or the algorithm level approaches. Data-level based approaches are independent of the classifier and use sampling methods to produce a well balanced dataset from the imbalanced training dataset. With regard to sampling, we can distinguish two methods which are undersampling and oversampling [14], [15], [16], [17]. Algorithm-level approach adapts existing classification learning algorithms to guide learning towards minority class. It requires a given specific knowledge of the corresponding classifier and of the application field [18], [19], [20], [21].

In reviewing the literature, the undersampling methods are the main focus in this paper. In this context, we refer to a popular work for undersampling, called "Condensed Nearest Neighbour Rule (CNN)" [22]. CNN works by eliminating the majority class samples that are distant from the decision border since these samples can be considered as less relevant for learning.

Another popular algorithm for undersampling, Tomek's Link removal (TL) was introduced in [23], This algorithm works by detecting pair of data points, called Tomek's Link, that are each other's nearest neighbor but have different class labels. Undersampling can be done by either removing all Tomek links or by removing the majority class data belonging to the Tomek Link.

In [24], Edited Nearest Neighbor Rule (ENN) was presented. It removes any instance whose class label is different from the class of at least two of its three nearest neighbors. The idea behind this technique is to remove the instances from the majority class near or around the borderline of different classes, in order to increase classification accuracy of minority instances rather than majority instances.

Another undersampling technique called Neighbourhood Cleaning Rule (NCR) was proposed by [25]. It uses Wilson's Edited Nearest Neighbour Rule (ENN) [24] to remove instances from the majority class when two out of three of the nearest neighbors of an instance contradict the class.

Two improvements to ENN are proposed in [26]: Repeated Edited Nearest Neighbor (RENN) and All-KNN (AKNN). Both methods make multiple passes over the training set repeating ENN. RENN just repeats the ENN algorithm until no further eliminations can be made from the edited set. AKNN repeats ENN for each sample using incrementation values of k(Number of Nearest Neighbor) each time and removing the sample if its label is not the predominant one at least for one value of k.

In [27] an undersampling approach called One-Side Selection (OSS) is proposed which is combination of Tomek's Link [23] followed by the application of CNN [22]. Tomek's Link is used to remove noisy and borderline majority class examples. Then, CNN will remove example from the majority class that are distant from decision border. Then a consistent subset of the majority class is formed.

III. Proposed Approach

Most classification learning algorithms are often biased toward the majority class due to the data imbalance distribution, which leads to a higher misclassification rate for the minority class instances [6]. Unfortunately, the minority class is usually the most important from a data analysis perspective. So, such misclassification of minority instances often has serious consequences in real applications. Another critical factor in real world imbalanced data concerns presence of a relatively large number of instances from the majority class located inside the minority class. This problem is commonly referred as overlapping data and we named it as behavioral noise. The behavioral noise implies the presence of a number of instances of the two classes being, to some degree, mixed in characteristic space. The presence of behavioral noise has a severe impact in learning processes. The generated models can become more complex, showing less generalization abilities, lower precision, and higher computational cost [28]. So, the classification performance depends on these two main problems: class imbalance and behavioral noise.

In this work, we focus on credit card fraud detection as a very imbalanced and overlapped data classification problem, where non-fraudulent samples are much more numerous than fraud samples [6]. As well, it is known that some users share a behavior where the transactions are similar. While the transaction behavior of some users resembles no behavior and may even behave like transactions unlike their labels. For example, if a user's credit card information is stolen, fraudsters will make several large transactions in a short time to maximize the benefits, while some normal users may also make large transactions in a short time for certain reasons. The normal behavior of an individual user is therefore close to fraud. We define this type of transaction as behavioral noise. The existence of behavioral noises pushes the system to judge certain fraudulent transactions as authentic transactions. This leads to an erroneous classification which can be costly. Undetected fraudulent transactions (false positive) are much

more serious and costly than detecting normal behavior as fraud (false negative). The cost of false positives is financial in nature, it varies according to the amount of the transaction. On the other hand, the cost of false negatives is measured in terms of customer dissatisfaction, and this latter can be resolved by strategies to compensate and retain customers.

The main objective of our approach, called: One Side Behavioral Noise Reduction (OSBNR), is to handle behavioral noise to improve the classification of the minority class instances. OSBNR consists of separating normal transactions (majority class instances) from fraudulent ones (minority class instances). Then, a cluster analysis is applied to group similar instances of the minority class containing the fraudulent transactions in several subsets, that form several behavior groups. The second step eliminates normal transactions behaviors, considered as behavioral noise, which overlap with the fraudulent transactions behaviors.

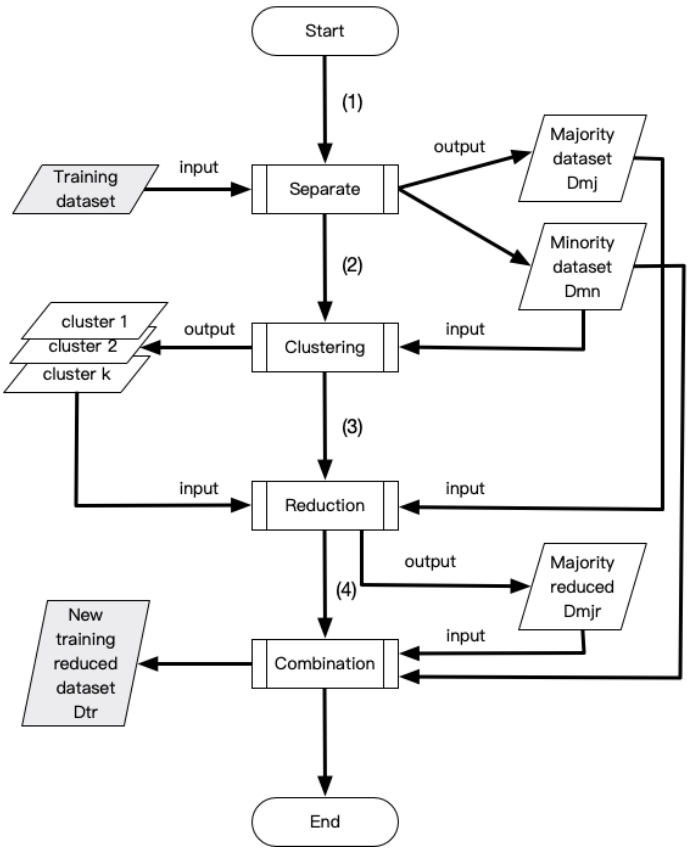

Fig. 1: Flowchart of the OSBNR approach

Figure 1 shows the main steps of OSBNR approach:

1) Separate: separates the majority Dmj and minority Dmn from the original training dataset.
2) Clustering: using k-means clustering algorithm [29] to form samples of similar minority instances into a number of behavior clusters. Each cluster seems to have distinct characteristics in the high-dimensional features space.
3) Reduction: carries out the behavioral noise reduction of the majority instances with those of the minority class. The Euclidean distance is used to measure the level of similarity between the different clusters centers of minority class and the majority class instances. So first, we calculate the furthest distance $dmax_i, 1 < i < k$, from minority instances to the cluster center $Cmin_i$ for each minority cluster according. Second, the distances $dmaj_{ij}$ between majority instances and different clusters centres of minority class C_i are obtained. As results, all majority instances in the minority clusters area are identified if $dmax_i \geq dmaj_{ij}$. So, they are considered as noisy instances and then eliminated.
4) Combination: we combine the reduced majority instances set $Dmjr$ with the minority instances set Dmn to have a new training dataset Dtr.

Accurate identification and elimination of these instances maximize the visibility of the minority class instances and at the same time minimize excessive elimination of data.

IV. Experiments and Results

A. Dataset Description

For this work, we use the Kaggle credit card fraud detection dataset [30]. It contains transactions made by credit card during two days of September 2013 by European card holders. Table I provides statistics for the dataset and shows that the minority class (fraud) accounts for 0.172% of all transactions. Therefore, this dataset is highly imbalanced [6]. It contains 31 numerical features. Since some of the input features contains financial information, the PCA transformation of 28 digital input features (named $V_1, \ldots, V_{28}$) were performed due to confidentiality issues. Three of the given features weren't transformed. Time feature shows the time between first transaction and every other transaction in the dataset. Amount feature is the amount's value spent in a single transaction made by credit card. Class feature represents the label, and takes only 2 values: value 1 in case of fraud transaction and 0 otherwise.

TABLE I: Kaggle credit card fraud dataset details

Transactions	Majority class	Minority class	Columns
284 807	284 315	492	31

B. Feature selection

Feature selection is a fundamental technique that selects the most relevant features from the given dataset. Choosing the right features wisely and removing the less important ones can reduce over-learning, improve accuracy, and reduce training time. Visualization techniques can be helpful in this process. Formally, we select a subset of features or attributes from the set of features and eliminate redundant features that do not contribute to the prediction performance. Thus, a feature is important when its data distribution of the two classes are

divergent. Therefore, this feature can potentially separate the two classes and improve prediction performance.

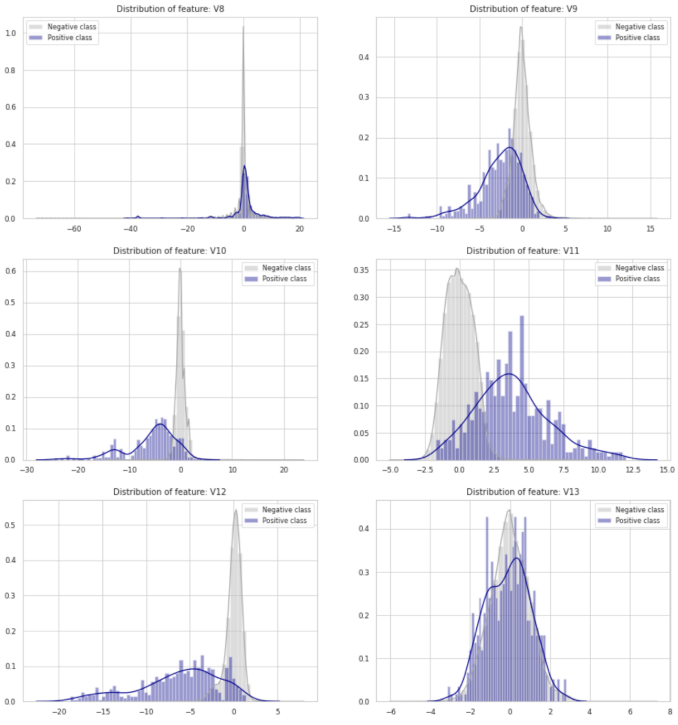

Fig. 2: Class distribution histogram on some features

Figure 2 shows the class distribution for some features of our dataset. We can see for V_9, V_{10}, V_{11} and V_{12} a significant divergence of class distribution. They are therefore features with a strong predictive power. So, we can keep them during the models construction. Similarly, we can see for feature V_8 and V_{13} that the distribution of normal transactions (majority class) corresponds to the distribution of fraudulent transactions (minority class). These features cannot effectively contribute to the separation between the two classes. We carried out this process for all 28 features. As a result, 11 relevant features were selected for our experiments: V_3, V_4, V_9, V_{10}, V_{11}, V_{12}, V_{14}, V_{16}, V_{17}, V_{18} and V_{19}.

C. Classifiers and resampling techniques

In this work, we applied various resampling techniques such as the Condensed Nearest Neighbour Rule (CNN) [22], Tomek's links (TL) [23], One-Side Selection (OSS) [27], Edited Nearest Neighbour Rule (ENN) [24], Repeated Edited Nearest Neighnor (RENN) [26], All-KNN (AKNN) [26] and Neighbor Cleaning Rule (NCR) [25]. We evaluated their performance with the proposed approach OSBNR using the best and widely used classifiers: Random Forest (RF) and Multilayer Perceptron (MLP) [31], [32], [33]:

- Random forest (RF) is an algorithm that consist of many decision trees. This algorithm works best when there are more trees in the forest. Each decision tree in the forest gives results. These results are merged in order to obtain a more precise and stable prediction [34].
- Multilayer perceptron (MLP) is an artificial neural network with direct action which is made up of at least 3 layers of nodes: entry layer, hidden layer and exit layer. Each node uses an activation function. The activation function calculates the weighted sum of its inputs and adds a bias. This allows us to decide which neuron should be removed and not taken into account in the external connections.

RF and MLP models parameters were determined from various preliminary tests carried out on the training data, as shown in Table II.

TABLE II: RF and MLP parameters used

Classifiers	Parameter
Random forest (RF)	Number of trees = 20 Depth of each tree = 8 Impurity = Gini
Multilayer perceptron (MLP)	Number of iterations = 100 Tolerance parameter = 1e-6

D. Evaluation Metrics

Evaluation metrics play an important role to assess and guide learning algorithms [11]. The common metric used is accuracy. However, accuracy is not a good indicator of the actual classification performance when the class distribution is not uniform, especially for the positive (minority) class. Indeed, because it has less effect on accuracy compared to the negative (majority) class. As in [35], we consider other metrics summarized as follows, where:

$$\begin{cases} FP & \text{false positive} \\ FN & \text{false negative} \\ TP & \text{true positive} \\ TN & \text{true negative} \end{cases}$$

- Precision or Positive Predictive Value (1): represents the proportion of positive samples that were correctly classified to the total number of positive predicted samples.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

- True Positive Rate (2): called Sensitivity or Recall, is the number of actual positives which are predicted positives.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

- F-measure or F1-score (3): represents the harmonic mean of precision and recall. The value ranges from 0 to 1, if the value is high then F-measure indicates high classification performance.

$$F1\text{-score} = 2\frac{Precision * Recall}{Precision + Recall} \tag{3}$$

- AUC (4): represents the ability to distinguish classes, which considers both the true positive rate TPR (2) and the false positive rate FPR (5). AUC is based on the consideration that the higher the true positive rate TPR,

and the lower false positive rate FPR, classification performance is better.

$$AUC = \frac{1 + TPR - FPR}{2} \tag{4}$$

Where False Positive Rate (FPR (5)) represents the proportion of legitimate samples that were wrongly predicted as fraud.

$$FPR = \frac{FP}{FP + TN} \tag{5}$$

V. Results Analysis

A. Training and test datasets used for the experiments

We present different experiments to compare the performance of our proposed OSBNR approach and the state-of-art resampling methods (CNN, ENN, AKNN, RENN, TL, OSS, and NCR). As there is no rule-of-thumb for how to divide a dataset into training and test sets, we have noticed that in the case of the $70/30$ rule, the percentages of the minority class of the training and test sets are: 80%, 20% respectively of the total fraud. In order to demonstrate the effectiveness of the proposed OSBNR method to deal with the imbalance and overlapping between classes problem, we studied 3 different divisions of the dataset by resampling the minority class based on the ratios: $80/20, 70/30, 60/40$, and the majority class based on the $70/30$ ratio. Table III presents training and test datasets used as input of our OSBNR approach (Fig. 1).

TABLE III: Training and test datasets used for the experiments

		Total	Majority class	Minority class
100%	**Dataset**	284 807	284315 - 100%	492 - 100%
Rule 80/20	**Training**	199 413	199020 - 70%	393 - **80%**
	Test	85 394	85295 - 30%	99 - **20%**
Rule 70/30	**Training**	199 364	199020 - 70%	344 - **70%**
	Test	85 443	85295 - 30%	148 - **30%**
Rule 60/40	**Training**	199 315	199020 - 70%	295 - **60%**
	Test	85 492	85295 - 30%	197 - **40%**

B. Performance study of the OSBNR: case of all features

In this section, we analyze the impact of the proposed OSBNR approach on the performance of each classifier by comparing it with existing resampling methods taking into account all the features and according to 3 different divisions of the dataset. The results are calculated for four metrics: AUC, Precision, Recall and F1-score.

Figures 3 and 4 show the results using the AUC metric for the RF and MLP classifiers respectively. The most interesting observation is that the proposed OSBNR offers significantly better performance compared to the other methods for the two classifiers from the AUC point of view for all the distributions of training and test sets. For the RF classifier, the best score is obtained by RF_OSBNR with an AUC value of ($AUC = 0.9341$), RF_CNN takes second place with a score of ($AUC = 0.9079$), when the training and test sets are set to $80/20$ rule. For the MLP classifier, the best AUC score rule is obtained by MLP_OSBNR with a value of ($AUC = 0.9487$), followed

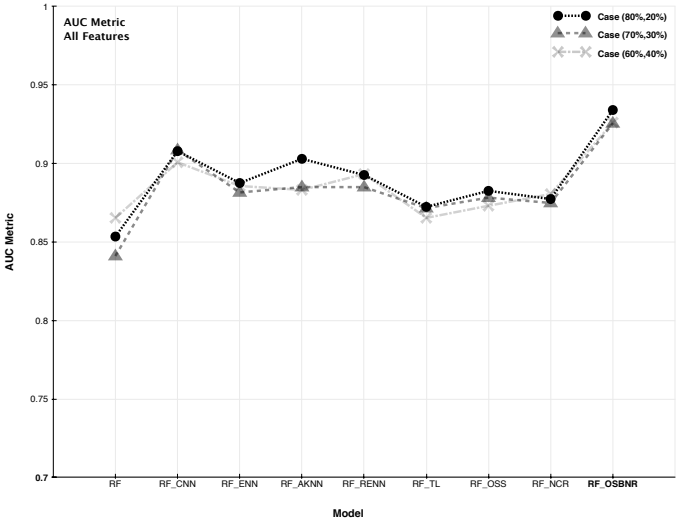

Fig. 3: AUC metric for OSBNR and the reference resampling approaches: RF as base classifier case of all features

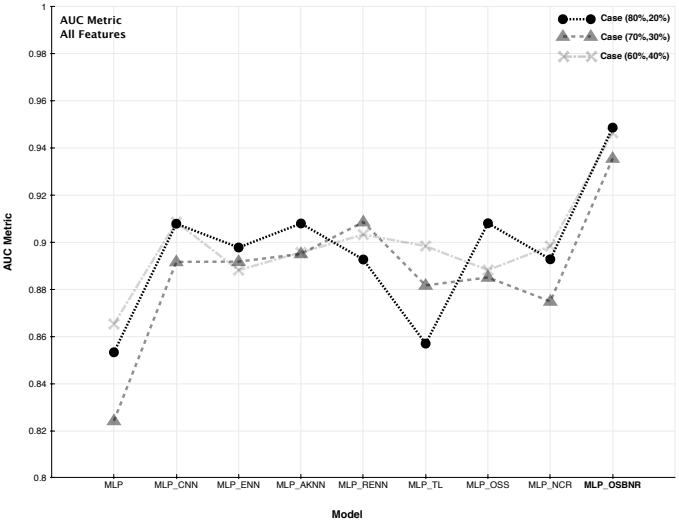

Fig. 4: AUC metric for OSBNR and the reference resampling approaches: MLP as base classifier case of all features

by MLP_OSS with a score of ($AUC = 0.9079$) when the training and test sets are set to $80/20$ rule.

Similarly, Figures 5 and 6 present the results in terms of precision. The results illustrated in Figure 5 show that OSBNR outperforms the other resampling methods for all the distributions of the training and test sets. The best precision score for the RF classifier is obtained by the OSBNR approach when the training and test sets are set at 80% and 20% fraud with a score of ($Precision = 0.8686$), while RF_CNN takes second place with a score of ($Precision = 0.8163$). Similar in MLP, as illustrated in Figure 6, it is clear that the best precision score is obtained by MLP_OSBNR with a score of ($Precision = 0.8979$), followed by MLP_OSS with a score of ($Precision = 0.8511$). Based on these results, we

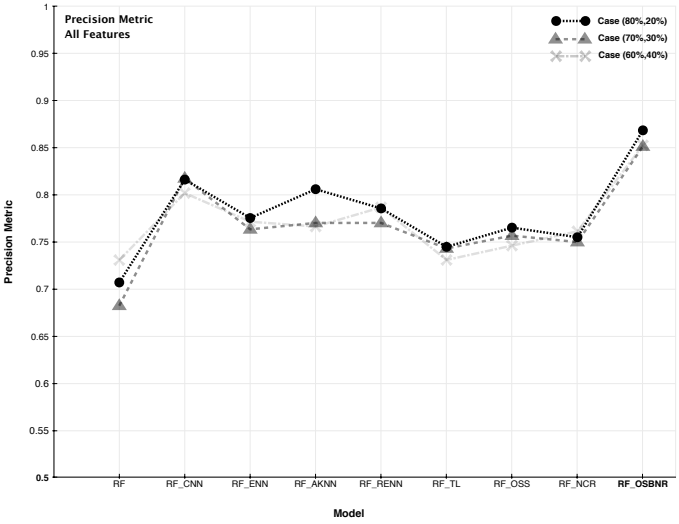

Fig. 5: Precision metric for OSBNR and the reference resampling approaches: RF as base classifier case of all features

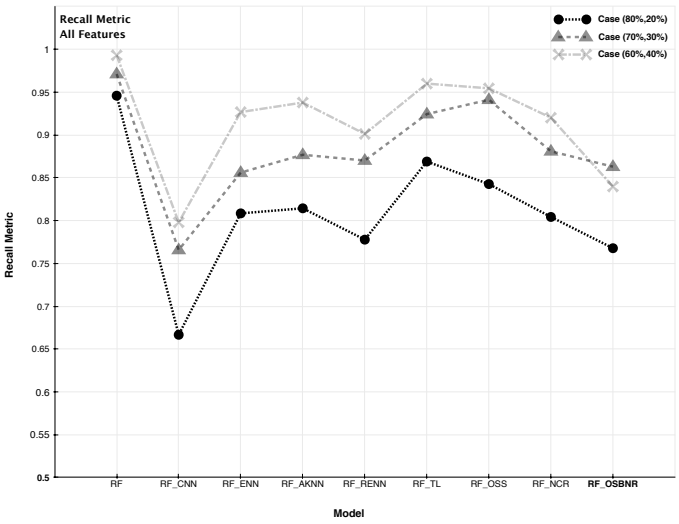

Fig. 7: Recall metric for OSBNR and the reference resampling approaches: RF as base classifier case of all features

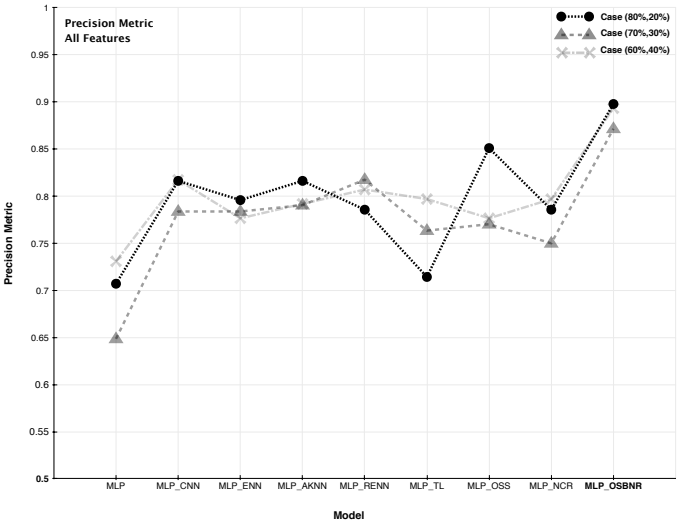

Fig. 6: Precision metric for OSBNR and the reference resampling approaches: MLP as base classifier case of features

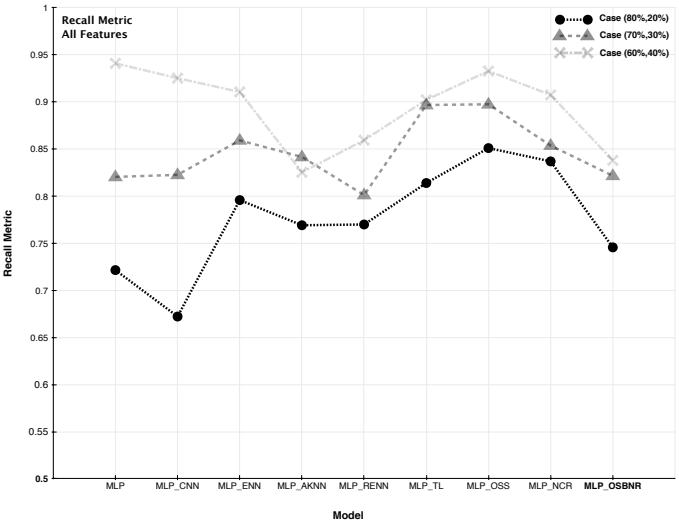

Fig. 8: Recall metric for OSBNR and the reference resampling approaches: MLP as base classifier case of all features

conclude that the proposed OSBNR can significantly improve the recognition rate of minority samples.

Figures 7 and 8 show the results obtained using Recall measure. The most interesting observation is that the RF and MLP classifiers without preprocessing clearly offer the best performance in terms of recall metric. Such a result was expected in a way, because the resampling methods were introduced to manage class imbalance and class overlap problems. They eliminate the instances of the majority class considered as noise to improve the prediction of instances of the minority class, and this can slightly increase the rate of false negatives. For the RF classifier, the second best recall score is obtained by RF_OSS with a value of ($Recall = 0.96$) when the training and test sets are set at 60% and 40% fraud. Similarly, for MLP the second place is occupied by MLP_OSS with a score of ($Recall = 0.9329$) with the $60/40$ rule.

As for measure F1-score, Figures 9 and 10 present the results of all the resampling methods by applying the two learning classifiers. We see that the OSBNR approach outperforms the other resampling methods for all the distributions of the training and test sets for the two classifiers. For RF, the best F1-score is obtained by RF_OSBNR with a score of ($F1 - score = 0.8572$) according to the $70/30$ rule, followed by RF_AKNN with a score of ($F1 - score = 0.8436$) when applying rule $60/40$ to divide the training and test sets. Regarding MLP, the best score is obtained by MLP_CNN with a F1-score of 0.8679 according to the $60/40$ rule, MLP_OSBNR takes second place with a value of 0.8648.

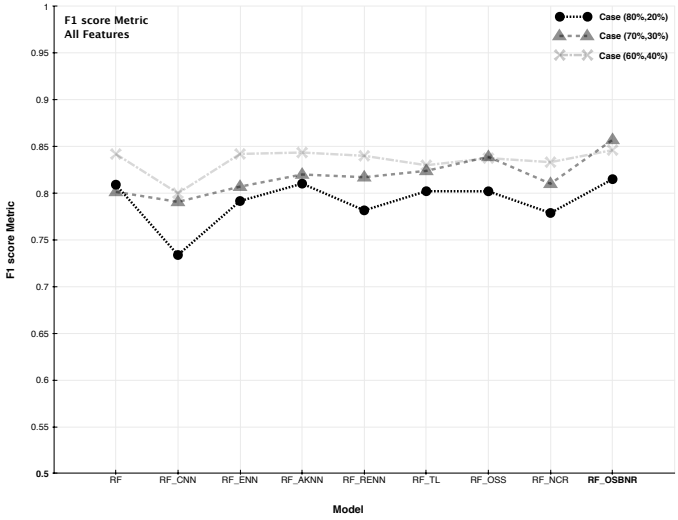

Fig. 9: F1-score metric for OSBNR and the reference resampling approaches: RF as base classifier case of all features

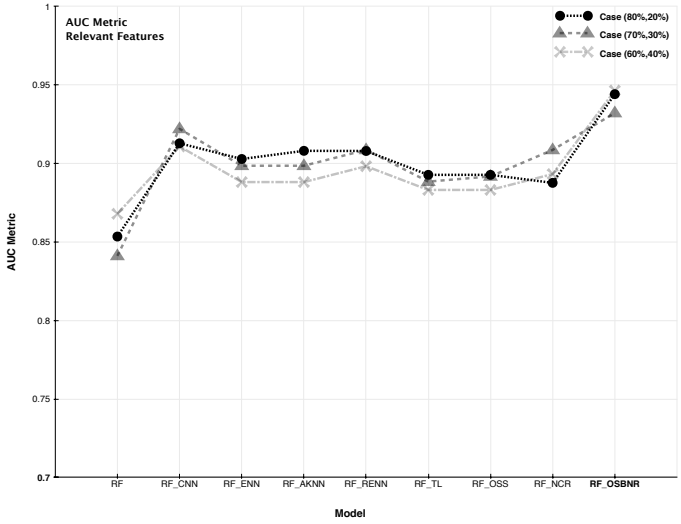

Fig. 11: AUC metric for OSBNR and the reference resampling approaches: RF as base classifier case of relevant features

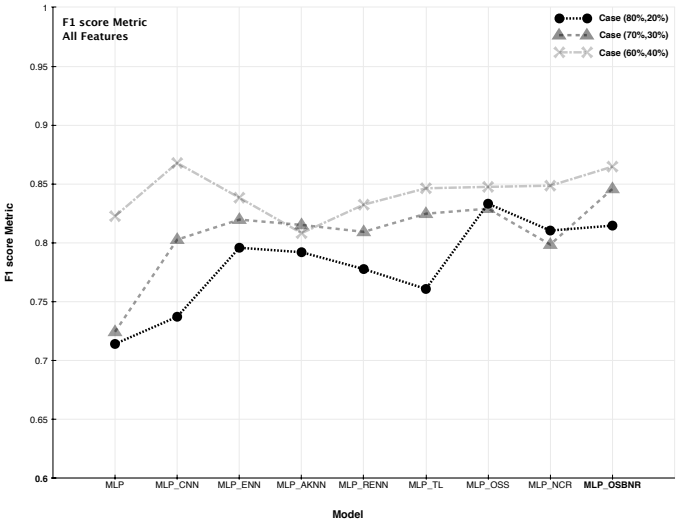

Fig. 10: F1-score metric for OSBNR and the reference resampling approaches: MLP as base classifier case of all features

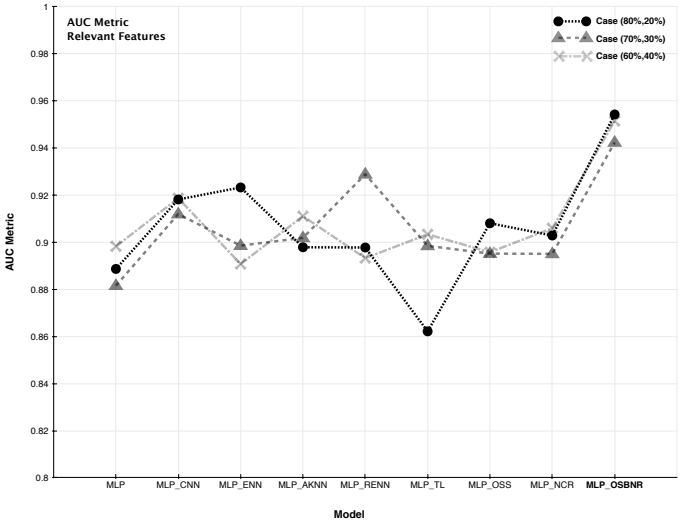

Fig. 12: AUC metric for OSBNR and the reference resampling approaches: MLP as base classifier case of relevant features

C. Performance study of the OSBNR: case of relevant features

In this section, we analyze the impact of the proposed OSBNR on the performance of each classifier by comparing it with existing resampling methods taking into account relevant features and also according to the 3 different divisions of the dataset. The results are calculated for four metrics: AUC, Precision, Recall and F1-score.

In order to better situate the results, we start by reporting on AUC metric from Figures 11 and 12. We can see that the best AUC scores are obtained by RF and MLP combined with OSBNR for all the distributions of the sets of training and testing. For the RF classifier, the best score is obtained by RF_OSBNR with an AUC value of ($AUC = 0.9442$), RF_CNN takes second place with a score of ($AUC = 0.9129$) according to the $80/20$ rule. Regarding the MLP classifier, the best AUC score is obtained by MLP combined with OSBNR when applying the $80/20$ rule with a value of ($AUC = 0.9543$), followed by MLP_RENN with a score of ($AUC = 0.9289$), while MLP_OSS maintained its score with the same value of ($AUC = 0.9079$). From these results, we can conclude that after eliminating redundant features that do not contribute to performance, we get continuity or even improvement in predictive performance.

Similarly, the figures 13 and 14 present the results in terms of precision. The results illustrated in Figure 13 show that the best precision score for the RF classifier is obtained by the OSBNR approach for all the distributions of training and test sets with a best value score of ($Precision = 0.8934$)

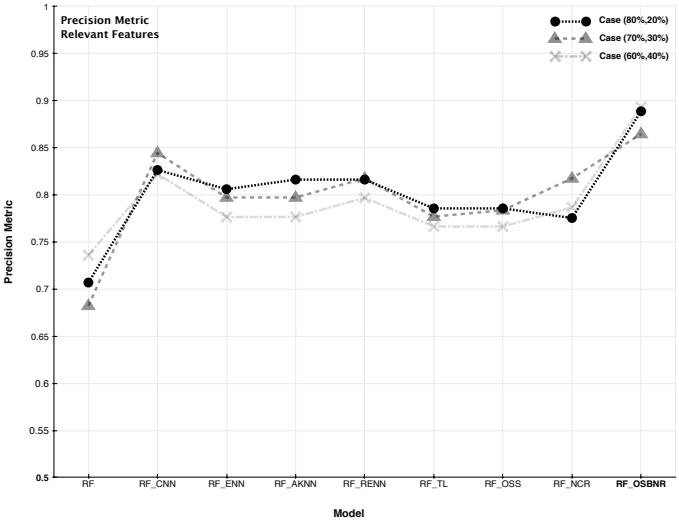

Fig. 13: Precision metric for OSBNR and the reference resampling approaches: RF as base classifier case of relevant features

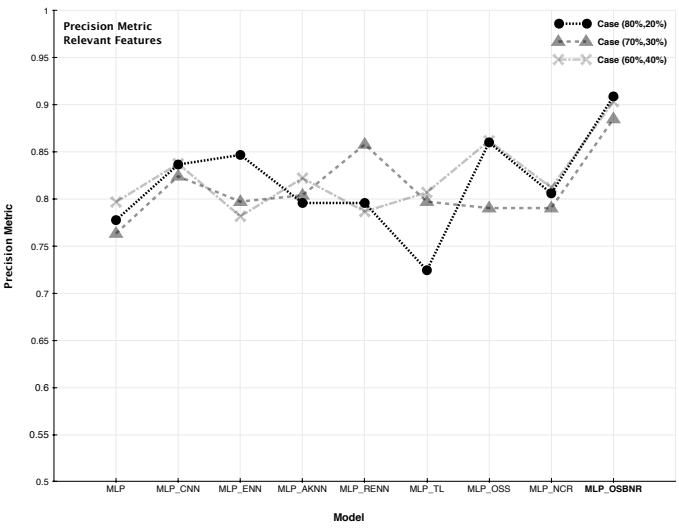

Fig. 14: Precision metric for OSBNR and the reference resampling approaches: MLP as base classifier case of relevant features

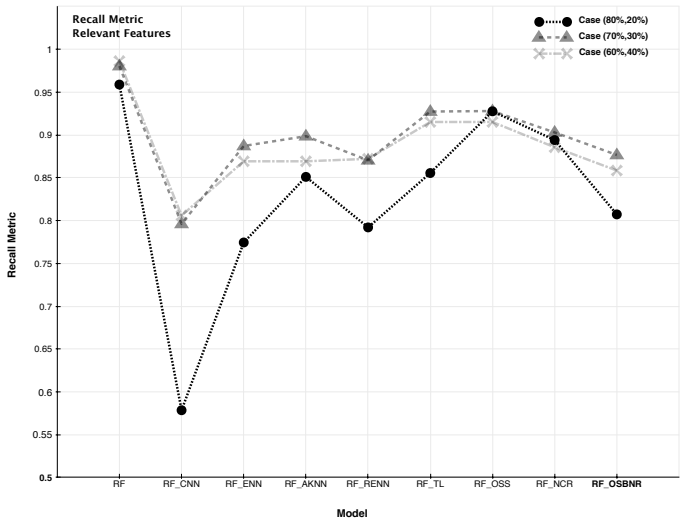

Fig. 15: Recall metric for OSBNR and the reference resampling approaches: RF as base classifier case of relevant features

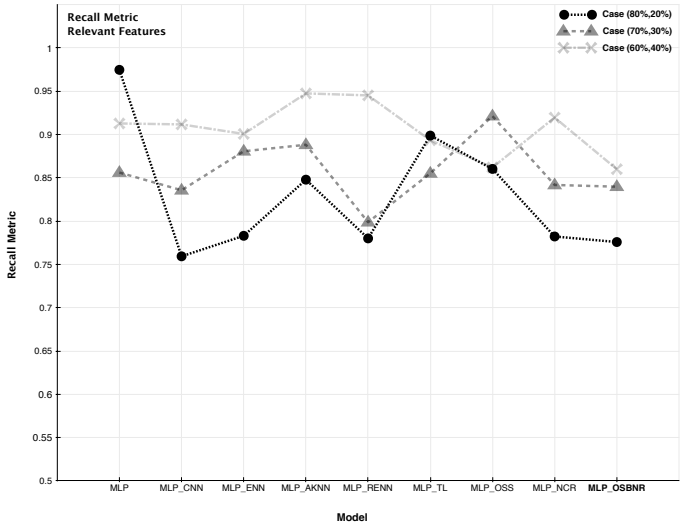

Fig. 16: Recall metric for OSBNR and the reference resampling approaches: MLP as base classifier case of relevant features

according to $80/20$ rule, which means that this model offers a better prediction of minority instances. Similar in MLP, As illustrated in Figure 14, it is clear that the best precision score is obtained by MLP_OSBNR with a best score of ($Precision = 0.909$).

With regard to the Recall measure and according to the $70/30$ rule, Figures 15 and 16 show the results obtained. we also notice that the RF and MLP classifiers without preprocessing clearly offer the best performance in terms of recall metric. For the RF classifier, the second best recall score is obtained by RF_OSS with a value of ($Recall = 0.928$). Similarly, for MLP the second place is occupied by MLP_AKNN with a score of ($Recall = 0.9474$).

As for measure F1-score, Figures 15 and 16 show the results of all the resampling methods by applying the two learning classifiers. For RF, we see that the OSBNR outperforms the other resampling methods for all the distributions of the training and test sets. Likewise for MLP, the best score is obtained by MLP combined with OSBNR for all the distributions of training and test sets.

VI. Conclusion and Future Work

In this paper, we present a new approach called: One Side Behavioral Noise Reduction (OSBNR) to deal with class imbalance problem, with a particular emphasis on the presence of

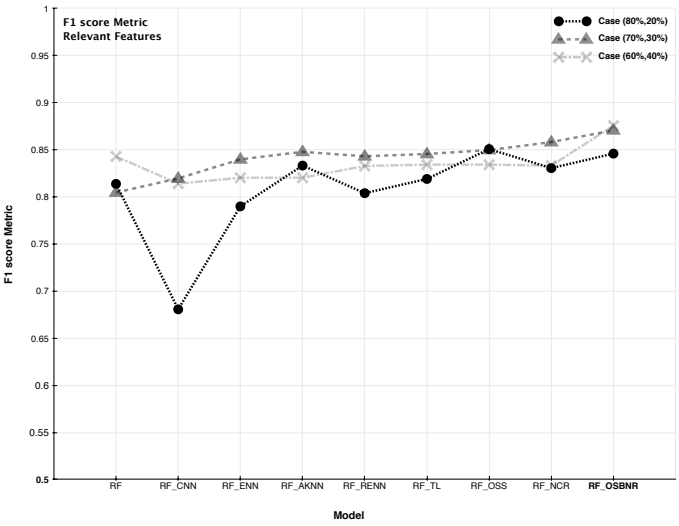

Fig. 17: F1-score metric for OSBNR and the reference resampling approaches: RF as base classifier case of relevant features

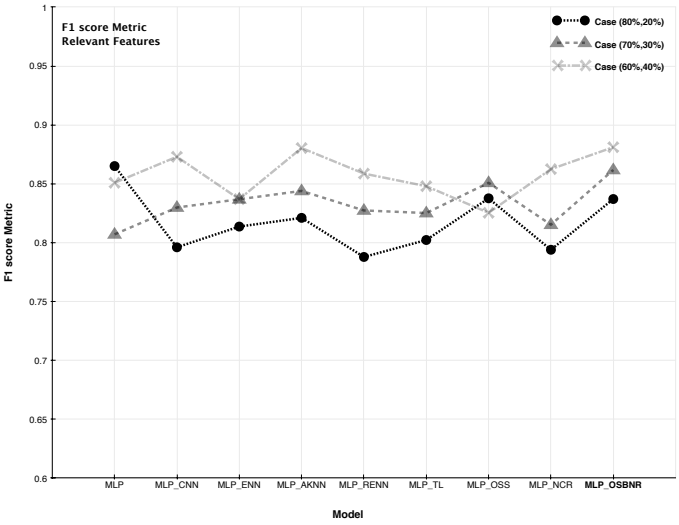

Fig. 18: F1-score metric for OSBNR and the reference resampling approaches: MLP as base classifier case of relevant features

behavioral noise. Our approach combines clustering analysis and a behavioral noise reduction process.

To study the effectiveness of the proposed method, we compare it to several state-of-the-art resampling methods. Two learning classifier, namely Random Forest and MultiLayer Perceptron, have been tested over these resampling methods. Experimental results measured using four metrics (AUC, Precision, Recall, F1-score) indicate that OSBNR achieves much better classification performance than the other compared methods to deal with noise data problem with a significant difference.

This work constitute an important part of the framework in development. Thus, we wish to study the behavior of the scaling of our approach in the context of real applications. This will raise two fundamental questions:

- Confidence and predictability of predictions for decision making. The main objective of our explanatory approach to machine learning is to propose methods to understand and explain how the system produces its decisions in case of real domain application.
- Notion of uncertainty in machine learning which is of major importance and constitutes a key element of modern machine learning methodology. It has gained in importance due to the increasing relevance of machine learning in real applications.

ACKNOWLEDGMENT

This work is carried out thanks to the support of the European Union through the PLAIBDE project of the FEDER-FSE operational program for the Nouvelle-Aquitaine region, France. The project is supported by aYaline company, with partners: LIAS-ENSMA laboratory in Poitiers, and the L3i laboratory at La Rochelle University.

REFERENCES

[1] A. Ali, S. M. Shamsuddin, and A. Ralescu, "Classification with class imbalance problem: a review," *Int. J. Advance Soft Compu. Appl*, vol. 7, no. 3, pp. 176–204, 2015.

[2] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.

[3] H. Alberto Fernández, L. Salvador García, G. Mikel, C. P. Ronaldo, K. Bartosz, and H. Francisco, *Learning from imbalanced data sets*. Springer International Publishing, 2018.

[4] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent data analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[5] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Special issue on learning from imbalanced data sets," *ACM SIGKDD explorations newsletter*, pp. 1–6, 2004.

[6] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating probability with undersampling for unbalanced classification," in *2015 IEEE Symposium Series on Computational Intelligence*, pp. 159–166, IEEE, 2015.

[7] C. Elkan, "The foundations of cost-sensitive learning," in *International joint conference on artificial intelligence*, pp. 973–978, Lawrence Erlbaum Associates Ltd, 2001.

[8] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Folleco, "An empirical study of the classification performance of learners on imbalanced and noisy software quality data," *Information Sciences*, vol. 259, pp. 571–595, 2014.

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

[10] R. C. Prati, G. E. Batista, and M. C. Monard, "Class imbalances versus class overlapping: an analysis of a learning system behavior," in *Mexican international conference on artificial intelligence*, pp. 312–321, Springer, 2004.

[11] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *International journal of pattern recognition and artificial intelligence*, vol. 23, no. 04, pp. 687–719, 2009.

[12] M. A. U. H. Tahir, S. Asghar, A. Manzoor, and M. A. Noor, "A classification model for class imbalance dataset using genetic programming," *IEEE Access*, vol. 7, pp. 71013–71037, 2019.

[13] Y. Sui, M. Yu, H. Hong, and X. Pan, "Learning from imbalanced data: A comparative study," in *International Symposium on Security and Privacy in Social Networks and Big Data*, pp. 264–274, Springer, 2019.

[14] W.-C. Lin, C.-F. Tsai, Y.-H. Hu, and J.-S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Information Sciences*, vol. 409, pp. 17–26, 2017.

[15] M. Rahman and D. N. Davis, "Cluster based under-sampling for unbalanced cardiovascular data," in *Proceedings of the World Congress on Engineering*, vol. 3, pp. 3–5, 2013.

[16] J. Zhang, T. Wang, W. W. Ng, S. Zhang, and C. D. Nugent, "Undersampling near decision boundary for imbalance problems," in *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, pp. 1–8, IEEE, 2019.

[17] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and smote," *Information Sciences*, vol. 465, pp. 1–20, 2018.

[18] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358–3378, 2007.

[19] P. Cao, D. Zhao, and O. Zaiane, "An optimized cost-sensitive svm for imbalanced data learning," in *Pacific-Asia conference on knowledge discovery and data mining*, pp. 280–292, Springer, 2013.

[20] C. L. Castro and A. P. Braga, "Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data," *IEEE transactions on neural networks and learning systems*, vol. 24, no. 6, pp. 888–899, 2013.

[21] M. Galar, A. Fernández, E. Barrenechea, and F. Herrera, "Eusboost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling," *Pattern recognition*, vol. 46, no. 12, pp. 3460–3471, 2013.

[22] P. Hart, "The condensed nearest neighbor rule (corresp.)," *IEEE transactions on information theory*, vol. 14, no. 3, pp. 515–516, 1968.

[23] I. Tomek, "Two modifications of cnn," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 7(2), pp. 679–772, 1976.

[24] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Transactions on Systems, Man, and Cybernetics*, no. 3, pp. 408–421, 1972.

[25] J. Laurikkala, "Improving identification of difficult small classes by balancing class distribution," in *Conference on Artificial Intelligence in Medicine in Europe*, pp. 63–66, Springer, 2001.

[26] I. Tomek, "An experiment with the edited nearest-neighbor rule.," *IEEE Transactions on Systems, Man, and Cybernetics*, 1976.

[27] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," in *Icml*, vol. 97, pp. 179–186, Nashville, USA, 1997.

[28] X. Zhu and X. Wu, "Class noise vs. attribute noise: A quantitative study," *Artificial intelligence review*, vol. 22, no. 3, pp. 177–210, 2004.

[29] K. Alsabti, S. Ranka, and V. Singh, "An efficient k-means clustering algorithm," 1997.

[30] K. Inc, "Credit card fraud detection: Anonymized credit card transactions labeled as fraudulent or genuine," 2013.

[31] E. H. Salma, M. Jamal, B. Mohammed, and F. Bouziane, "Machine learning for anomaly detection. performance study considering anomaly distribution in an imbalanced dataset," in *2020 5th International Conference on Cloud Computing and Artificial Intelligence Technologies and Applications (Cloudtech'20)*, IEEE, 2020.

[32] V. Jonnalagadda, P. Gupta, and E. Sen, "Credit card fraud detection using random forest algorithm," 2019.

[33] A. M. Mubarek and E. Adalı, "Multilayer perceptron neural network technique for fraud detection," in *2017 International Conference on Computer Science and Engineering (UBMK)*, pp. 383–387, 2017.

[34] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[35] C. Priscilla and D. Prabha, "Credit card fraud detection: A systematic review," in *International Conference on Information, Communication and Computing Technology*, pp. 290–303, Springer, 2019.