International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V69I11P228

Volume 69 Issue 11, 216-226, November, 2021
©2021 Seventh Sense Research Group®

Review Article

Composition of Feature Selection Methods And Oversampling Techniques For Banking Fraud Detection With Artifical Intelligence

Bouzgarne Itri[1], Youssfi Mohamed[2], Bouattane Omar[3] , Qbadou Mohamed[4]

[1,2,3,4]*Lab. Computer Science, Artificial Intelligence & Cyber Security (2IACS), Enset Mohammedia, Hassan II University Of Casablanca, Morocco.*

[1]itri.bouzgarne-etu@etu.univh2c.ma, [2]med@youssfi.net , [3]o.bouattane@gmail.com, [4]qbmedn7@gmail.com

Abstract *- The digital age is accompanied by a proliferation of crimes and attacks against institutions handling banking data, such as card fraud and electronic payments. The traditional protection systems used by banks based on rules and signatures are proving increasingly insufficient and ineffective in the face of constantly evolving attack techniques. Artificial intelligence and machine learning becoming dominant problem-solving techniques to fill these gaps. Thus, this article proposes a new approach for optimising the performance of prediction models for fraud detection in the case of credit cards. Although fraud prediction algorithms have been developed to deal with the problem, they still encounter some very common difficulties due to the imbalance data set. Hence, this study proposes a new composition-based algorithm, an approach that combines oversampling and feature selection methods to find the best combination of several supervised classification algorithms. This work aim to maximise the performance of the fraudulent transaction detection model in the presence of an imbalanced Dataset, while illustrating the impact of oversampling methods on the relevance of features. This research obtains the best performance in comparison to the pervious results on the same scope.*

Keywords — *Machine Learning, Oversampling, Feature Selection, imbalanced dataset, credit card, Fraud.*

I. INTRODUCTION

A. Challenge facing credit card fraud detection

Since 2018, according to European Central Bank (ECB) [1], the total volume of electronic payment transactions in Europe has organically grown by 9.7% per year. Bank card fraud is an extremely critical issue for both banks and individuals, and is a complex and exciting topic from an artificial intelligence perspective. Obviously, we can recognize that inadequate fraud management can have extremely damaging consequences. There is the financial impact of an undetected fraudulent transaction, the impact on image and customer confidence, and operational impact (fraud processing unit, crisis management, etc.). Among the frauds we find:

- Skimming : A technique whereby the banking data stored on the magnetic strip of the card and sometimes the 4-digit secret code are duplicated by means of a camera or a hijacked numeric keypad.
- Phishing, also known as scamming, consists of sending emails pretending to be a banking, insurance or health organisation to request bank account passwords or credit card numbers.
- Fraudulent use of the bank card following robbery or loss.
- Formjacking : Fraudulent use of bank cards on the Internet, a technique which consists of infiltrating websites and install malware that intercepts and transfers credit card data.

In summary, fraud occurs when a third party uses your credit card or credit account to process a transaction without authorization. However, According to Javelin Strategy, the probability of card-not-present (CNP) fraud is now 81% higher online than at the point of sale [2]. Levels of CNP fraud are still increasing year on year in Europe, in the Single Euro Payments Area (SEPA. CNP fraud reached €1.43 billion losses from 2018.

According to The Nilson Report “Fig. 1”, $28.65 billion was lost due to payment card fraud worldwide in 2019, an increase of 19.5% from $23.97 billion in 2017[3]; Losses are projected to rise to $35.67 billion in five years and $40.63 billion in 10 years. While the United States holds the top spot with a 38.6% loss based on 2018 statistics, increased by 18.4 percent and continues to climb over time [3]. Total fraud losses on UK-issued cards amounted to £671.4 million in 2018, up 19% from 2017. E-commerce fraud still represents 50% of total card fraud losses at £310.2 million. In 2020, the total number of annual fraud cases of payment cards amounted to roughly 2.83 million. In France, the amount of bank card fraud in 2018 was approximately 439 million euros, in the same year, 57300 credit card forgeries were recorded.

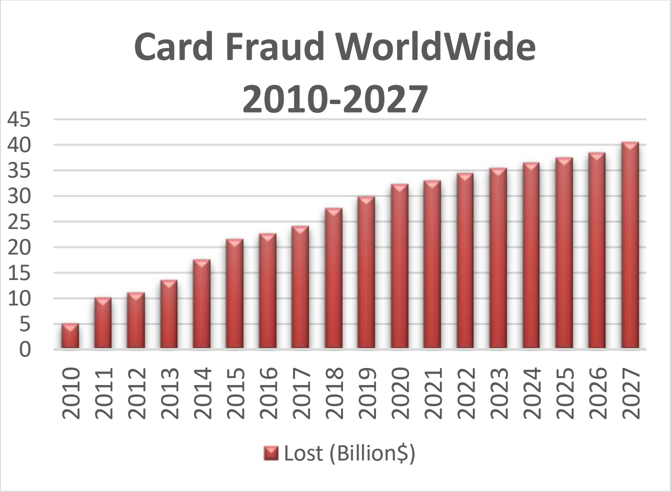


Fig. 1.Evolution of payment card fraud losses from Nilson Report

In order to secure the customer journey, 3 major axes can be envisaged:

- Protection of the customer journey: in order to secure the performance of sensitive operations via the implementation of protection solutions and customer awareness;
- Reaction to fraud: in order to alert, investigate and react rapidly in the event of fraud, following the alerts resulting from detection.
- Fraud detection: the objective of which is to detect fraud that has been or is being committed.

B. *Study approach*

a) Machine learning approach in fraud detecetion: The classical approach to fraud detection, which is widely used in the world today, consists of detecting fraud patterns that have been encountered in the past. This approach is mainly based on the application of pre-established rules (simple or advanced) on the transaction flow, which remains insufficient given the mass of transactions and the proliferation of methods that appear each time. The methods are evolving from day to day and are accentuated with the technology progress, ranging from phishing to identity theft, from "simple" password theft to data breach.

Thus, This field of study will focus on the axis of fraud detection automation through data mining techniques [4] or machine learning advances. This is a field of use of algorithms capable of learning from examples and developing a statistical model based on correlations discovered within representative samples of the dataset, which gives the computer the ability to study and progress through experience without being explicitly programmed, and offering more accurate results with less effort. Therefore, the machine-learning algorithm is developed to solve the complexity behind the data and attributes that make up a data set, a complexity that depends on and changes from one source to another. These algorithms are characterized by their low time consumption and their more accurate results. As illustrated in diagram “Fig. 2”, the machine learning process in the context of online bank card payment, often means creating customer profiles on the basis of historical data and information collected (terminals used, times and places of habitual connections, connection and transaction paths). Based on these profiles the relevant data is formed on which the algorithm will learn to build a predictive model, so that the detection of fraudulent behavior of the current transaction is deduced by comparing the actual behavior of the customer with the model. Finally, this profile is continuously updated based on new transactions made by the customer.

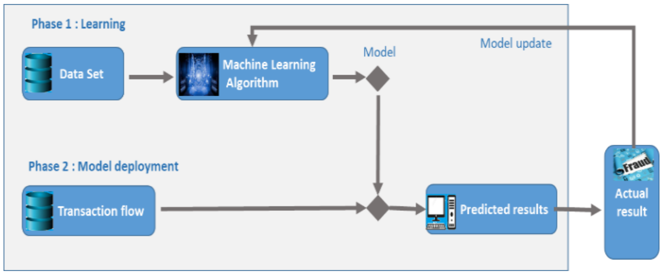


Fig. 2.How machine learning works for credit card fraud detection.

b) Challenge handled: In this paper we will address the credit card fraud detection problem for a real use case based on the dataset, containing real transactions made through credit cards in September 2013 by European cardholders. The goal is to perform the fraud prediction model through classification algorithms according to the supervised method. However, one of the main weaknesses of machine learning models in the context of fraud resides in the unbalanced dataset between the fraudulent and non-fraudulent classes, in this case, the positive class (frauds) account for 0.172% of all transactions. Indeed, machine-learning algorithm will consequently learn to overlook the minority class and to classify all the cases in the majority class. Sequentially the learning and the prediction model generated will be biased. However, there are very common statistical techniques to solve this problem of imbalance, such as the Synthetic Minority Oversampling Technique [5] and its variants.

On the other hand, there are other problems related to feature quality, in fact, in each dataset there are features more salient than others for the machine learning result. Theoretically, more discrimination can be achieved by increasing the number of features; However, feedback on the results of machine learning algorithms proves the otherwise [6]. However, if we consider the factor of massive data and the injection of new data by oversampling methods, the classification model design will not be optimal if we do not adopt the feature selection process.

Thus, the multi-faceted nature of the problem imposes the development of specific or hybrid techniques combining several algorithms and optimization processes, which we will develop in this article through a new hybridization approach of oversampling and feature selection algorithms.

c) Artcile organisation: This paper is organized as follows. In section 2, we present the background to the comprehension of the paper. We first provide a description of the dataset in study, then we present the methods conducted to solve the class imbalance problem, including the SMOTE-based technique and some of its variants. We discuss the methods that will be used for feature selection and we end this section with the performance measures of the models adequate to the fraud detection problems. In section 3, we detail the new approach and the proposed algorithm. Section 4 presents the experiment: we discuss the results of the different oversampling and feature selection methods, as well as the performance obtained with proposed algorithm. A comparison of study results with previous works on the same scope will also be presented in this section. Finally, in section 5, we conclude with discussion.

II. RELATED WORK

Fraud detection systems (FDS) have been widely studied in the literature. Several data mining techniques have been used to solve the problem of credit card fraud. The approaches pursued operate according to a supervised or unsupervised strategy [7]. Some authors have handled the same use case as this case study, using the same dataset. Of these, several authors [8][9][10][11][12][13] have used oversampling methods to deal with an imbalanced dataset, some of them adopting a technique using additional synthetic data called Synthetic Minority Over-sampling TEchnique (SMOTE), first introduced by [5]. It is a technique whereby synthetic data is added to extend the minority class and revolutionizes the old techniques of oversampling based on the reinjection of the initial data. This method was subsequently improved by several authors who proposed new variants selecting a particular area of synthetic data generation such as [14] which proposed the Borderline-SMOTE method that selects synthetic data only on the border area. [15] proposed the ADASYN method which generates synthetic data on a safe and border area based on harder level ratio and their distribution ratio. [16] proposed Majority Weighted Minority Oversampling MWMOTE, aimed to improve the sample selection scheme by generating synthetic sample based on their selection probability. [17] proposed edge detection algorithm Egde-Det, this method generates synthetic data based on the sample weight calculated by the overall magnitude of gradient. [18] proposed a Safe-Level-SMOTE method which improves SMOTE and Borderline by synthesizing the minority instances more around the larger safe level, although in this study the Borderline's performance is better than Safe-Level-SMOTE.

Regardless of the imbalanced datasets problem, these contain a large number of features that also need to be addressed; A data mining model built using all features is generally not efficient because the machine learning algorithms will be impacted by insignificant or even disturbing features in the training process. This problem can be solved by adopting three significant approaches: filtering methods, wrapping methods and embedded methods [19]. Since 2009, feature selection problem has been solved using evolutionary heuristic computing techniques [20].

III. BACKGROUND

A. Handling class imbalance with the sampling method

The main challenge of machine learning in fraud detection, especially for credit cards is the highly imbalanced distribution of two classes: normal and fraudulent transactions. This class imbalance is a problem that is relative to the degree of imbalance, they often give wrong results and they can be misleading with too optimistic scores. One of the causes of these failures is that the points of the minority class are considered as outliers that contain no information. In order to account for class imbalance, different training strategies can be used such as oversampling, undersampling, membership probability thresholding, and cost-sensitive learning [21][22]. We propose in this study the oversampling method based on the SMOTE technique [5] and its variants, a method that increases the minority class synthetically, by generating new examples of the minority class according to specific methods such as the nearest neighbor and Euclidean distance. The method provides a set of simple rules to generate new "synthesized" examples. Although each new synthetic data is built from its parents, the generated data is never an exact duplicate of one of its parents. We also used a package that implements 85 variants of the SMOTE technique [23]. All of them competed in order to find the best result.

B. Pertinent predictive features selection

Feature selection is a technique for selecting the most interesting features, variables or measures of a given system that are relevant to the achievement of the task for which it was designed. In classification of a problem, irrelevant or partially relevant characteristics can have a negative impact on the model, so their unnecessary inclusion leads to:

- Hinders interpretation by researchers/users.
- Increases the learning process time.
- Raise the probability of overfitting.

Thus the feature selection method automatically selects the most relevant features of a dataset. The training process will proceed through the highest scored feature sub-set. Here are three objectives of the Feature Selection process:

- Improving the reliability of performance prediction and avoid overfitting.

- Increasing the speed of the model's training.
- Reducing Overfitting.
- Avoid the dimensional scourge.

Feature selection algorithms are classified into three classes: the filter methods, the wrapper methods and embedded methods.

a) Filter Methods: The "filter" was the first method used for feature selection. The methods assign a score to each feature according to statistical techniques, evaluate the relationship between a predictor and the target variable. They are often univariate and consider the feature independently. Filter procedures is independent of the learning algorithm. They are generally less computing time consuming since they avoid repetitive executions of the learning algorithms on different subsets of variables. On the other hand, their major disadvantage is that they ignore the impact of the chosen sub-sets on the performance of the learning algorithm.
We used two implementations of filter methods in this study: Pearson correlation and ANOVA F-Statistic.

1) Pearson correlation*:* The full name is the Pearson Product Moment Correlation (*PPMC*), and is used to measure how strong a relationship is between two variables, notably the linear relationship between the data set and the class target. Through its correlation coefficient, the strength of this relationship between the data can be calculated. The formula returns a value between -1 and 1. When the value approaches 1, the relationship becomes strong. A zero value indicates no relationship at all, while a value close to -1 implies a strong negative correlation.

2) ANOVA: Analysis of variance method, was originally developed by Sir Ronald A. Fisher. It represents a parametric statistical hypothesis test for compare the means of different groups and demonstrates the existence of statistical differences between the means. It is a type of F-statistic, which calculates the ratio of two variances, or technically two mean squares. Variances measure the dispersal of the data points around the mean.

b) Wrapper methods: Wrapper methods define feature relevance through a prediction of the final system performance. They wrap a machine learning model by fitting and evaluating the model with different subset of predictive variables. The subset with the best performance is then selected. Here are the most commonly used techniques: Forward selection, Backward elimination, Bi-directional elimination and Recursive Feature Elimination (RFE).

We decide to address the implementation of the backward elimination method and the RFE method according to their effectiveness and reliability.

1) Backward Elimination*:* We feed all the possible features to the model at first. We check the performance of the model and then iteratively remove the worst performing features one by one until we have the final set of significant features. We used ths **OLS** model in this study, which stands for “Ordinary Least Squares”.

2) RFE: This method is used to select the most relevant features to predict the target variable in a predictive model - regression or classification. RFE applies a backward selection process to find the optimal combination of features. First, it builds a model based on all features and calculates the relevance of each feature in the model. Then, it ranks the features and removes the least relevant ones in an iterative way based on the evaluation score-model. We used two approaches for this method, one named "Opt-RFE" consists in calculating the optimal number of features according to the score for which the accuracy-model is above the average. The second one named "Fix-RFE" consists in fixing the maximum number of features according to the feedback on several training iterations.

c) Embedded methods: Takes on the "qualities" of filtering and wrapping methods. The method is iterative and includes feature selection during the learning process and carefully extracts the features and searches for the best subset that offers the best classification performance. The most common methods are Lasso and Ridge regression. Based on the results studied, we chose the Lasso regression for this work.

C. Classification Evaluation

This section presents the techniques for evaluating the trained model through the learning algorithms, including performance measures.

a) Confusion Matrix: In a training process for solving a machine learning classification problem, the Confusion Matrix is a summary of correct and incorrect prediction results compared with the actual values of the input data, divided by class as shown in "Fig. 3". Each column of the table refers to a predicted class, and each row refers to an actual class.

Prediction Actual	Fraud	No Fraud
Fraud	TP - true positive	FN - false negative
No Fraud	FP - false positive	TN - true negative

Fig. 3.Confusion Matrix

However, the most common way to derive interesting information from this kind of table is through the measures obtained by the indicators of the fusion matrix, which are calculated as derived metrics, the following are the most important of these measures:

b) Accuracy : The proportion of correct predictions (TP, TN) among all transactions that were predicted (TP, FP, TN, FN). The formula is given by

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Accuracy Paradox: Many machine-learning models are evaluated by the Accuracy measure, except that in a fraud detection problem that generally deals with unbalanced data. The accuracy is no longer adequate [26], because the training exercise and the evaluation according to the accuracy measure will generate a model that will tend to predict the no fraudulent class (majority) for all the test examples by increasing the percentage of TN. Given the other indicators TP, FP and FN will be negligible considering the result of a probability based on a huge unbalanced class percentage, which allows to reach an accuracy of 99%. Unfortunately, many studies use this measure to judge the model performance without considering the paradox problem for unbalanced data.

c) F-Measure: To evaluate a trade-off between recall and precision, we use the "F-measure". It is a harmonic mean or the weighted average of precision and recall. Created by Van Rijsberjen [24], it performs well on an imbalanced dataset. The formula is given by

$$F - Measure = \frac{\left((1 + \beta^2) \times Precision \times Recall\right)}{\left((\beta^2 \times Precision) + Recall\right)}$$

- Precision measure calculates what percentage is truly positive (TP) among all transactions predicted positive (TP, FP). It is given by

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

- Recall, also knowen as Sensitivity, is a measure that calculates the proportion of transactions actually predicted as fraudulent (TP) among those actually predicted as fraudulent or non-fraudulent (TP, FN). This is the capacity of the model to detect all frauds. It is given by

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

- β parameter give the degree of the recall's importance regarding the precision. If Recall is considered more important than Precision then β should be > 1, if twice as important then β = 2.

d) AUC-ROC: Area Under Receiver Operating Characteristic Curve. This measure represent a probability curve that plots the Recall versus (1- specificity) at different threshold values. The Area at the bottom of the curve (AUC) is a measure of the ability of a classifier to distinguish between classes. The greater the value (or area), the better the model is.

e) G-mean: called Geometric Mean, proposed by Kubat et al. [25]. It is a widely used metric for unbalanced classification problems. The metric uses two opposite measures: recall and specificity. This measure calculates the balance as:

$$G - mean = \sqrt{Recall \times Specifity}$$

Where specificity is a measure that tells us what proportion of predicted transactions are actually non-fraudulent (TN) among all fraudulent transactions (TN, FP). It is the opposite of Recall given by

$$\text{Specificity} = \frac{TN}{TN + FP}$$

The G-mean is introduced in this study to calculate the percentage and threshold of oversampling of the minority class [26].

IV. PROPOSED METHODOLOGY

A. Data Collection & Data Pre-processing

We have chosen to work on a real case for this study. The data set is represented by credit card transaction data in September 2013 provided by European cardholders. The data set comprises 284,807 transactions recorded over 2 days, with 30 predictor variables and one target variable [0,1], the values 1 means "Fraud" and 0 means "No fraud. Fraudulent transactions represent only 0.172% of all transactions, with 492 cases, revealing a huge imbalance of the two classes in Data set “Fig. 4”. As part of the data processing, we have only processed the Time feature given in seconds, and in fact, we have transformed it into minutes and hours to improve relevance.

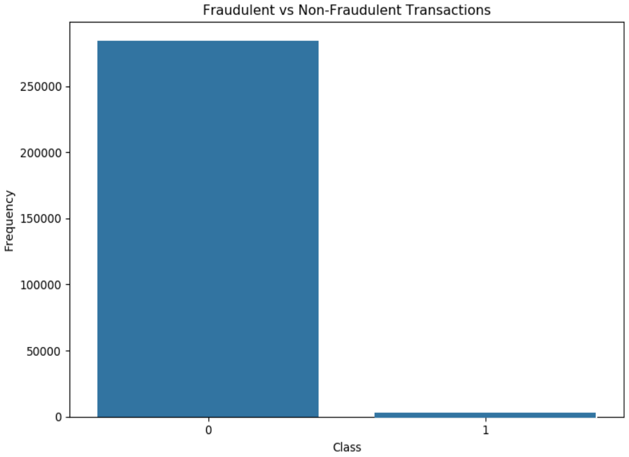


Fig. 4.Data distribution by index variable (normal=0, fraud=1)

The dataset “is in CSV format, obtained after transformation of the original attributes using the Principal Component Analysis (PCA) method. The features are kept anonymous except for three features: the index, the amount of transactions and the time when a transaction took place. As part of the data processing, we have only processed on the time feature. In fact, as it is represented in seconds, we transform it into minutes and hours to get a better relevance, so we have 31 features in total without the index.

B. Classification algorihtm

In order to choose between the learning algorithms, we have highlighted six of the best known algorithms in this study : Random Forest (RF), Multilayer Perceptron (MLP), AdaBoost (ADB), Gaussian Naive Bayes (GNB), k-Nearest Neighbors (KNN) and Decision Tree (DT). For the model training, we privileged the cross-validation technique. On the other hand, several evaluation methods are discussed in order to measure the performance of one algorithm against another in each step.

C. Hybrid Algorithm composition

To develop a credit card fraud detection system, it is not wise to train the classification algorithms on an unbalanced data set, otherwise the classification model will tend to predict any new transaction as non-fraudulent, and indeed, it will be more likely to fall into the fraudulent class due to the percentage of unbalance.

Thus, we propose to introduce oversampling methods, namely the variants of the SMOTE technique symbolized by the set "S" in the algorithm Fig. 4, whose objective is to balance the two classes by choosing the best variant offering very high model performance.

After the application of the oversampling method on the initial dataset, we focused only on the instances to generate a new data set where the two classes are balanced. However, the features must be also optimized and well-chosen to keep only the relevant ones and offer a better performance, except that the feature selection methods symbolized by the set «*F*» in "Fig. 5" depend on the instances and the distribution of the new dataset. According to this study-hypothesis, each oversampling algorithm will act on the choice of the feature selection algorithms. This hypothesis will be validated in the experimentation section. As a result, after applying the oversampling methods (SMOTE variant), we will test all the feature selection methods and choose the best «*SoF*» composition between the "S» set and the "F» set, a hybridization of the algorithms will generate a new dataset.

After the generation of the new instances by the oversampling method and having chosen the relevant features, the training of the model and the evaluation of the performances will be carried out through the classification algorithms symbolized by «*C*» by choosing the best classifier which offers a performing model on the basis of the Data Set generated by the « *SoF*» composition, constituting a new hybridization symbolized by «*W = SoFoC*». The following is the detail of the proposed algorithm:

Suppose a training dataset D with N examples:

$D(m) = \{(\mathbf{x}_n, y_n)_i, i=1..N\}$, where x_n is the n_{th} data sample containing m features, and y_n is the corresponding class label (0,1) in n_{th} sample.

We denote:

- $S = \{Si, i=1 \ldots P\}$: list of P smote variant method.
- $F = \{Fj, j=1 \ldots Q\}$: list of Q Feature selection method.
- $C[D(m)] = \{C_k, k=1 \ldots R\}$: list of R machine learning classifiers, applied to the training data set $D(m)$ with the m selected features.
- $Wi,j,k = S_i o F_j o C_k$: a composition of three algortihm to constitute an hybrid algorithm of SMOTE variant, feature selection and classifier. W_o is noted as the best composition offering the best performance.
- $Perf(D)i,j,k$: calculate the performance of the composition $W_{i,j,k}$ after cross-validation experiments. *MaxPerf* is the best performance found.

```
Input:   Data $D = \{(\mathbf{x}_n, y_n)\}$,  $S = \{S_i\}$, $F = \{F_j\}$
Output:  $W_o(D)$, $MaxPerf(D)$   : Best Composition Oversampling & Feature
                                    Selection Algorithm.
Initialization:
For $i=1$ to $P$ // $P$ is the number of oversample SMOTE variant algorithm S.
    $D_i \leftarrow S_i(D)$ // apply the SMOTE variant (i), $D_i$ is a new  balanced
                     training dataset.
    For  $j=1$ to $Q$ // $Q$ is the number of feature selection algorithm F.
        $m_j \leftarrow F_j(D_i)$ // apply feature selection algorithm j, $m_j$ is a
                        new selected feature after balancing dataSet.
        For $k=1$ to $R$ // $R$ is the number of machine learning classifiers.
            $Perf(D)_{i,j,k} \leftarrow Evaluate[Cross\_validate(C_k[D_i(m_j)])]$
            // apply 10-fold cross-validation and evaluate the model for
               each iteration.
            //MaxPerf  initialized by the first iteration $(i,j,k)$.
            If     $Perf(D)_{i,j,k} >$  $MaxPerf$
            Then
                $MaxPerf \leftarrow Perf(D)_{i,j,k}$
                $W_o \leftarrow W_{i,j,k}$
RETURN $W_o$ & $MaxPerf$
```

Fig. 5.Proposed hybridisation algorithms

V. EXPERIMENTAL ANALALYSIS

This section is organised in two parts. The first part presents the results of experiments and lists the performance measures covered by the classification problems for fraud prediction. Thereafter details the new approach combining the oversampling and feature selection methods and their impact on the performance measures. The second part presents results comparison with previous works in the literature on the same use case.

A. Experimental Results Discussion

a) Measure analysis and performance comparative:

1) Measure analysis and classifiaction model *:* In this paragraph, we proposed a comparative analysis between the different classification models before applying the oversampling method and feature selection, according to the performance measures Accuracy, Precision, Recall, Specifity, G-mean, F-Measure and *AUC-ROC*.

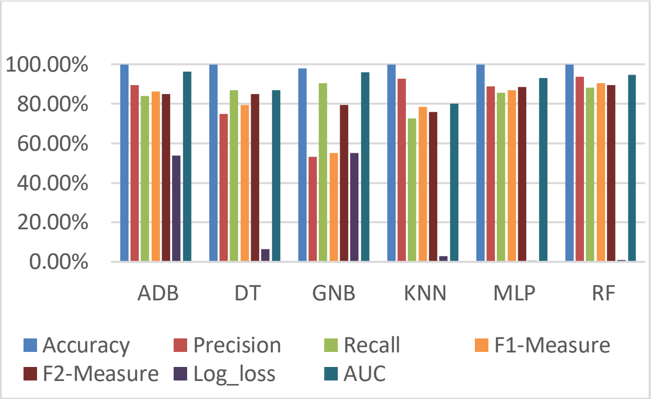


Fig. 6.Performance comparison of classification algorithms

We trained the dataset and evaluated the most known machine learning models: Ababoost (ADB), Random Forest (RF), Multilayer perceptron (MLP), k-nearest neighbors (KNN), Decision Tree (DT), Gaussian Naive Bayes (GNB). We used the Corss validation method for this evaluation. The results obtained in "Fig. 6" show interesting performances according to the measures accuracy 99% and Specifity at 99% despite the dataset imbalance problem. Although Accuracy is the most common method used to evaluate the performance of a classification model, some studies [26][27] have explained that Accuracy is not enough to measure the performance for data imbalance problems. Indeed, the model will have a tendency to predict a fraudulent case of the majority class, where a high accuracy of 99% is achievable by predicting the majority class for all examples. Consequently, the false positives will have a lower percentage, which also explains the high values of Specifity like Accuray. Thus other measures are being adopted to overcome these parameters and make performance measurement more credible, notably the F-measure, AUC-ROC, G-mean and Log_Loss.

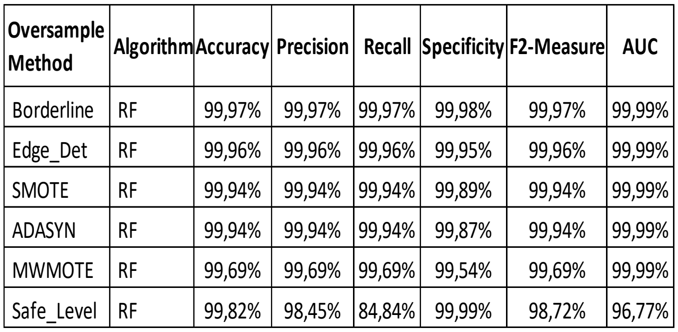
In the field of fraud processing, the Recall measure holds a major importance compared to the other basic measures like precision, specificity and accuracy. Indeed, in the process of fraud prediction and detection in the banking sector, it is important to detect all possible probabilities that could produce a fraud. The banker can tolerate the processing of a false positive case than to miss a false negative one; as it is more risky to let a fraudulent case pass and generate losses or for the bank to have a bad reputation. Conversely, if a transaction is predicted to be fraudulent when it is not in reality (false positive), this only adds to the costs of checking and analysing the case without major loss or risk. Thus, false negatives are much more significant than false positives in this case, for which reason we give more importance to recall.

The F-score is a way of combining the precision and recall but also has a parameter β which offers the possibility to weigh the Recall and Precision values according to their importance, in fact in this context we will need to give weight to Recall by assigning the value β to 2, giving importance to Recall twice as much.

As presented in "Fig. 6" the best recall score is obtained by the GNB algorithm with a score (90.561%), but the precision score (53.051%) remains the lowest compared to the other algorithms. If we look at the results according to F2-measure, RF and MLP have better performances (resp. 89.43%, 88.65%), but the Recall values are (resp. 88.10%, 85.65%). In the following paragraphs, we will present an oversampling approach to deal with the imbalance problem and feature selection methods to improve the performance of these indicators.

2) Sampling methods

To address the dataset imbalance problem, we adopted variants of the SMOTE method as an oversampling technique for generating new samples by interpolation until the two minority and majority classes are balanced. We have chosen five variants of SMOTE as the best ranked for this case study among 85 variants [23]. We have trained each of these methods with six classification algorithms. In "Fig. 7", we illustrated the best score - according to F2-Measure - of each oversampling method and the classification algorithm holding this score.



Oversample Method	Algorithm	Accuracy	Precision	Recall	Specificity	F2-Measure	AUC
Borderline	RF	99,97%	99,97%	99,97%	99,98%	99,97%	99,99%
Edge_Det	RF	99,96%	99,96%	99,96%	99,95%	99,96%	99,99%
SMOTE	RF	99,94%	99,94%	99,94%	99,89%	99,94%	99,99%
ADASYN	RF	99,94%	99,94%	99,94%	99,87%	99,94%	99,99%
MWMOTE	RF	99,69%	99,69%	99,69%	99,54%	99,69%	99,99%
Safe_Level	RF	99,82%	98,45%	84,84%	99,99%	98,72%	96,77%

Fig. 7. Smote variant method result

We can observe that Random Forest algorithm takes the first position for all the Oversampling tehnique, Borderline has the best score for all the measures, their values are clearly improved compared to the previous results close to 100% (F2-measure = 99,97, AUC-ROC = 99,99).

a) Feature selection *:*

Removing irrelevant features is pertinent for better performance. Thus, we chose to train data set through the six Classification Algorithms for six feature selection methods. "Fig. 8" illustrates the performance results with the number of features selected, given unbalanced dataset in the study contains 32 features before training. We presented the best performing algorithm for each feature selection method according to the value of the F2-measure indicator.

FS Method	Future Number	Classifier	Accuracy	Precision	Recall	Specificity	G-mean	F2-Measure	AUC
ANOVA	23	RF	99,94%	93,77%	88,30%	99,98%	93,96%	89,34%	94,64%
Embedded	16	RF	99,94%	93,82%	88,20%	99,98%	93,91%	89,27%	94,08%
PPMC	9	RF	99,94%	93,78%	88,51%	99,98%	94,07%	89,51%	93,00%
OLS	30	RF	99,94%	93,94%	88,40%	99,98%	94,01%	89,46%	94,51%
Opt-RFE	31	RF	99,94%	93,69%	88,61%	99,98%	94,12%	89,58%	94,73%
Fix-RFE	17	RF	99,94%	93,93%	88,51%	99,98%	94,07%	89,54%	93,99%

Fig. 8. Result Feature Selection Method

The results show that Random Forest (RF) still has the best score for all Feature Selection methods. The results are slightly improved from the initial results, the Opt-RFE method has the best score according to F2-measure and AUC-ROC (89,58 and 94,73).

B. Hybrid method approach

1) Algorithm composition

The injection of new samples of the minority class through the oversampling methods during the training of the model involves an impact on the relationship between each input feature and the target variable and their relevance.

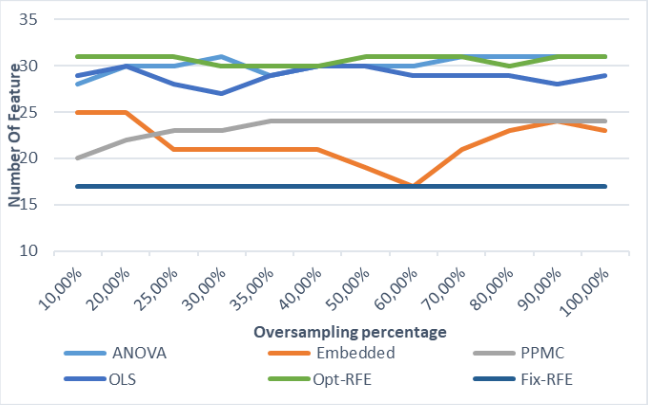


Fig. 9. Evolution of feature numbers per Future Selection method and oversampling percentage.

"Fig. 9" demonstrates that the number of relevant attributes varies through the oversampling iterations of the minority class, but also according to the feature selection method. In this graph, we used the example of Borderline as an oversampling method and Random Forest as a training classification algorithm.

Thus, this study approach will combine and compose the two functions Oversampling (First function) and Feature Selection (Second function) in order to take into consideration the evolution of the dataset distribution, the training model process through cross-validation method will then be applied according to the chosen classification algorithms as outlined in "Fig. 4". "Fig. 10" shows the result of the 10 best compositions, according to the F2-Measure and the loss_Log indicator. We introduced the Log Loss measure in support of the F2-Measure to rule in case of an equality of scores.

Oversampler	Feature selection	Feature Number	Algorithm	Precision	Recall	Specificity	F2-Measure	Log_loss	AUC
Borderline	Fix-RFE	17	RF	99,97%	99,97%	99,97%	99,967%	**-0,51%**	99,99%
Borderline	PPMC	19	RF	99,97%	99,97%	99,97%	99,966%	-0,58%	99,99%
Borderline	Opt-RFE	30	RF	99,97%	99,97%	99,98%	99,966%	-0,60%	99,99%
Borderline	PPMC	21	RF	99,97%	99,97%	99,98%	99,971%	-0,62%	99,99%
Borderline	OLS	29	RF	99,97%	99,97%	99,98%	**99,974%**	-0,64%	99,99%
Borderline	Fix-RFE	17	RF	99,97%	99,97%	99,98%	99,972%	-0,64%	99,98%
Edge_Det	Fix-RFE	17	MLP	99,90%	99,90%	99,81%	99,897%	-0,66%	99,99%
ADASYN	OLS	21	RF	99,96%	99,96%	99,93%	99,963%	-0,72%	99,99%
Edge_Det	Fix-RFE	17	RF	99,95%	99,95%	99,93%	99,947%	-0,72%	99,99%
Edge_Det	Opt-RFE	30	RF	99,96%	99,96%	99,94%	99,958%	-1,27%	99,99%

Fig. 10. Top ten result through new algorithm.

According to the results, we notice that this approach has clearly improved the performance of the model and capitalizes on the impact of oversampling method for selecting the relevant features. The best performance according to F2-Measure is held by the composition (Borderline, OLS, Random Forest) with a score (99.974%), whereas the best performance according to Log Loss is held by the composition (Borderline, Fix-RFE ,RadomForest) with a score (-0.51%).

2) Oversampling threshold

Finally, in order to improve the quality of the model and to reduce the risks of overfitting that can be generated by the new (non-real) minority class samples injected by the oversampling methods; we opted to add to the study the oversampling threshold method TH-SMOTE defined in a previous study [26]. This consists of increasing the percentage of the minority class up to the threshold where the evolution of the performance over the oversampling iterations takes an approximate stagnant curve.

Thus, in this section we will apply the TH-SMOTE method to the two best compositions found (Borderline, Fix-RFE and RadomForest) and (Borderline, OLS, Random Forest). The two graphs "Fig. 11" and "Fig. 12" present the evolution of the indicators as a function of the percentage of oversampling for the two chosen compositions. We gradually increase the percentage of the minority class then train the model and check the performance until we reach 100% where the percentage of the two classes become balanced.

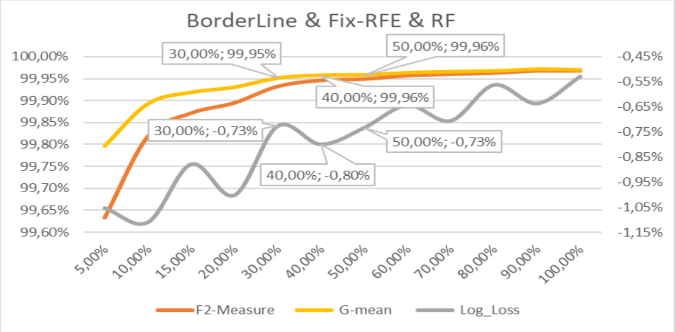


Fig. 11. Evolution of Measures by BordLine oversampling percentage iterations in composition with Fix-RFE feature selection and Random Forest classifier.

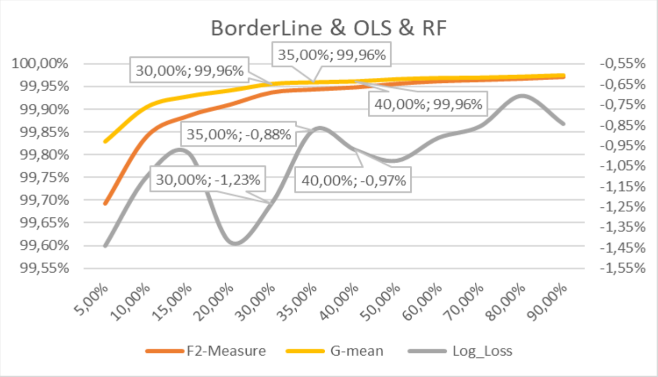


Fig. 12. Evolution of Measures by Borderline oversampling percentage iterations in composition with OLS feature selection and Random Forest classifier.

According to the previous approach [26], the optimal threshold according to the G-mean indicator is reached at 40% for the composition (Borderline, OLS, Random Forest) with 30 selected features, the values of G-mean, F2-measure and Log_Loss are respectively 99.96%, 99.95% and -0.97%. Whereas the threshold for composition (Borderline, Fix-RFE ,RadomForest) is also 40%, but with only 17 features selected, the G-mean, F2-measure and Log Loss indicators take the values 99.96%, 99.95% and -0.80% respectively. Hence, at 40% oversampling, the F2-measure and G-mean values are equal for both compositions while the Log_Loss value favours the composition (Borderline, Fix-RFE and RadomForest) with 16 features instead of 30 for the other composition.

C. Comparison with previous literature

According to the literature that addresses the same use case, this study outperformed the other results mentioned in the "Fig. 13". Indeed, these performances are obtained with a new method, using an unbalanced data set where the minority class represents only 40% of instance compared to the majority class. In addition, several works only use the accuracy indicator to evaluate the model. In others, they push the oversampling percentage to 100%, which negatively affects the credibility of the results and their representation.

Research Articles	Accuracy	Precision	Recall	Specificity	AUC-ROC
Rtayli and Enneyaa, 2019 [28]	95%	-	87%	-	91%
Sohony et all (2018) [29]	99.95%	85.85%	86.73%	-	-
Saia and Carta (2017) [30]	95%	-	91%	-	98%
Kittidachanan et all (2020) [31]	-	-	98.53%	87.50%	93.01%
Zamini et all (2019) [32]	98.90%	-	81.63%	-	96.10%
Fiore et al. 2017 [33]	99.96%	97.87%	70.23%	99.99%	-
Randhawa 2018 [34]	97.70%	-	83%	-	-
Nayak et all 2020 [35]	99.25%	-	-	-	-
El hlouli et all 2020 [36]	97.84%	99.32%	96.35%	-	-
Sailusha et all (2020) [37]	99.83%	99.96%	99.87%	15.91%	94.29%
Ummul et all (2019) [38]	97.69%	-	-	-	-
Kumar et all (2020) [10]	98.69	98.41	98.98	-	-
Itri and Youssefi, this study: 100% Oversampling.	99.97%	99.97%	99.97%	99.98%	99.98%
Itri and Youssefi, this study: 40% Oversampling.	99.96%	99.95%	99.95%	99.97%	99.97%

Fig. 13. Comparative Performance Analysis with previous work.

VI. CONCLUSION

The current study proposes a machine learning solution for solving the credit card fraud detection problem and dealing with highly imbalanced data. This paper demonstrated an approach drastically improved the performance of the model. The obtained indicators outperformed the results obtained in previous research in the same scope. Illustrating the impact of oversampling methods on the relevant feature selection and model performance. A hybrid algorithm succeeded in finding the best combination of algorithms, both triplets (Borderline, Fix-RFE, Radom Forest) and (Borderline, OLS, Random Forest) are ranked higher than other composing algorithms. The measurement indicators and their veracity are also discussed in order to evaluate the performance and quality of the model, focusing on adequate metrics for fraud detection and imbalanced data problem.

ACKNOWLEDGMENT

This project is a part of a project supported by co-financing from the CNRST (Centre National pour la Recherche Scientifique et Technique) and the Hassan II University of Casablanca, Morocco. The project is selected in the context of a call for projects entitled “Scientific and Technological Research Support Program related to the theme of COVID-19”.

REFERENCES

[1] European Central Bank, Payments Statistics: (2018) .Press release. (2019).https://www.ecb.europa.eu/press/pr/stats/paysec/html/ecb.pis2018~c758d7e773.en.html

[2] Javelin Strategy & Research. Identity Fraud Hits All-Time High With 16.7 Million U.S. Victims in (2017).https://www.javelinstrategy.com/press-release/identity-fraud-hits-all-time-high-167-million-us-victims-2017-according-new-javelin

[3] Nilsonreport.com. [online]. Source of news and analysis of the global card and mobile payment industry Available at: https://shiftprocessing.com/credit-card-fraud-statistics/

[4] Sayyed Shifanaz, S.Muzaffar, V.Kshirsagar,P.Kadlag, N.Kadam.Fraud Detection in Online Transactions using Data Mining Technique. SSRG International Journal of Computer Science Engineering (SSRG - IJCSE) - Special Issue ICIETEM (2019).

[5] Chawla et al., SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 16 (2002) 321–357.

[6] Mark A. Hall: Correlation-based Feature Selection for Machine Learning. Université de Waikato, NewZeland, (1999).

[7] Phua, C., Lee, V. C. S., Smith-Miles, K., and Gayler, R. W., A comprehensive survey of data mining-based fraud detection research. CoRR, abs/1009.6119. (2010).

[8] Saia R, Carta S. Evaluating credit card transactions in the frequency domain for a proactive fraud detection approach. ICETE 2017 - Proc. 14th Int. Jt. Conf. E-bus.Telecommun. 4(Icete) (2017) 335–42.

[9] Fiore U, De Santis A, Perla F, Zanetti P, Palmieri F. Using Generative Adversarial Networks for Improving Classification Effectiveness in Credit Card Fraud Detection. Inf. Sci. (NY). (2017).

[10] Varun Kumar K S, Vijaya Kumar V G, Vijay Shankar A, Pratibha K. (2020) Credit Card Fraud Detection using MachineLearning Algorithms.International Journal of Engineering Research & Technology. 9(7) (2020).

[11] Muhammad Syafiq Alza bin Alias et al. Improved Sampling Data Workflow Using Smtmk To Increase The Classification Accuracy Of Imbalanced Dataset. European Journal of Molecular & Clinical Medicine, 8(2) (2021) 91-99

[12] Yang W., Zhang Y., Ye K., Li L., Xu CZ., FFD: A Federated Learning-Based Method for Credit Card Fraud Detection. In: Chen K., Seshadri S., Zhang LJ. (eds) Big Data – BigData 2019. BIGDATA 2019. Lecture Notes in Computer Science, Springer, Cham. 11514 (2019)

[13] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, Credit Card Fraud Detection - Machine Learning methods, 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH), (2019) 1-5, doi: 10.1109/INFOTEH.2019.8717766.

[14] H. Han, W. Wang and B. Mao, Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning, Proc. of International Conference on Intelligent Computing, Part I, Hefei, China, (2005) 878-887.

[15] H. He, Y. Bai, E. A. Garcia and S. Li, ADASYN: Adaptive synthetic sampling approach for imbalanced learning, IEEE International Joint Conference on Neural Networks, (IEEE World Congress On Computational Intelligence), 3 (2008) 1322-1328.

[16] S. Barua, M. M. Islam, X. Yao, and K. Murase, MWMOTE – Majority weighted minority oversampling technique for imbalanced data set learning, IEEE Trans. Knowl. Data Eng., 26(2) (2014) 405-425,.

[17] Kang, Y. and Won, S., Weight decision algorithm for oversampling technique on class-imbalanced learning, ICCAS (2010) 182-186

[18] Bunkhumpornpat et al.Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem, Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, (2009) 475–482

[19] Xu, H., Yu, S., Chen, J. et al. An improved firefly algorithm for feature selection in classification.Wireless Pers. Commun. 102 (2018) 2823–2834.

[20] Yusta, S.C., Different metaheuristic strategies to solve the feature selection problem. Pattern Recognition Letters. 30(5) (2009) 525–534.

[21] Japkowicz N & Stephen S., The class imbalance problem: A systematic study. Intelligent Data Analysis 6(5) (2002) 42-449.

[22] Buda M, Maki A & Mazurowski MA., A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks 106 (2018) 249-259.

[23] Kovács, G.: An empirical comparison and evaluation of minority oversampling techniques on many imbalanced datasets. Applied Soft Computing. 83 (2019) 105662.

[24] C. J. V. Rijsbergen, Information Retrieval. Butterworth-Heinemann, Newton, MA, USA, 2nd edition. (1979)

[25] Miroslav Kubat and Stan Matwin: Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. Proceedings of the 14th International Conference on Machine Learning, (1997) 179-186.

[26] Itri BouzgarneandYoussfi Mohammed., Empirical Oversampling Threshold Strategy for Machine Learning Performance Optimisation in Insurance Fraud Detection International Journal of Advanced Computer Science and Applications(IJACSA), 11(10) (2020).

[27] P. Branco, L. Torgo, R.P. RibeiroA Survey of Predictive Modelling on Imbalanced Distributions. ACM Computing Surveys (CSUR), 49 (2) (2016) 1-50

[28] Rtayli N, Enneya N. Selection Features and Support Vector Machine for Credit Card Risk Identification. Procedia Manuf (2020) 46:941–8

[29] Sohony I, Pratap R, Nambiar U. Ensemble learning for credit card fraud detection. In: ACM International Conference Proceeding Series, (2018).

[30] Saia R, Carta S. Evaluating credit card transactions in the frequency domain for a proactive fraud detection approach. ICETE 2017 - Proc. 14th Int. Jt. Conf. E-bus.Telecommun. 4(Icete) (2017) 335–42

[31] Kittidachanan K. Anomaly Detection based on GS-OCSVM Classification. In: 202012th Int. Conf. Knowl. Smart Technol, (2020) 64–9

[32] Zamini M, Montazer G. Credit Card Fraud Detection using autoencoder based clustering. In: 9th International Symposium on Telecommunication: With Emphasis on Information and Communication Technology, IST 2018, (2019)

[33] Fiore U, De Santis A, Perla F, Zanetti P, Palmieri F. Using Generative Adversarial Networks for Improving Classification Effectiveness in Credit Card Fraud Detection. Inf. Sci. (NY). (2017).

[34] Randhawa, Kuldeep, Chu Kiong Loo, Manjeevan Seera, CHEE PENG Lim, and Asoke K. Nandi, Credit card fraud detection using AdaBoost and majority voting IEEE ACCESS, (2018) 14277-14284.

[35] Nayak H.D., Deekshita, Anvitha L., Shetty A., D'Souza D.J., Abraham M.P., Fraud Detection in Online Transactions Using Machine Learning Approaches—A Review. Advances in Intelligent Systems and Computing, Springer, Singapore 1133 (2021) 589-599.

[36] Fatima Zohra El hlouli, Jamal Riffi, Mohamed Adnane Mahraz, Ali El Yahyaouy, Hamid Tairi., Credit Card Fraud Detection Based on Multilayer Perceptron and Extreme Learning Machine Architectures.2020 International Conference on Intelligent Systems and Computer Vision (ISCV) (2020).

[37] Ruttala Sailusha, V. Gnaneswar, R. Ramesh, G. Ramakoteswara Rao. 4th International Conference on Intelligent Computing and Control Systems (ICICCS) (2020).

[38] M. Ummul Safa , R. M. Ganga., Credit Card Fraud Detection Using Machine Learning. International Journal of Research in Engineering, Science and Management 2(11) (2019) 2581-5792 .ISSN (Online)