Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Detection of fraudulent credit card transactions: A comparative analysis of data sampling and classification techniques

Detection of fraudulent credit card transactions: A comparative analysis of data sampling and classification techniques

Konduri Praveen Mahesh[1], Shaik Ashar Afrouz[1], Anu Shaju Areeckal[1]*

[1]Department of Electronics & Communication Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Karnataka, India

*Corresponding author: anu.areeckal@manipal.edu

Abstract- Every year there is an increasing loss of a huge amount of money due to fraudulent credit card transactions. Recently there is a focus on using machine learning algorithms to identify fraud transactions. The number of fraud cases to non-fraud transactions is very low. This creates a skewed or unbalanced data, which poses a challenge to training the machine learning models. The availability of a public dataset for this research problem is scarce. The dataset used for this work is obtained from Kaggle. In this paper, we explore different sampling techniques such as under-sampling, Synthetic Minority Oversampling Technique (SMOTE) and SMOTE-Tomek, to work on the unbalanced data. Classification models, such as k-Nearest Neighbour (KNN), logistic regression, random forest and Support Vector Machine (SVM), are trained on the sampled data to detect fraudulent credit card transactions. The performance of the various machine learning approaches are evaluated for its precision, recall and F1-score. The classification results obtained is promising and can be used for credit card fraud detection.

Keywords: Fraud detection, Classification, Unbalanced data, Data sampling, Credit card,

1.Introduction

In the consumer industry, fraud transactions are the illegal usage of credit or debit card of a person without his knowledge. Credit card fraud transactions may occur either in an authorized manner where the authorised owner is misled to make payments to a fake account, or in an unauthorized manner where the transaction is carried out without authorization from the card holder. In 2018, the United Kingdom reported unauthorized fraudulent transactions amounting to £844.8 million [1]. In India, the number of fraudulent transactions in 2019 is over 365 cases [2]. Hence, it is the need of the hour to develop efficient methods for detection and prevention of fraudulent transactions.

Generally, the credit card fraud transactions are much lesser in number when compared to the non-fraudulent transactions. Hence, the user data is skewed in nature. This imbalance in data poses a challenge when employing machine learning approaches for fraud detection in credit card transactions.

The aim of this paper is to explore data sampling techniques to balance the dataset. Different classifier models are trained with the balanced dataset and their performance are compared. This work provides a clear insight on a particular classifier which outperforms the other classifiers in predicting the fraud cases accurately.

The remaining paper discusses related work in the research topic in section 2, proposed approach for fraud detection in section 3, classification results and comparison with related work in section 4, and conclusion in section 5.

2.Related work

There are many research work done on fraud detection in credit card transactions. Dal Pozzolo A et al. worked on incremental learning along with using sampling techniques and final prediction was done using ensemble of those models [3]. They observed better results with random forest classifier along with Synthetic Minority Oversampling Technique (SMOTE) sampling technique.

Najadat H et al. compared different decision tree splitting criteria and they came up with a new skew measure, Hellinger distance [4]. By using Hellinger distance, they proposed its application in forming decision tree for a better performance.

Drummond C and Holte RC provided a new insight on the two sampling techniques [5]. It directs us to adapt the machine learning algorithms to imbalanced data by balancing it. They observed that under-sampling technique beats over-sampling. Their research signified the use of random forest classifier with data balancing, to achieve better prediction accuracy.

Kumar MS et al. used random forest for detecting fraudulent transactions and accuracy [6]. They have evaluated its performance based on confusion matrix and have obtained an accuracy of 90%.

Sadineni PK dealt with various machine learning techniques to detect fraudulent transactions [7]. An accuracy of 99.21% was obtained for random forest, 98.47% for decision tree, 95.55% for logistic regression, 95.16% for support vector machine (SVM) and 99.92 % for Artificial Neural Network (ANN).

Sailusha R et al. employed random forest and Adaboost algorithm to detect fraudulent transactions [8]. Performance was analysed by accuracy and F1-score. They obtained accuracy of 100% for random forest and 99.9% for Adaboost.

In our paper, we have sampled our datasets using three different sampling techniques, to overcome the imbalance in the data. The balanced dataset was then trained on four different classifier models and validated on a public dataset.

3.Proposed methodology

The proposed methodology is subdivided into the following stages: input dataset, data sampling to balance the data, training classifier models and detection of fraudulent cases. An overview of the methodology is given in figure 1.

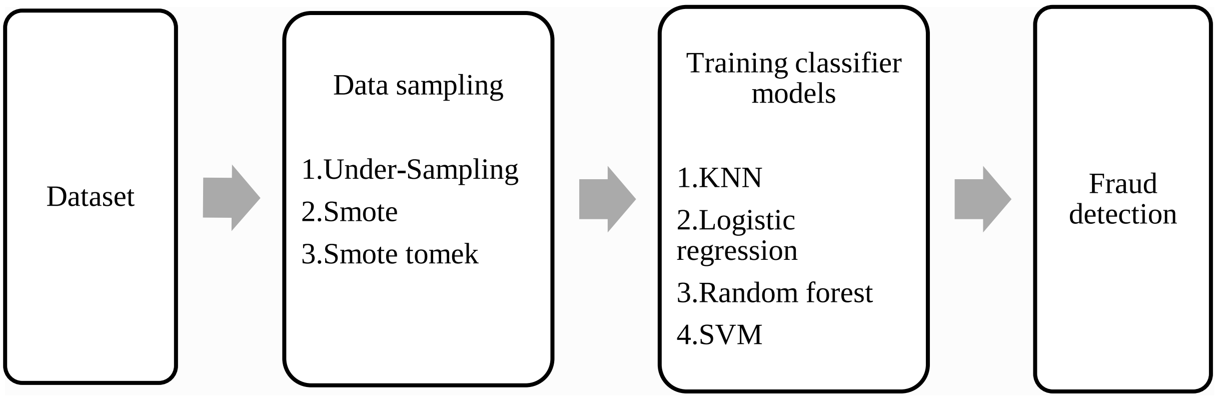

Figure 1. Overview of the proposed approach.

3.1 Dataset used

This work uses a public dataset which is available as part of a challenge task in Kaggle [3,9]. It contains 284,807 credit card transactions collected for two days in 2013, with 492 fraudulent cases. Since the fraudulent transactions present in the dataset are very less, we can differentiate the fraudulent and non-fraudulent transactions as minority and majority classes respectively. Hence, the dataset is highly unbalanced. Due to privacy concerns, the data sample characteristics was transformed using Principal Component Analysis (PCA) in the available dataset. The characteristics of the data samples are available as 28 principal components, V1 to V28, obtained with PCA. The only characteristics of the transactions that are not modified with PCA are 'Time' and 'Amount'. The target label is given as class 0 for genuine transaction and class 1 for fraudulent transaction.

3.2 Data sampling techniques

The dataset used is highly imbalanced. In order to balance the data of both the classes, data sampling techniques can be used. In the proposed methodology, three data sampling techniques are explored, namely under-sampling, Synthetic Minority Oversampling Technique (SMOTE) and SMOTE-Tomek.

3.2.1 Under-sampling. In this sampling method, we deal with imbalance in dataset by considering random samples of majority class in equal ratio with minority class. This technique reduces the number of data samples present in the majority class. The random sampling is done till all the classes get an equal number of data samples each. This technique has the disadvantage of reducing the overall size of the dataset.

3.2.2 SMOTE. SMOTE data augmentation is an oversampling technique for the minority class [10]. In this technique, new samples are created in the minority class by synthesizing the old samples. The basic principle is to draw a line between the data samples of the minority class and then a new data sample is created at a point on the line drawn, as shown in figure 2. It thus selects data samples that are close together in the minority class. SMOTE helps to overcome the overfitting problem posed by random oversampling.

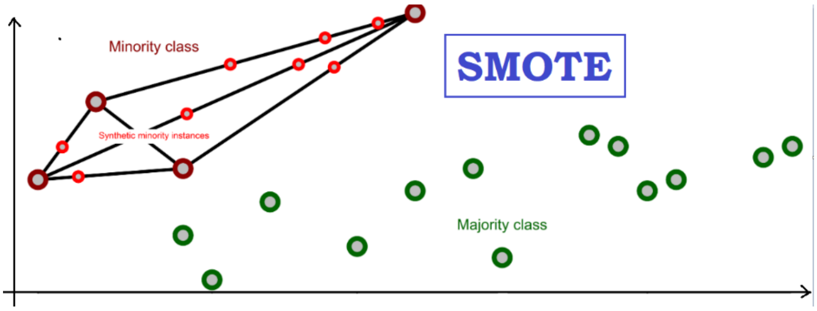

Figure 2. An illustration of the SMOTE sampling technique [11].

3.2.3 SMOTE-Tomek. To get the best of the under-sampling and over-sampling techniques, we use SMOTE-Tomek. SMOTE-Tomek is a hybrid technique that tries to clear overlapping data points in sample space for each of the classes [12]. The class clusters may be invading each other's space after the over-sampling by SMOTE. As a result, the model of the classifier will be overfit. Tomek linkages are paired samples of the opposite class that are the nearest neighbours to each other. As a result, the bulk of class observations from these linkages have been eliminated because it is thought that this will increase class separation around the decision borders. Tomek links are now applied to oversampled minority class samples created by SMOTE, in order to achieve better class clusters. Figure 3 shows an illustration of the SMOTE-Tomek sampling method.

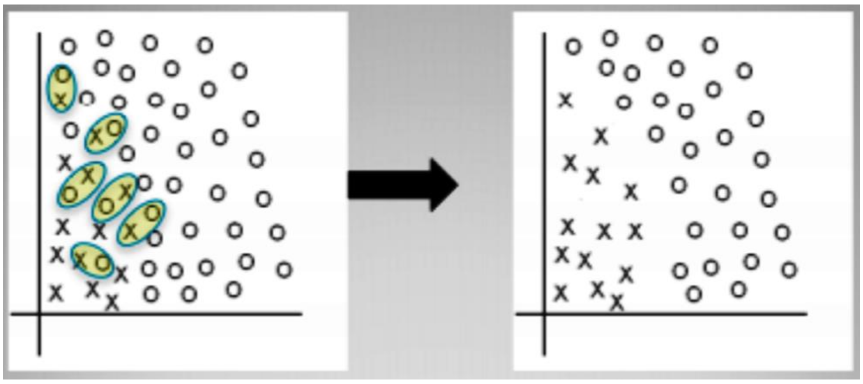

Figure 3. An illustration of the SMOTE-Tomek method [13].

3.3 Classifiers

After balancing the majority and minority classes in the dataset using data sampling techniques, the data is used to train classifier models. The classifiers used are k-Nearest Neighbour (KNN), logistic regression, random forest and SVM.

3.3.1 KNN. In KNN algorithm, we first calculate the distance between the unknown data sample and all the labelled samples. We then look for the k nearest samples and label it by the class of dominant number of sample surrounding the unknown sample. Euclidean distance is used as the distance measure. We decrease the Euclidean distance between samples by normalizing or scaling the data. Figure 4 shows the basic working of the KNN classifier.

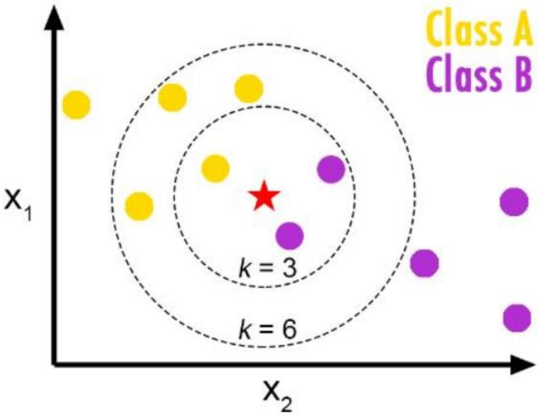

Figure 4. An illustration of KNN [14].

3.3.2 Logistic Regression. It is represented by a sigmoid function, as shown in figure 5. Input values are combined linearly using weights or coefficients to predict output value. It is used to model a binary variable which takes only two values 0 and 1. Its objective is to develop a mathematical equation that can give us a score between 0 and 1.

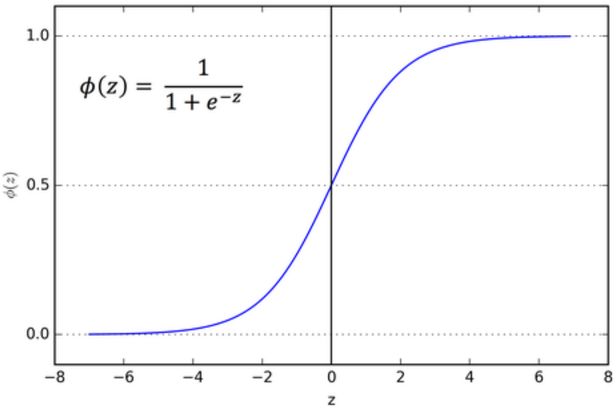

Figure 5. Logistic regression [14].

3.3.3 Random forest. This machine learning algorithm is an ensemble bagging technique. In this classification method, we split the training dataset into various bootstrap samples, which contain random *m* rows and *n* columns from the training dataset. These bootstrap samples are fed into multitude of decision trees. The majority vote of the decision tress is considered as the output prediction. Figure 6 shows the basic structure of the random forest classifier.

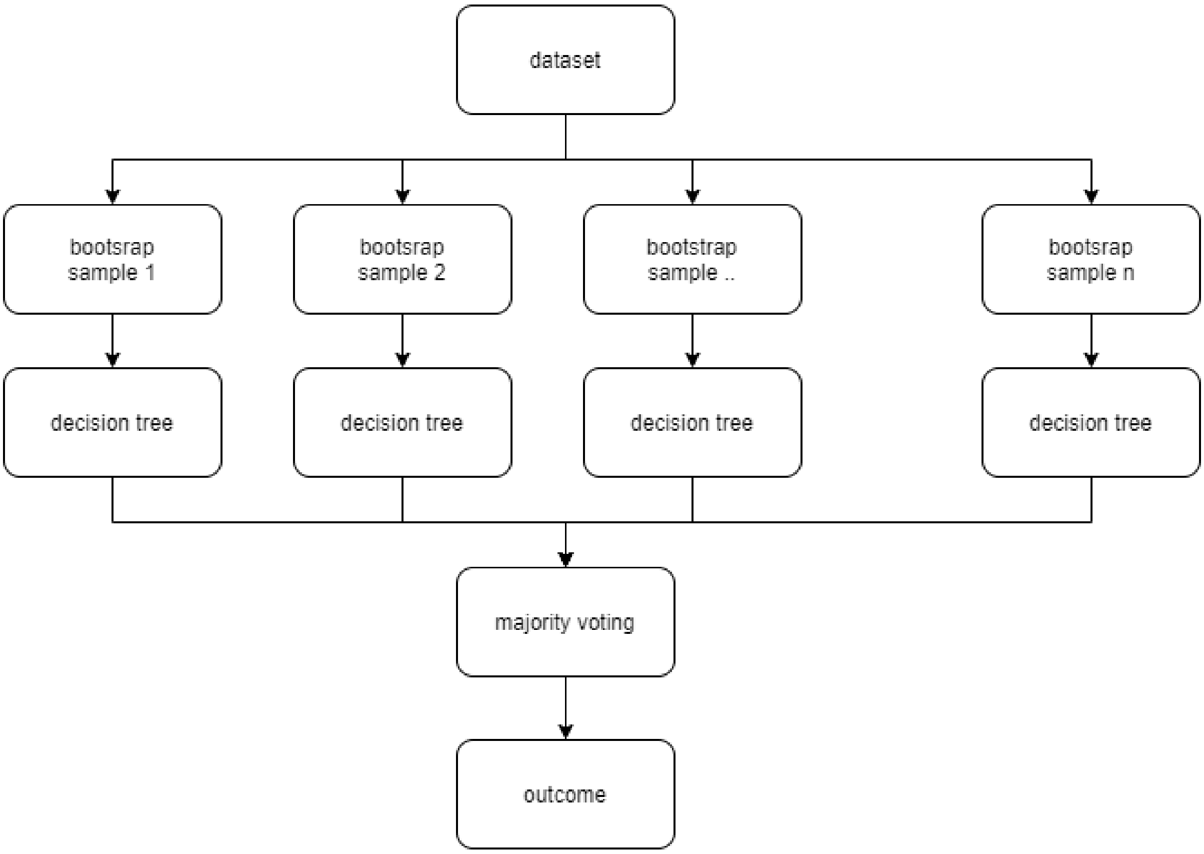

Figure 6. Decision tree of random forest classifier.

3.3.4 SVM. In this technique, a hyperplane is created to separate the two classes. Two parallel hyperplanes are created next to the main hyperplane in such a way that it passes through the nearest points in both the classes. The distance between the two hyperplanes is called margin. The optimal hyperplane is selected such that the marginal distance between the parallel planes is maximum, as shown in figure 7. Marginal planes pass through the points of the classes, known as support vectors. The class to which the data sample belongs to is predicted by using the hyperplanes. Since our data is not linearly separable, the radial basis function kernel has been used for our work.

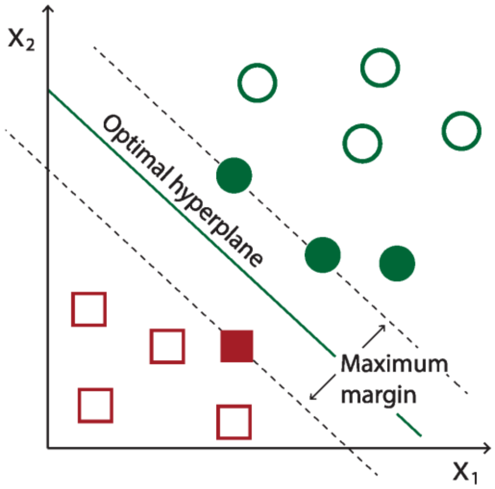

Figure 7. A denotation of Support Vector Machine [15].

3.4 Evaluation metrics

To determine the efficiency of the trained classifiers, precision, recall and F1-score are used as the evaluation metrics.

3.4.1 Precision. Precision is the number of positive class values predicted to the total positive predictions.

$$Precision = \frac{True\ Positives}{(True\ Positives)+(False\ Positives)} \tag{1}$$

3.4.2 Recall. Recall is also known as the True Positive Rate or sensitivity. It is the number of positive predictions to the total positive class values present in the data.

$$Recall = \frac{True\ Positives}{(True\ Positives)+(False\ Negatives)} \tag{2}$$

3.4.3 F1-score. The F1-score is the weighted harmonic mean of precision and recall.

$$F1\ score\ = 2 * \frac{(Precision*Recall)}{(Precision)+(Recall)} \tag{3}$$

4.Results and discussion

In this work, data sampling techniques such as under-sampling, SMOTE and SMOTE-Tomek are used to handle the unbalanced data. The dataset is divided into training and test set (75% and 25% respectively). The pre-processed data is then trained on classification models, such as KNN, logistic regression, random forest and SVM.

4.1 Data sampling results

The number of data samples obtained in each class after applying data sampling techniques is shown in table 1. Under-sampling technique reduces the number of samples considerably to just 492 samples, which are less for training machine learning algorithms efficiently. Both SMOTE and SMOTE-Tomek methods result in 227845 samples in both the classes.

Table 1. Balanced data samples used for training the classifiers.

Sampling technique	Class 0	Class 1
Under-sampling	492	492
SMOTE	227,845	227,845
SMOTE-Tomek	227,845	227,845

4.2 Classification results

The precision, recall and F1-score values of the trained classifier on the test data predictions is shown in table 2. It is observed that out of the data sampling techniques, under-sampled data shows the best results for classification predictions. However, the number of under-sampled data is very less and cannot be generalized with a high confidence level. Among SMOTE and SMOTE-Tomek, the best results are obtained for SMOTE-Tomek method with F1-score of 0.93.

A comparison of the classifiers show that random forest works the best in all the data sampling techniques, with the best result being F1-score of 0.94 obtained for the under-sampled data. Logistic regression classifier shows the worst performance for SMOTE and SMOTE-Tomek sampling methods. SVM classifier also shows poor performance for SMOTE and SMOTE-Tomek methods. It can be observed that both logistic regression and SVM classifiers show good sensitivity but very low precision.

Table 2. Performance of classifiers trained using balanced data.

Data sampling	Classifier	Precision	Recall	F1-score
Under-sampling	KNN	0.91	0.90	0.90
	Logistic Regression	0.92	0.91	0.91
	Random forest	**0.94**	**0.94**	**0.94**
	SVM	0.94	0.93	0.93
SMOTE	KNN	0.82	0.86	0.84
	Logistic regression	0.52	0.92	0.53
	Random forest	**0.91**	**0.92**	**0.92**
	SVM	0.66	0.92	0.73
SMOTE-Tomek	KNN	0.82	0.91	0.86
	Logistic regression	0.53	0.93	0.55
	Random forest	**0.92**	**0.94**	**0.93**
	SVM	0.67	0.90	0.74

4.3 Comparison with related work

The classification results obtained for the random forest classifier in our proposed approach are compared with the related work and the best results of the Kaggle challenge [16]. As observed in table 3, our proposed classifiers show the best recall values and F1-score values, when compared to the related work. Although some related work show accuracy of 1.00, accuracy is not the right performance metric to measure the performance of the model. Hence, our proposed approach shows promising results with recall and F1-scores.

Table 3. Comparison of classification performance with related work.

Related work	Classifier	Sampling technique	Precision	Recall	F1-score	Accuracy
[6]	Random forest	None	NM	NM	NM	0.90
[7]	ANN	None	0.99	NM	NM	0.99
[8]	Random forest	None	0.97	NM	0.93	1.00
Merryyundi [16]	Random forest	SMOTE	0.96	0.89	0.92	0.99
Hassan Amin [16]	Random forest	None	0.99	0.88	0.93	1.00
Our proposed methods	**Random forest**	**Under-sampling**	**0.94**	**0.94**	**0.94**	**0.94**
	Random forest	**SMOTE**	**0.91**	**0.92**	**0.92**	**0.99**
	Random forest	**SMOTE-Tomek**	**0.92**	**0.94**	**0.93**	**0.99**

*NM- Not mentioned

4.4 Limitations and future work

In this work, we have used only three data sampling techniques. The proposed approach could be developed further with a combination of different data sampling techniques and cross-validation techniques to improve the classification performance.

5.Conclusion

This paper proposes different machine learning methods to detect credit card fraudulent transactions. This work explores under-sampling and over-sampling techniques to solve the imbalance of classes in the dataset. A number of classifiers were trained with the sampled data and evaluated for precision, recall and F1-score. Random forest classifier showed the best results for classification with all the data sampling techniques. Hence, the results obtained are promising and the proposed approach can be used for fraud detection.

References

[1] Fraud The Facts 2019- The definitive overview of payment industry fraud *UK Finance Report*

[2] Number of credit and debit card fraud incidents reported across India in 2019, by leading state- https://www.statista.com/statistics/1097927/india-number-of-credit-debit-card-fraud-incidents-by-leading-state/

[3] Dal Pozzolo A, Caelen O, Le Borgne YA, Waterschoot S and Bontempi G 2014 Learned lessons in credit card fraud detection from a practitioner perspective *Expert systems with applications* **41** pp 4915-28

[4] Najadat H, Altiti O, Aqouleh AA and Younes M 2020 Credit card fraud detection based on machine and deep learning *11th Int. Conf. Information and Communication Systems* (IEEE) pp 204-08

[5] Drummond C and Holte RC 2003 C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling *Workshop on learning from imbalanced datasets II* (Washington DC: Citeseer) **11** pp 1-8

[6] Kumar MS, Soundarya V, Kavitha S, Keerthika ES and Aswini E 2019 Credit card fraud detection using random forest algorithm *3rd Int. Conf. Computing and Communications Technologies* (IEEE) pp 149-53

[7] Sadineni PK 2020 Detection of fraudulent transactions in credit card using machine learning algorithms *4th Int. Conf. IoT in Social, Mobile, Analytics and Cloud* (I-SMAC, IEEE) pp 659-60

[8] Sailusha R, Gnaneswar V, Ramesh R and Rao GR 2020 Credit card fraud detection using machine learning *4th Int. Conf. Intelligent Computing and Control Systems* (IEEE) pp 1264-70

[9] Dal Pozzolo A, Caelen O, Johnson RA and Bontempi G 2015 Calibrating probability with undersampling for unbalanced classification *IEEE Symposium Series on Computational Intelligence* pp 159-66

[10] Chawla NV, Bowyer KW, Hall LO and Kegelmeyer WP 2002 SMOTE: synthetic minority over-sampling technique J Artificial Intelligence Research 16 pp 321-57

[11] SMOTE for imbalanced data- https://iq.opengenus.org/smote-for-imbalanced-dataset/

[12] Batista GE, Prati RC and Monard MC 2004 A study of the behavior of several methods for balancing machine learning training data ACM SIGKDD Explorations Newsletter 6 pp 20-9

[13] Dealing With Imbalanced Datasets- https://www.datasciencecentral.com/profiles/blogs/dealing-with-imbalanced-datasets

[14] Comparative Study on Classic Machine learning Algorithms- https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222

[15] Toledo-Pérez DC, Rodríguez-Reséndiz J, Gómez-Loenzo RA and Jauregui-Correa JC 2019 Support vector machine-based EMG signal classification techniques: A review Applied Sciences 9 p 4402

[16] Kaggle Credit card fraud detection- https://www.kaggle.com/mlg-ulb/creditcardfraud