Assessing the feasibility of machine learning-based modelling and prediction of credit fraud outcomes using hyperparameter tuning

Zhixin Tang

School of Economics, Guangdong University of Technology, Guangzhou, Guangdong, China
1225707834@qq.com

Abstract: Both the actual theft of a credit card and the deletion of private credit card data are considered forms of credit card fraud. For detection, there are numerous machine learning algorithms accessible. So, several algorithms that can be used to categorize transactions as fraudulent or lawful are illustrated in this study. In this experiment, the credit card fraud prediction dataset was utilized. The dataset is extremely skewed, hence undersampling is used rather than oversampling. The dataset is separated into test and training data portions, and feature selection is made. The experiment uses the methods of Logistic Regression, Random Forest, SVM, ADABoost, XGBoost, and LightGBM. Moreover, the SMOTE and Optuna's hyperparameter tweaking ways provide model customization. The findings suggest that specific algorithms may be capable of accurately recognizing credit card fraud.

1. Introduction

The intersection of finance and technology has received more attention from the financial sector since the start of the 21st century, and addressing problems in the financial pitch by combining the two has been the favored method. The global spread of COVID-19 has encouraged the transformation of several traditional businesses that have been severely impacted into industries of the digital economy. This has also contributed to the rise of e-commerce and credit card-based online services. Nonetheless, it cannot be denied that credit card fraud has grown to be one of the most challenging problems to handle during this time. This type of illegal activity occurs when credit card authentication information is taken or withdrawn from accounts without the owner's authorization.[1] According to numerous studies, unauthorized transactions and credit card fraud account for 10-15% of all fraud instances, but 75-80% of their financial value, according to numerous studies. [2]

Banks and financial institutions are under enormous strain from rising personal and company defaults at a time when fraud research is predominantly focused on the credit insurance sector; it is therefore far more crucial to discover how to prevent fraud than how to cure it after it has occurred. S.Vaithyasubramanian[3] has put forward a number of remedies to the credit card fraud issue, including the application of   Primary PIN and Multifactor Authentication. These methods' limitations, nevertheless, stem from the difficulty of time management and database upkeep.

Nonetheless, there are a number of reasons why it is so challenging to identify credit theft. One of the biggest challenges to using Machine Learning(ML) techniques to detect credit card fraud is the inability to reproduce the majority of published work, which implies that the datasets used for detection have unknown attributes. [4] Not only is the extent of credit fraud so vast, but also its manifestations are so diverse as to render it unpredictable. Existing prediction approaches are insufficiently precise, and it is difficult to match the constantly shifting form and pattern of credit card fraud. This issue has been the topic of substantial research and analysis employing a range of approaches, yet it persists, necessitating the quest for a more effective preventive solution as opposed to a corrective one.

It is consequently advantageous to address the current anti-fraud prediction issue to build a risk control model that accurately anticipates fraud, along with machine learning and modern financial theory to confirm the model's usefulness. The goal of ML, a subset of computer science and artificial intelligence, is to accurately imitate human learning through the use of data and algorithms. [5] This suggests that it makes it possible for computers to gain knowledge from the past in order to make more accurate forecasts. [6] By learning from previous datasets, ML may be utilized to adjust to the unpredictable and covert nature of credit card fraud. In the meantime, the accuracy of prediction may be greatly improved to reduce the risk of fraud to relevant institutions by routinely comparing the prediction results of various classifiers and using a more efficient feature selection approach.

This paper examines the usage of the supervised ML algorithms in credit card fraud detection, including Logistic Regression (LR), Random Forest (RF), adaptive boosting (AdaBoost), Support Vector Machine (SVM), XGBoost, and lightGBM (LGBM). The primary objective is to determine which machine learning models are the most effective at detecting credit card fraud by comparing their performance on the dataset. The dataset for this article is compiled from information of genuine cardholders throughout Europe. Since the dataset is highly unbalanced, feature selection will be performed before to the experiment in order to facilitate subsequent performance and scoring by cleaning the data. In addition, the research's scope will include data cleansing, feature selection, hyperparameter tuning, autoencoder model building, and model evaluation.

These are the remaining sections of the essay: Several types of classifiers will be used in Section II. Previous efforts in the same field are presented in Section III. The methodology of the investigation is then presented in Section IV, which includes both the overall experimental design and the dataset source processing. The procedure for estimating the models utilized in this paper is outlined in Section V. Part VI will report the trials, while Section VII will provide the conclusion.

2. Classifiers

2.1 Logistic Regression—LR

For supervised machine learning, Logistic Regression (LR) is a highly used and prevalent model. It constructs the dataset's solution using dependent (features) and independent (target) variables.

2.2 Random Forest—RF

Frequently, a Decision Tree (DT) algorithm generates the random forest ML technique. In addition, it is utilized generally to address a variety of classification and regression issues, accurately forecasting the output of large datasets. Several classifiers are integrated into RF technology to provide various solutions for a wide range of difficult scenarios. The RF plays a key role in estimating mean values from other data. As the number increases, the precision of the results will increase. Moreover, the RF approach facilitates the elimination of the Decision Tree algorithm's [7] limitations. Moreover, it minimizes the upscaling of datasets to increase precision. There are numerous Decision

Trees in a forest, with each person serving as a slow learner. In contrast, their union results in an intelligent learner. Random Forest technology provides the advantages of great processing speed and efficiency while coping with huge, unbalanced datasets.

2.3 Adaptive boosting--AdaBoost

AdaBoost is a technique for iterative ensemble learning designed to improve binary classifier performance by identifying weak points and strengthening them [8]. Sequential learning is used by AdaBoost to gradually create new models, and successors learn from mistakes that take advantage of model dependencies by giving mislabeled instances a lot of weight [9].

2.4 Support Vector Machine--SVM

Support vector machine (SVM) is acknowledged as a method for the analysis of regression and classification in numerous circumstances. Using this strategy, researchers routinely assess the credit card usage patterns of clients. SVM algorithms are utilized to categorize consumer behaviors as either fraudulent or legitimate transactions. The SVM approach is useful when fewer features from the dataset are utilized, and as a result, accurate results can be obtained. Yet, complications arise while utilizing huge datasets (at least over 100,000). [10]

2.5 XGBoost

Using machine learning, XGBoost (XGB) is a scalable approach to optimal tree boosting. This technique can be downloaded free of charge as part of a publicly accessible source package. Its significance in numerous machine learning and data mining issues [11] is well-known. OpenMP parallel processing is just one of the many functionalities included in XGBoost. In most cases, it offers a speed boost of more than 10 times compared to Gradient Boosting. It helps with individualized goal-setting and evaluation processes. The results are superior across multiple data sets.

2.6 LightGBM

This algorithm is a more sophisticated version of gradient boosting. With a tree-based training technique, LGBM is employed to raise the gradient. This approach differs from others due to the tree's depth growth or leaf growth. And should also take note of the name of this algorithm. The term "light" refers to a quick rate of execution. LightGBM manages massive volumes of data while using the least amount of memory. The method's emphasis on forecast accuracy is another advantage. [12]

3. Related work

Varmedja et al. [13], who choose to utilize the dataset on credit card fraud from Kaggle [14], suggested an ML-based method for identifying credit card fraud. This dataset contains two days' worth of European credit card transactions. To solve the issue of class imbalance in the database, the researcher used the Synthetic Minority Oversampling Technique (SMOTE). The recommended method's effectiveness was assessed using NB, RF, and Multilayer Perceptron (MLP) ML approaches. The findings of the experiments revealed that the RF technique accurately detected fraud at a rate of 99.96%. The MLP and NB ratings for accuracy were 99.93% and 99.23%, correspondingly. The authors acknowledge that additional investigations, including diverse forms of stacked classifiers and exhaustive feature selection, is required to attain improved results.

To detect credit card fraud, Salekshahrezaee et al. [15] utilized four integrated learning classifiers based on Decision Tree (DT) classifiers in conjunction with distinct feature selection approaches. Also, their dataset was produced by the Kaggle community. Convolutional Autoencoders and Principal Component Analysis (PCA) were utilized to pick features from imbalanced data sets (CAEs). In addition, SMOTE, SMOTE Tomek, and Random Undersampling (RUS) were used for data processing. It is based on data estimating the results of the five aforementioned comparisons. The results indicate that sampling using RUS data produces the best outcomes. The authors stress that future research will yield additional classifiers, feature selection methods, and datasets that mix audio and visuals.

To improve fraud detection's precision, Fana et al. [16] devised a two-stage method that combines deep AE models such as dimensionality reduction approaches with three classifiers based on deep learning, including RNN, CNN, and CNN_RNN. To select the appropriate model hyperparameters, Bayesian optimization procedures are applied. The results of the trials indicate that the suggested solutions improve the efficiency of deep learning-based classifiers. In the meanwhile, the authors use PCA to assess the deep autoencoder model's capacity for dimension reduction. Tests have proven that AE-created models outperform PCA-created models. Future data preprocessing will incorporate supervised autoencoders to improve processing efficiency and prediction accuracy.

Khatri et al. [17] evaluated DT, KNN, LR, RF, and Naive Bayes (NB) ML approaches and then investigated the effectiveness of those algorithms for detecting credit card fraud. To examine the efficacy of each ML approach, the authors utilized an extremely unbalanced dataset of European cardholders. During the studies, the precision achieved by each classifier served as one of the key performance criteria. According to the experimental results, the precisions of LR, DT, RF, and KNN were, correspondingly, 87.5%, 85.11 %, 89.77 %, and 91.11%.

Alamri et al. [18] propose employing effective sampling approaches during the data preprocessing phase to balance the data and train a classifier model to recognize fraudulent transactions. Before using confusion metrics to measure the performance of the system and guarantee reliable outcomes, experiments examine sampling methodologies and various approaches. When used with the asymmetrical dataset, the SMOTE approach produces the best classifier results, according to experiments. Future researchers will investigate the most efficient mixed sampling procedures for tackling identification-related data imbalances.

Baker et al. [19] offer an ensemble learning technique that blends voting with machine learning classifiers. Several classifiers, including LR, NB, Bagging, DT, RF, AdaBoost, and SVM, use SMOTE and the large percentage of voting ensemble learning techniques to address difficulties. Increasing the performance of the model will require further incorporation of datasets including more fraudulent transactions.

4. Methodology

4.1 Dataset

The study makes use of a European credit card dataset provided from Kaggle, a popular platform for discovering data sets on a variety of topics. In September 2013, two-day transaction data from Western Europe's credit card dataset was gathered. The data collection contains 284,808 transactions, however, it is unbalanced because only 492 of them are marked as fraudulent, or even just 0.172% of total transactions. The majority of this dataset's characteristics were modified using Principal Component Analysis (PCA) to conceal sensitive information. Nonetheless, some characteristics are not hidden.

Initially, the "time" function displays the quantity of seconds that have transpired since the original transaction was executed. Secondly, the "amount" feature specifies the transaction's total value. The

"class" designation denotes the authenticity of the transactions. The transaction is considered to be valid if the value is 0, and fraudulent if the value is 1. [20]

4.2 The detective framework

The detective framework is illustrated by Figure1. The data must be cleaned and preprocessed before the model can be built. The first stage is to compare the correlation between the data, identify the parameters that have a high correlation with the "class" category, and establish thresholds for these parameters to assist in the future elimination of outliers. In the data cleaning process, null values and outliers are eliminated from the data. A training set and a test set are created from the dataset before the data is formally forecasted using different ML approaches to develop models. The data set was predicted in three ways: models with scale (No Smote) in Original Data, models with scale in df-filter, and model building using SMOTE and scale. These three techniques of processing the data were applied to different machine learning methods and confusion matrices were produced to compare their scores. The first method was found to overfit the data, but the third method produced superior results. LightGBM was then used to build the model for prediction, comparing the scores before and after hyperparameter tweaking using Optuna, and ultimately using feature importance to identify which characteristics contributed most to the model.
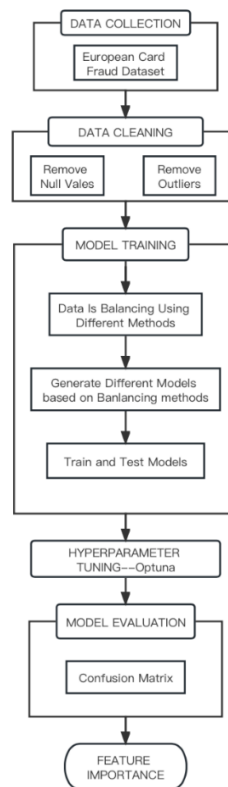

Figure 1 Research Framework for Detecting Credit Card Fraud

4.3 Data Analysis

In this section, we will analyze and verticalize each piece of data so that it is in close proximity to our business target, and then we will use deployed models to discover the best component that has the largest impact on our business objective. Using Box Plot and Hist Plot, it is straightforward to conclude that the majority of columns are not normally distributed, but columns v4, v9, v20, v18,

v24, and v26 are normally distributed. In addition, there is an imbalance between the volume of fraudulent and non-fraudulent data, a problem that must be addressed later by sampling approaches.

4.4 Data Pre-processing

Data preprocessing was done to ensure that the data was correct, consistent, and completely relevant for training autoencoders. To avoid overfitting during the model training phase, outliers were removed from the valid data. In order to verify that all of the data had the same scale, the data were then scaled, which is required because the model may favor characteristics with greater values. After employing undersampling to balance data, there are a total of 984 datasets, of which 50% are fraudulent. Before developing the model, the dataset will be split in two, with 70% of the information used for training and 30% used for testing.

4.5 Hyperparameter Tuning

Training ML models requires tuning hyperparameters. With several parameters to modify, a lengthy training period, and k-folds to prevent data leakage, hyperparameter adjustment is a laborious process. There are several ways to approach the problem, including random search, grid search, and Bayesian techniques.

Optuna is a framework for hyperparameter optimization of the next generation [19] and is a modified version of such a preceding system. [21] It offers an easy-to-setup, flexible design, a define-by-run API that enables users to dynamically expand the parameter search space, as well as effective implementation of both pruning and searching algorithms.

5. Model Evaluation

5.1 Confusion matrix

We duplicate data in unbalanced datasets to reduce the likelihood of prediction bias. Owing to this duplication process, we employ synthetic data for modeling to guarantee that forecasts are not skewed towards the majority target class value. Hence, judging models based on their precision is deceptive. Instead, we will evaluate the model using a confusion matrix that includes recall, precision, and accuracy ratings.

Typically, the confusion matrix is used to illustrate that a machine learning model's prediction does not correspond to the dataset's underlying truth. The confusion matrix has the following items: [5] True positive, False positive (FP), and False negative (FN). In light of this, we will estimate the model using the Recall, Precision, Accuracy, and F1-score of the confusion matrix score.

5.2 Area under Curve --AUC

The area under the ROC curve is the AUC. Given a randomly selected positive sample and a randomly selected negative sample, a classifier classifies and predicts the probability that the positive sample's score will be greater than the negative sample's score. And by extension, the classification of the model is better the bigger the AUC (the closer it is to 1) and the closer it is to 1.

6. Result and Comparison

Several criteria for comparing models have been used to the challenge of identifying transactions that are fraudulent in order to decide which algorithm is best suited for the task. Accuracy, recall, and precision is the most prominent metrics used to estimate the effectiveness of ML systems. A

Confusion Matrix can be used to calculate each of the aforementioned indications. These measures were utilized to estimate a model's performance. The outcomes of testing models with both original and oversampled data demonstrate that sampling is of critical importance. 30% of the overall dataset is comprised of the test set.

Prior to the hyperparameter tweaking, the data from both a testing and a training set were modeled and scored using six distinct ML techniques. Three distinct procedures were used in this study to balance the data. The first was to directly model the actual data, however, the results were discovered to be overfitting. The reason for overfitting is that the algorithm is over-learning or that some of the assumptions (e.g., sample independent identical distribution) may not be accurate, which can also result in incomplete prediction accuracy. In order to address this issue, the research employs a data filter to filter the data, upon which the model is reconstructed, and the matrix score is derived. The data was then balanced using SMOTE on the basis of the data filter, and the resulting model scores were compared to those achieved by the two preceding approaches.

The multiple matrix scores for the test set and the training set for each of the three aforementioned methods of model construction are listed below in table 1-3.

Table 1 Comparison of models with scale (No Smote) in Ordinal Data

Models	Accuracy Score Test	Accuracy Score Train	Recall Score Test	Recall Score Train	F1 Score Test	F1 Score Train
LR	0.999166	0.999194	0.552239	0.622419	0.675799	0.725086
RF	0.999507	1.000000	0.723881	1.000000	0.822034	1.000000
SVM	0.999366	0.999698	0.634328	0.837758	0.758929	0.904459
ADA	0.999248	0.999230	0.634328	0.684366	0.726496	0.752026
XGBOOST	0.999554	1.000000	0.746269	1.000000	0.840447	1.000000
LGBM	0.993632	0.996022	0.641791	0.716814	0.240896	0.380878

Table 2 Comparison of the models with scaling in Df- filter

Models	Accuracy Score Test	Accuracy Score Train	Recall Score Test	Recall Score Train	F1 Score Test	F1 Score Train	Precision Score Test	Precision Score Train
LR	0.995356	0.995414	0.787879	0.780059	0.818898	0.833856	0.852459	0.895623
RF	0.998890	1.000000	0.94697	1.000000	0.957854	1.000000	0.968992	1.000000
SVM	0.997476	0.998572	0.833333	0.920821	0.897959	0.950076	0.973451	0.98125
ADA	0.999192	0.999481	0.962121	0.982405	0.969466	0.982405	0.957854	0.982405
XGB	0.998789	1.000000	0.931818	1.000000	0.953488	1.000000	0.97619	1.000000
LightGBM	0.997779	1.000000	0.886364	1.000000	0.914062	1.000000	0.943548	1.000000

Table 3 Comparison of the models with SMOTE and scale

Models	Accuracy Score Test	Accuracy Score Train	Recall Score Test	Recall Score Train	F1 Score Test	F1 Score Train	Precision Score Test	Precision Score Train
LR	0.985151	0.983319	0.977983	0.974539	0.985044	0.983171	0.992208	0.991957
RF	0.995648	1.000000	0.993856	1.000000	0.995640	1.000000	0.997431	1.000000
SVM	0.988991	0.991989	0.982079	0.987050	0.988915	0.991949	0.995846	0.996896
ADA	0.993856	0.997586	0.99232	0.997805	0.993846	0.997586	0.995378	0.997367
XGB	0.996672	1.000000	0.995904	1.000000	0.996669	1.000000	0.997436	1.000000
LightGBM	0.996416	1.000000	0.995904	1.000000	0.996414	1.000000	0.996925	1.000000

Analyzing the data in the three tables above reveals that, of these ways, the majority of models constructed by filtering the data and balancing it with SMOTE have higher matrix scores, are more efficient, and do not suffer from over-fitting, which reduces prediction accuracy. In order to optimize the model, this experiment will compare the model's scores before and after hyper parameterization using Optuna to see whether the hyper parameterization has an optimizing effect on the mode.
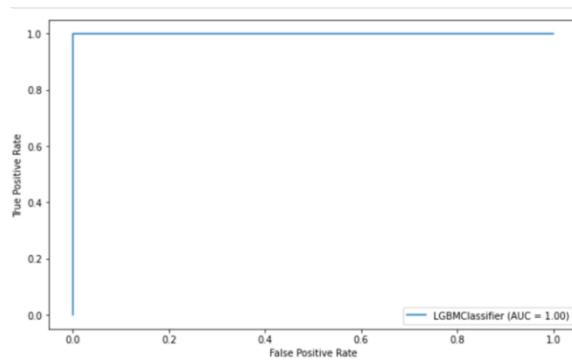
Figure 2. The AUC curve plotted after using LGBM prediction without hyperparameter tuning.

Figure 2 depicts the AUC curve for the training set predicted by the model constructed with LGBM prior to hyperparameter adjustment. The area under this AUC, or ROC curve, is 100%, indicating that the model constructed with this dataset is overfitted and inapplicable. Figure 3 depicts the recall accuracy curve for the dataset after hyperparameter adjustment. It is not straightforward to determine that the dataset performs better after hyperparameter tweaking, given that it produces high recall and precision values. And the AUC is close to 1, indicating that the model is quite effective at classification.
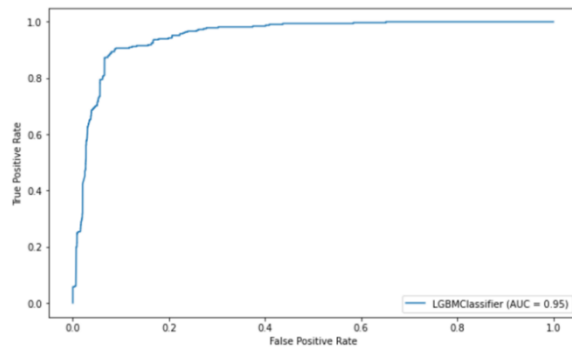

Figure 3. The AUC curve plotted with hyperparameter tuning.

7. Conclusions

This paper examines the sampling techniques that can be used to manage unbalanced data in the whole dataset which is credit card transactions. It highlights the impact of inconsistent data on the classification algorithm and the significance of sampling methodologies.

The primary objective of this study is to make comparisons of several ML techniques for the identification of fraudulent transactions. Hence, a comparison was conducted, and It was determined that in this way most of the methods generated well performance, i.e., the well categorization for determining if a transaction was fraudulent or not. This was established using a wide range of variables, including precision, recall, accuracy and F1-score. A strong recall value is essential for this sort of scenario. The assignment of characteristics and the balancing of the dataset were crucial for producing significant results. In order to provide better results, future research should concentrate on various machine learning algorithms and methods of data processing, such as genetic algorithms and various forms of stacking classifiers.

However, as a consequence of technical breakthroughs and the passage of time, it is possible that these approaches may no longer be appropriate to the current forecasts, necessitating the use of more recent datasets to review the results in the future.

References

[1] M. Valavan and S. Rita. (2023). *Predictive-Analysis-based Machine Learning Model for Fraud Detection with Boosting Classifiers.* COMPUTER SYSTEMS SCIENCE AND ENGINEERING, 45(1), pp. 231-245.

[2] Saia, R., & Carta, S. (2019). *Evaluating the benefits of using proactive transformed-domain-based techniques in fraud detection tasks.* Future Generation Computer Systems, 93, 18-32.

[3] Vaithyasubramanian S. (2020). *Authentication using robust primary PIN (Personal Identification Number), multifactor authentication for credit card swipe and online transactions security.* International Journal of Advanced Computer Science and Applications, 11(4).

[4] Ileberi E., Sun Y. & Wang Z. (2022). *A machine learning based credit card fraud detection using the GA algorithm for feature selection.* Journal of Big Data, 9(1), 1-17.

[5] Sasikala G., Laavanya M., Sathyasri B., Supraja C., Mahalakshmi V., Mole S. S. & Dejene M. (2022). *An Innovative Sensing Machine Learning Technique to Detect Credit Card Frauds in Wireless Communications.* Wireless Communications and Mobile Computing, 2022.

[6] Vemuri, V.K. (2020). *The Hundred-Page Machine Learning Book.* Journal of Information Technology Case and Application Research, 22(2), pp. 136-138.

[7] Darwish S. M. (2020). *An intelligent credit card fraud detection approach based on semantic fusion of two classifiers.* Soft Computing, 24(2), 1243-1253.

[8] Shahraki, A., Abbasi, M., & Haugen, Ø. (2020). *Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost.* Engineering Applications of Artificial Intelligence, 94

[9] Baker M. R., Mahmood Z. N., & Shaker E. H. (2022). *Ensemble Learning with Supervised Machine Learning Models to Predict Credit Card Fraud Transactions.* Revue d'Intelligence Artificielle, 36(4), 509-518.

[10] Sasank J. S., Sahith G. R., Abhinav K., & Belwal M. (2019, July). *Credit card fraud detection using various classification and sampling techniques: a comparative study.* In 2019 International Conference on Communication and Electronics Systems (ICCES) (pp. 1713-1718). IEEE.

[11] Chen T., & Guestrin C. (2016, August). *Xgboost: A scalable tree boosting system.* In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[12] Izotova A., & Valiullin A. (2021). *Comparison of Poisson process and machine learning algorithms approach for credit card fraud detection.* Procedia Computer Science, 186, 721-726.

[13] Varmedja D., Karanovic M., Sladojevic S., Arsenovic M., & Anderla A. (2019, March). *Credit card fraud detection-machine learning methods.* In 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-5). IEEE.

[14] The Credit card fraud [Online]. https://www.kaggle.com/mlg-ulb/creditcardfraud

[15] Salekshahrezaee Z., Leevy J. L., & Khoshgoftaar T. M. (2023). *The effect of feature extraction and data sampling on credit card fraud detection.* Journal of Big Data, 10(1), 1-17.

[16] Fanai H., & Abbasimehr H. (2023). *A novel combined approach based on deep Autoencoder and deep classifiers for credit card fraud detection.* Expert Systems with Applications, 119562.

[17] Khatr S., Arora A., & Agrawal A. P. (2020, January). *Supervised machine learning algorithms for credit card fraud detection: a comparison.* In 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 680-683). IEEE.

[18] Alamri M., & Ykhlef M. (2022). *Survey of Credit Card Anomaly and Fraud Detection Using Sampling Techniques.* Electronics, 11(23), 4003.

[19] Akiba T., Sano S., Yanase T., Ohta, T., & Koyama, M. (2019, July). *Optuna: A next-generation hyperparameter optimization framework.* In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2623-2631).

[20] Rosley N., Tong G. K., Ng K. H., Kalid S. N., & Khor, K. C. (2022, December). *Autoencoders with Reconstruction Error and Dimensionality Reduction for Credit Card Fraud Detection.* In International Conference on Computer, Information Technology and Intelligent Computing (CITIC 2022) (pp. 503-512). Atlantis Press.

[21] Pugliese V. U., Hirata C. M., & Costa R. D. (2020). *Comparing Supervised Classification Methods for Financial Domain Problems.* In ICEIS (1) (pp. 440-451).