Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Applying Data Mining Techniques for Fraud Detection

Nawaf Farhan M Alsarrah, Ibrahim M. El-Hasnony, Hazem M. El-Bakry

Dept. of Information Systems, Faculty of Computer & Information Sciences, Mansoura University, Egypt

Recently after the Crona outbreak, the number of online transactions has increased many folds. This is due to lockdown orders and lack of face-to-face transactions. Even before the pandemic online commerce has grown steadily for the last few years. The number of fraud transactions is the number one cause of losses in the modern banking industry. There are many approaches for solving this problem, but after the development of the state of art machine learning based approaches for many fields, the utilization of these approaches in fraud detections has risen considerably. In this paper we proposed four stage machine learning based frameworks, the first stage is the feature selection based one support vector machine, the second stage is balancing techniques including random under sampling and borderline smote techniques, third stage is the hyper parameters optimization base on grid search approach and finally the classification based one logistic regression. The proposed approach achieved state-of-the-art performance accuracy of 99.74%. precision of 99.68%, recall 99.81% and finally f1-measure of 99.75%.
Keywords: Credit card fraud detection dataset.

1. Introduction

In recent years, the adaptation of utilization of smart payments systems and online payments systems has been blown up. Due to many factors including the pandemic and ease of use of contact less and smartwatch enabled approaches [1,2,3].

With increasing in the legal transitions comes an increase in the fraudulent transactions which lead to financial loss for all stakeholders. These include credit card holders, banks, payment processors and payment processors. This loss and increasing number of these types of transactions lead to widespread and huge investments in the fraud detections systems that can help fight back the fraud transition and limit the amount of loss with not affecting the ease of use and speed of the conduction the transactions. As a result of the huge popularity of credit

cards for conducting commercial activities online and advancement in the approach for making fraudulent transactions, the amount of loss affects all parties increasing exponentially [4,5,6].

With the advancement of new disruptive technologies like IoT and utilization of near field communications in smartphone and smartwatches in payment, due to the ease of use and popularity of non-touch applications due to Convd-19, this lead to new kinds of fraud utilizing new attacks and techniques to steal information and conduct fraud transactions [7,8,9].

For fighting back on the fraud transaction online, there is a huge arsenal of tools that can be used in this battle. Some of them are traditional and easily fooled by the new approaches of fraud and others are state of the art like datamining and machine learning and deep learning approaches. We are going to discuss some of these approaches, from these approaches the risk models which are one of the traditional approaches that use the type of transaction and the amount of money, date and time and the location for assign a score of fraud and after that take actions depending on this score which lead to assessment and reduction in fraud [10,11]

Another popular approach is utilization of biometric data like the face print and fingerprint and voice print, individually of collectivity for providing fraud free transactions online but these faces come challenges for device dependent sensors. This means if you do not have the right device that has the sensors you cannot utilize these services which make you at risk [12,13].another approach is the usage of the latest developments in the field of Machine learning artificial intelligence, a machine learning model that can predict future fraud transactions depending on past fraud transactions. This type of model needs to be updated regularly for adaptation for new approach of fraud and evolution approaches and techniques [14,15,16]. Another approach is based on anomaly detection which can help adapt to changes in approaches of fraud that is developing without the need for updating the model itself [17,18].

One of the most common approaches for handling huge amounts of data, and the same time fight back on fraud activities of all kinds is datamining, it Is the process of searching though huge volumes of data for useful and valuable insights. This information can be patterns and trends and valuable relation between data that can be used in many fields, like social media research, law enforcement and online fraud and insurance fraud detection [19,20,21].

The proposed model the utilization of datamining approaches for detection of online fraud transactions utilizing a benchmark dataset and comparing performance and providing a comprehensive solution for problems that are common in fraud transactions. We provided a solution for the imbalance nature of fraud dataset. We also provided approaches for improving performance of model by utilization features selection to reduce training and testing time and improve overall model performance, not only this but at the same time using automated hyper parameter optimization using grid search, that enables the model to choose best parameters. This enables our model to be able to achieve the best possible results compared to other approaches as shown in the next sections.

The main contribution points of the paper are the following: first we proposed using borderline-smote for oversampling the minority class, second we used support vector machine for feature selection, third we used grid search for hyper parameters optimization and fourth we used logistic regression for classification.

The rest of the paper is organized as follows: section two will focus on the related work about

using data mining approaches for fraud detection, third section will focus on background of the proposed approach, and fourth section will focus on the proposed approach and dataset , and fifth section will focus on metrics and the results of the proposed approach, and finally sixth section will focus on results and discussion the conclusion and future work.

2. Related Work

Jiang et al [22] proposed UAAD-FDNet employs unsupervised learning to identify fraudulent credit card transactions as anomalies. Autoencoders, enhanced with feature attention, and Generative Adversarial Networks (GANs) are utilized to effectively distinguish abnormal transactions from legitimate ones within large datasets. Rigorous evaluations on standard datasets confirm the superior performance of UAAD-FDNet compared to existing fraud detection techniques.

Salekshahrezaee et al [23] evaluate PCA and CAE for feature extraction and RUS, SMOTE, and SMOTE Tomek for data sampling on a credit card fraud dataset. Four ensemble classifiers (Random Forest, CatBoost, LightGBM, and XGBoost) are employed. The F1 score and AUC are used to assess performance. Our findings indicate that RUS followed by CAE yields the best results for credit card fraud detection.

Sisodia et al. [24] proposed a novel hybrid data sampling technique, SMOTEOSS, to address class imbalance. By combining SMOTE and OSS, SMOTEOSS generates synthetic minority class instances while removing noisy majority class instances, leading to improved model performance.

Bakhtiari et al. [25] investigated the efficacy of ensemble learning methods for a specific task/problem area. They combined gradient boosting algorithms like LightGBM and LiteMORT through averaging techniques to achieve enhanced performance and reduced error rates.

Scott et al. [26] introduced a novel BGP anomaly detection technique using Matrix Profile (MP). MP, a time series data mining approach, can be applied to various network data types without requiring domain-specific knowledge. It offers efficient and accurate anomaly detection in BGP event data.

Massi et al. [27] proposed a two-stage algorithm to identify outliers in hospital behavior related to specific disease treatments. The first stage groups of hospitals use k-means clustering, while the second stage employs a human-decision support system to validate outliers.

Deng et al. [28] investigated a transaction fraud detection system combining a random forest classifier with manual detection methods. This hybrid approach achieved high accuracy and AUC scores, demonstrating its effectiveness in identifying fraudulent transactions.

Aftabi et al. [29] introduced a novel approach for fraud detection in financial statements. They leveraged GANs to generate synthetic fraudulent data and ensemble models to handle high-dimensional feature spaces. This approach effectively addresses the challenges of limited non-fraudulent data and complex relationships.

Sahu et al. [30] investigated the efficacy of various classification algorithms for credit card

fraud detection. To address class imbalance, they employed data resampling and cost-based learning techniques, improving the accuracy of fraud detection models.

Ali et al. [31] provided a comprehensive review of fraud detection techniques and their applications in various sectors. They discussed methodologies, data pre-processing, visualization, and real-world applications, highlighting the importance of data mining and machine learning in combating financial crimes.

Settipalli et al. [32] introduced WMTDBC, a novel unsupervised multivariate analysis approach for detecting fraudulent healthcare claims. It analyzes categorical and continuous data to identify anomalies, improving fraud detection performance.

Seera et al. [33] investigated the efficacy of statistical and machine learning models for payment card fraud detection. By comparing models using original and aggregated features, they found that aggregated features offer superior discriminative power, leading to improved fraud detection performance.

Fanai et al. [34] presents a two-stage framework for fraud detection. The first stage utilizes a deep Autoencoder to learn robust representations of transaction data. The second stage employs supervised deep learning techniques to classify transactions as fraudulent or legitimate. Experimental results demonstrate that the proposed framework significantly outperforms baseline models trained on original or PCA-transformed data, achieving superior performance in terms of various evaluation metrics.

Almarshad et al. [35] proposes a novel approach to address the class imbalance problem in credit card fraud detection. By leveraging GANs to generate synthetic fraudulent transactions, the model significantly improves the performance of classification models. This approach outperforms traditional methods and demonstrates the potential of GANs in enhancing fraud detection accuracy.

Chu YB et al. [36] investigates the effectiveness of fundamental machine learning models, particularly SVM, in credit card fraud detection without the use of sampling techniques. The results demonstrate high accuracy and suggest that direct application of machine learning models on original datasets can yield promising results, simplifying the fraud detection process.

Damoun et al. [37] introduces G-HIN2VEC, a novel approach to learn graph-level representations for heterogeneous graphs. By leveraging unsupervised learning techniques and negative sampling, G-HIN2VEC effectively captures semantic information from the graph structure. The model's effectiveness is demonstrated through its application to real-world credit card data, outperforming traditional methods in various downstream tasks such as age, income, and gender prediction.

Jemai et al. [38] investigates the performance of ensemble learning methods, specifically XGBoost, in detecting credit card fraud on real-world and synthetic datasets. The results indicate that while ensemble models perform well on real-world data, they struggle with synthetic datasets, highlighting the importance of real-world data for effective fraud detection.

Alshutayri et al. [39] proposes the use of machine learning, specifically logistic regression, to detect fraudulent credit card transactions in the context of movie ticket purchases. By

analyzing a dataset of European cardholder transactions, the study demonstrates high prediction accuracy, highlighting the potential of machine learning in mitigating credit card fraud risks.

Dang et al. [40] proposes a federated learning approach for fraud detection in credit card transactions. By collaboratively training models without sharing sensitive data, federated learning enables banks to improve fraud detection accuracy while safeguarding privacy. The proposed FDS model achieved a high accuracy rate of 97% on the ECC dataset, demonstrating the effectiveness of federated learning for this task.

Feng et al. [41] addresses the critical issue of credit card fraud. By leveraging machine learning techniques and introducing a novel feature reduction method, Compact Data Learning (CDL), the study aims to improve the accuracy and efficiency of fraud detection systems. The findings contribute to the advancement of fraud detection and provide practical implications for the financial sector.

Abdullahi et al [42] explores the application of Artificial Neural Networks (ANN) to identify fraudulent and legitimate transactions in a European transaction dataset. To address the class imbalance issue, the study employs the SMOTE-Tomek technique and utilizes PCA for dimensionality reduction. The developed ANN model achieves an impressive accuracy of 97.86%, outperforming existing methods. The model's ability to minimize false positives and negatives, as demonstrated by high precision, recall, and F1-scores, highlights its potential for real-world fraud detection applications.

Table 1 summary of list of related work

PAPER	CONTRIBUTION	MODELING
[22]	Detection of anomalies of fraud transactions using autoencoders and Gans	Unsupervised learning ,Gans, auto encoders
[23]	Ensemble model and hybrid	PCA,CAE,Smote tomek
[24]	Hybrid data sampling technique to address class imbalance	SMOTE, OSS
[25]	Ensemble learning for improved performance	Gradient boosting algorithms, averaging techniques
[26]	Novel BGP anomaly detection technique using Matrix Profile	Time series data mining, Matrix Profile
[27]	Two-stage algorithm to identify outliers in hospital behavior	K-means clustering, human-decision support system
[28]	Transaction fraud detection system combining random forest and manual detection methods	Random forest, manual detection
[29]	Novel approach for fraud detection in financial statements using GANs and ensemble models	GANs, ensemble models
[30]	Investigation of classification algorithms for credit card fraud detection, addressing class imbalance	Classification algorithms, data resampling, cost-based learning
[31]	Comprehensive review of fraud detection techniques and applications	Various techniques, data mining, machine learning
[32]	Novel unsupervised multivariate analysis approach for detecting fraudulent healthcare claims	Weighted MultiTree, Density-Based Clustering

[33]	Comparison of models using original and aggregated features for payment card fraud detection	Statistical and machine learning models, feature engineering
[34]	Two-stage framework for fraud detection using deep learning	Deep Autoencoder, Supervised Deep Learning
[35]	Addressing class imbalance in credit card fraud detection using GANs	Generative Adversarial Networks (GANs)
[36]	Investigating SVM for credit card fraud detection without sampling	Support Vector Machine (SVM)
[37]	Learning graph-level representations for heterogeneous graphs using G-HIN2VEC	Graph Neural Networks (GNNs)
[38]	Evaluating ensemble learning (XGBoost) for credit card fraud detection	XGBoost
[39]	Using logistic regression for credit card fraud detection in movie ticket purchases	Logistic Regression
[40]	Federated learning for fraud detection in credit card transactions	Federated Learning
[41]	Improving credit card fraud detection using machine learning and feature reduction (CDL)	Machine Learning, Compact Data Learning (CDL)
[42]	Using ANN for credit card fraud detection with SMOTE-Tomek and PCA	Artificial Neural Networks (ANN), SMOTE-Tomek, PCA

3. Background

In this section we discuss the general view of the steps of the proposed framework, and the main motivation related to usage of each approach in the overall framework.

3.1 Feature selection

The goal of this stage is to reduce the number of features for each entry, this can lead to improved performance in two areas, first one is time consumed during training and testing and the second one is the accuracy of producing results. feature selection filters out the unnecessary features that do not correlate between dependent (entry feature) and independent (class label) features [65-82].

With binary and multi class classification, it is utilized for deciding which feature is correlated more with class label, the more correlation between the two the higher the chances that the feature is kept into the entry feature.

There are several approaches for feature selection, filter based [43] , Wrapper based [44], embedded based [45], online based [46], and hybrid [47] which consists of two or more approaches from the previously mentioned.

3.2 Sampling

It is a well-known approach for treating the problem of biased datasets or unbalanced datasets, usually there are three main approaches for sampling data, down sampling is simply focusing on the majority class, we reduce the number of the majority class to match the classes of the monitory class, do achieving there are many approaches like random under sampling [48] , oversampling and hybrid a mix between many types of over and down sampling into one

approach [49]. The difference between the two approaches is illustrated in figure 1 below.

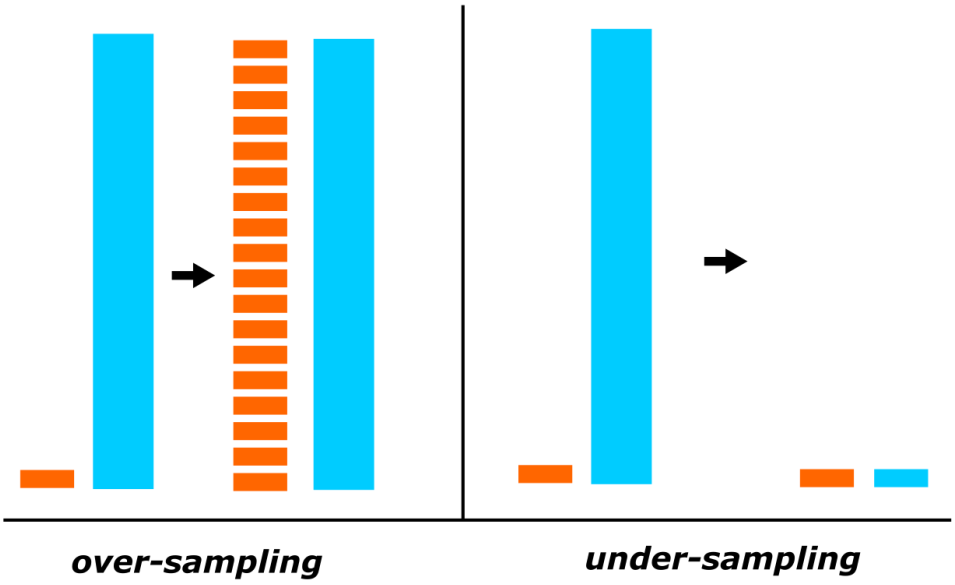

Figure 1 comparison between over sampling and under sampling

3.3 Hyper parameters optimization

One of the most important activities in machine learning, is the selecting best possible set or a group of hyperparameters, these parameters can have a huge impact on the performance of the model, but usually ignored. There are many approaches for hyperparameters optimization [50] approaches like search approaches and naturally inspired approaches, Bayesian approaches, and surrogate model approaches, we will focus on the searching approaches the grid search and random search approaches.

3.4 Grid search

We test every possible combination of parameters [51] and choose the one with the highest performance. This approach is ideal when we have plenty of time and resources to handle this task as sometimes this can take much longer than the training of the model itself, so it is usually a one-time operation.

3.5 Random search

A set of new hyper parameters [52] is chosen at random from the set of possible parameters, this reduces the time of the search operation considerably but choosing the best possible combination is not guaranteed as it usually depends on the parameters chosen in the search set. The difference is illustrated in the figure 2

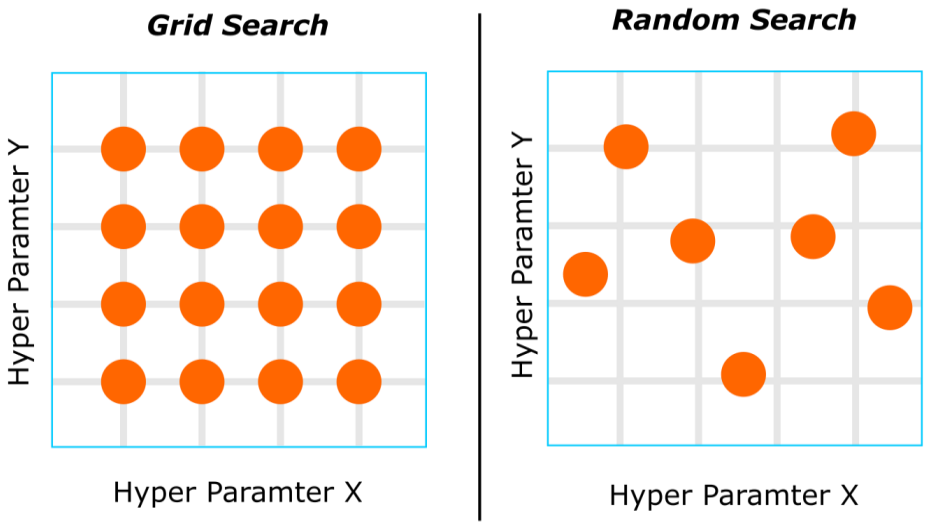

Figure 2 comparison between grid search and random search techniques

3.6 Logistic regression

Logistic regression [53] is simply one of the simplest machine learning approaches and best used in binary classification, as in our cases we have either fraud or non-fraud transactions. It is simply using mathematical analysis to define the relation between dependent (features) and independent variables (class label). In equation 1 we can see the mathematical basis of the logistic regression

$$f(x) = \frac{1}{1 - e^{-x}} \tag{1}$$

In the following figure 3 we show how logistic regression works,

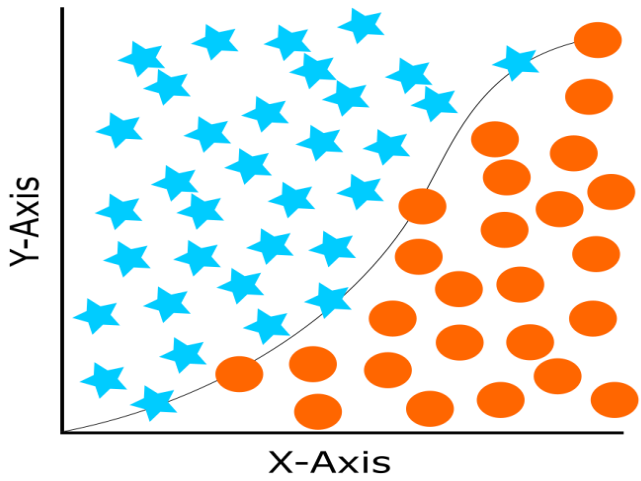

Figure 3 logistic regression

4. Proposed Model

Our proposed approach consists of serval stages, each stage takes part in the overall improvement and enhancing the performance of approach. Our approach consists of up to four main stages, first stage is the feature selection, second stage is sampling, third stage is hyper parameters optimization and final stage is the classification.

4.1 Sampling

In our approach we utilized the hybrid of random under sampling and borderline smote. Which in our case leads to improved overall performance. In random down sampling, the approach selects samples from majority class and removes until the minority and majority class have the same number of samples, we fixed the results of the function by using a fixed random_state 42 as shown in table 2.

Table 2 parameters of random undersealing functions

Parameter	Value
random_state	42

In figure 4 we illustrated the mechanism of the random under sampling, it is simply removes at random samples from the majority class to make it the same size as the minority class.

Figure 4 Random under sampling

Borderlines Smote [54] , samples from the borderline between classes will be used. The process of the generation of new samples to improve the performance of the original smote. Exactly the samples of the minority class near the borderline with the majority classes improves accuracy of the generated samples considerably , in our approach we used the function with the default parameters but with the random state of 42. And it is presented in table 3.

Table 3 parameters of the borderline Smote

Parameter	Value
random_state	42

In figure 5 we illustrate how it works, simply considers the samples near the border of the majority in the generation process of a new synthesized samples, this provides huge improvement over the ordinary smote approach.

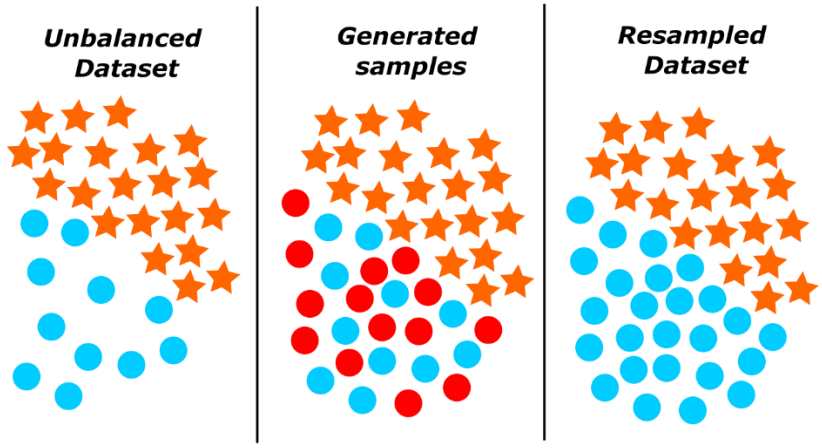

Figure 5 borderline smote mechanism

Figure 6 Borderline Smote Oversampling

In figure 6 we simply show the borderline effects on an imbalanced dataset by generation of new samples that lead in the end to newly balanced dataset.

4.2 Feature selection

We utilized support vector machine [55] into feature selection process. SVM can work as an approach for selecting the best possible set of features that help improve the overall performance of the model. In the following table are the hyper parameters that are used in the

linear SVM model for feature selection in table 4 we present the used hyper parameters of SVM

Table 4 table of hyper parameters of the linear support vector machine used for features selection

Parameter	Value
C	0.01
penalty	l1
dual	False
Random_State	42

Where c is the regularization parameters, the lower the regularization parameters the higher the regularization effect, penalty is the norm that is used for penalization, false for dual means that the model is working with data that is number samples is larger compared to the number of features. Manual hyperparameters optimization where used for the selection of parameters.

And in figure 7, we present the mechanism of working of SVM, as the features in are reduces or the same as features out which improves the performance as we explained before

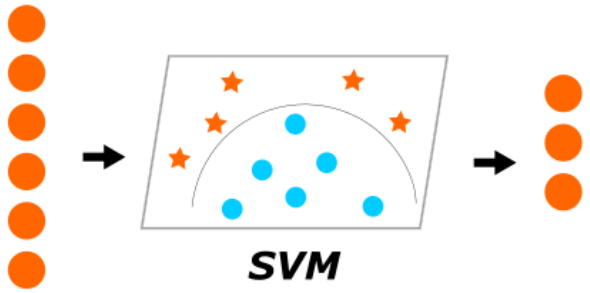

Figure 7 SVM Feature Selection

3.3 Hyper parameter optimization

Grid search [56] is one of the basic approaches for hyper parameters optimization or tuning. Simply it is creating a grid or all different possibilities of the parameters and gives us in the end a model with the best parameters, it means it tests every combination of parameters in the grid, it is slow, but it is much better than doing this manually.

The pseudocode is presented in the listing 1, simply put it searching for the best possible results based on given combination of set of parameters as shown in the listing, simply put it will iterate though the entire combination using the nested looping structures and then present in the end best performance in terms of given combination of parameters which achieved these goals.

```
param_grid = {
    'parameter_1': [value1, value2, value3],
    'parameter_2': [valueA, valueB],
    # Add more parameters as needed
}

for param1_value in param_grid['parameter_1']:
    for param2_value in param_grid['parameter_2']:
        # Set the model parameters
        model.set_params(parameter_1=param1_value, parameter_2=param2_value)

        # Train the model (e.g., on training data)
        model.fit(X_train, y_train)
        # Evaluate the model (e.g., on validation data)
        score = model.evaluate(X_val, y_val)
        # Update the best score and parameters if the current score is better
        if score > best_score:
            best_score = score
            best_params = {'parameter_1': param1_value, 'parameter_2': param2_value}
```

Listing 1 shows the pseudocode of the grid search algorithms

Table 5 set of hyperparameters

Parameter	Values
penalty	'l1' , 'l2'
C	0.001, 0.01, 0.1, 1, 10, 100, 1000

In tables 5 , 6 we list all hyperparameters that we used for the penalty and for the c parameters.

Table 6 grid search and the selected best parameters

	0.001	0.01	0.1	10	100	1000
l1	{l1 , 0.001 }	{l1 , 0.01 }	{l1 ,0.1 }	{l1 , 10}	{l1 , 100 }	{l1 , 1000 }
l2	{l2 , 0.001 }	{l2 , 0.01 }	{l2 ,0.1 }	{l2 , 10 }	{l2 , 100 }	{l2 , 1000 }

In table 6 we presented the list of all parameters in the grid, and we highlighted the selected hyperparameters for our model. In figure 8 we present the grid and the search set of the proposed parameters.

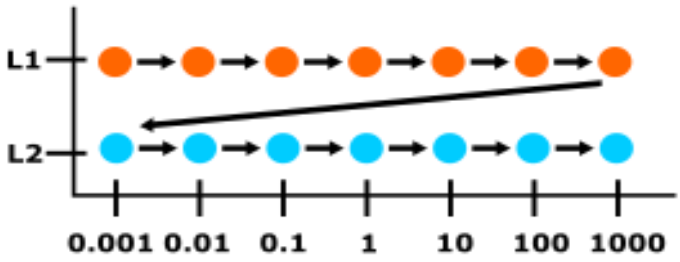

Figure 8 grid search illustration

4.4 Classification

We focus on finding the best fit for our application we choose classical machine learning model as our data is simply tabular, classical models are best known for find hidden pattern of relation in this type of data.

We use logistic regression as a classifier which provides great performance with the help of other enhancing steps like sampling, feature selection and hyperparameter optimization. The selections of the model for classification approaches will add little to the overall performance of the proposed approach. But logistic regression is known to be good at finding relation between label class and the sample features which are based on explored the hidden pattern of relation between every feature and the class label. The proposed framework is illustrated in figure 9.

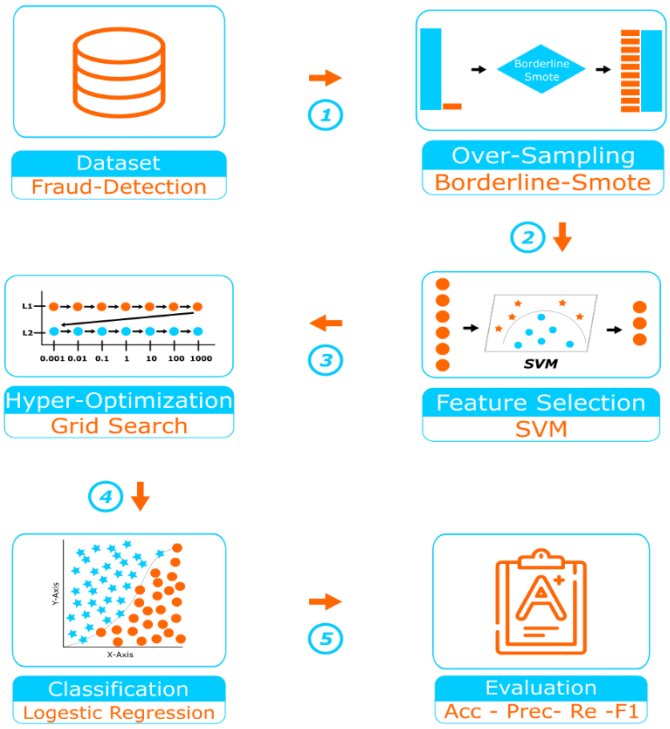

Figure 9 proposed framework

5. Results and discussion

5.1 Dataset description

The data was collected by European stakeholders [29] in the month of September in 2013 over two days. This dataset contains up to 31 different features per case. And it had up to 284,807 cases. With as little as 492 this accounts for up to 0.172% of fraud transactions. This means

the dataset is hugely unbalanced toward the legal transactions. Which makes the task of detection of the fraud transactions much harder task. Due to confidentiality of the data most features are results of Principal Component Analysis PCA for the protection of user data and their privacy. Only three features were not changed using PCW, they are class, amount and time. And the huge difference is shown in the following. One important feature of this dataset has many names online, it is called European cardholder dataset, and credicard fraud detection dataset.

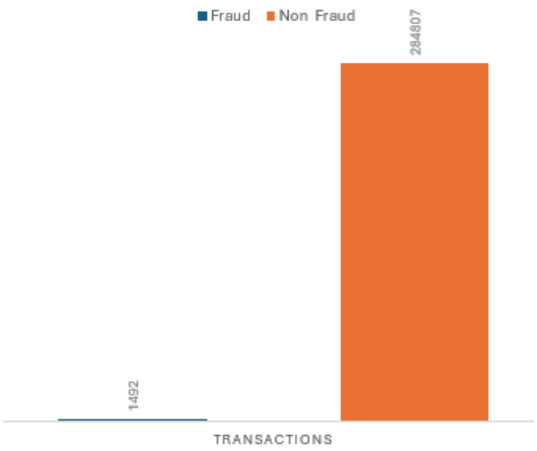

Figure 10 dataset distribution between fraud and not fraud classes

5.2 Comparison Metrics

Accuracy [39]

One of the most common metrics for assessing the accuracy of classification or prediction models is accuracy. It is computed by dividing the sum of correct predictions (true positives and true negatives) by the total number of predictions (both correct and incorrect). The accuracy can be calculated using the following formula.

$$\text{Accuracy} = \frac{(\text{TP}) + (\text{TN})}{(\text{TP}) + (\text{FP}) + (\text{TN}) + (\text{FN})} \quad (2)$$

Precision [40]

Precision is a classification metric that focuses primarily on the positive cases. It is calculated by dividing the number of true positive predictions by the sum of true positive and false positive predictions. This metric is also known as positive predictive value in fields outside of computer science and machine learning.

$$\text{Precision} = \frac{(\text{TP})}{(\text{TP}) + (\text{FP})} \quad (3)$$

Recall [40]

Recall, also known as sensitivity, is a classification metric that measures the ability of a model to correctly identify positive cases. It is calculated by dividing the number of true positive predictions by the total number of actual positive cases.

$$\text{Recall} = \frac{(\text{TP})}{(\text{TP}) + (\text{FN})} \tag{4}$$

F1-measure [40]

F1-score is a combined metric that considers both precision and recall. It is calculated as the harmonic mean of precision and recall. While F1-score is widely used, it can be less informative in cases of imbalanced datasets.

$$\text{F1} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{5}$$

Our model performance compared to other approaches shows state of the art performance compared to other models on benchmark datasets as shown in table 2.

5.3 Results and discussion

Our proposed approach achieved state-of-the-art performance as shown in the next table 7 and table 8. First, we conducted 6 different experiments that showed a steady improvement of the performance after adding each stage of our proposed approach. We divided out dataset into 20% testing and 80% training split. And the results of the different steps are illustrated in figure 11.

Table 7 Credit card fraud detection dataset experiments

Approach	Step	Accuracy	Precision	Recall	F-1
Logestic Regression (LR)	1	99.9%	71.3%	73.2%	72.2%
LR + grid search (GS) hyper parameters optimization	2	99.98%	69.6%	66.98%	68.2%
LR + GS +RUS	3	92.38%	93.87%	91.08%	92.46%
LR + GS +SVM+RUS	4	93.90%	97.75%	89.69%	93.54%
LR+GS+SVM	5	99.93%	88.88%	74.76%	81.21%
LR+GS+bl-smote	6	98.81%	98.75%	98.87%	98.81%
LR+GS+SVM+bl-smote	7	99.74%	99.68%	99.81%	99.75%
LR+GS+RUS +bl-smote	8	98.79%	98.72%	98.87%	98.79%
LR+GS+ SVM+ RUS +bl-smote	9	99.71%	99.63%	99.80%	99.71%

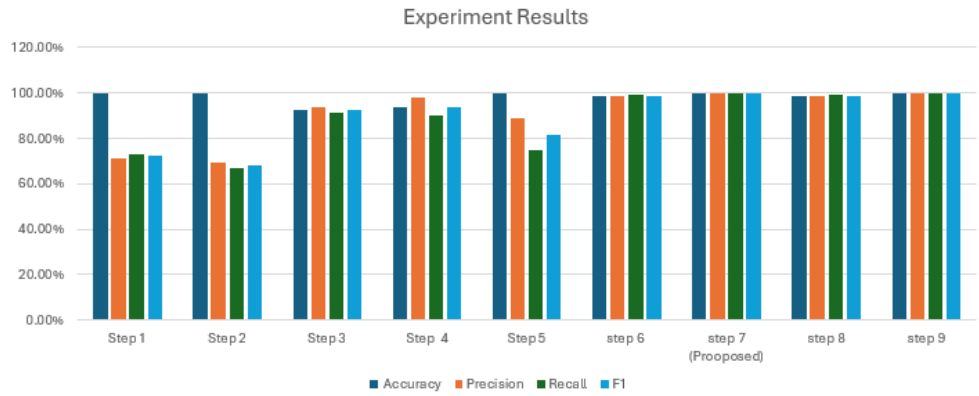

Figure 11 results of the 6 experiments

Our proposed framework was able to outperform other models and beat the next highest model which achieved up to 98% accuracy and 98% precision and recall 98% and finally F1 99%. Our framework achieved up to 99.74% accuracy, 99.69% Precision and 99.8% recall and finally achieved up to 99.75% f1-measure. The results are illustrated in table 8 And illustrated in figure 12.

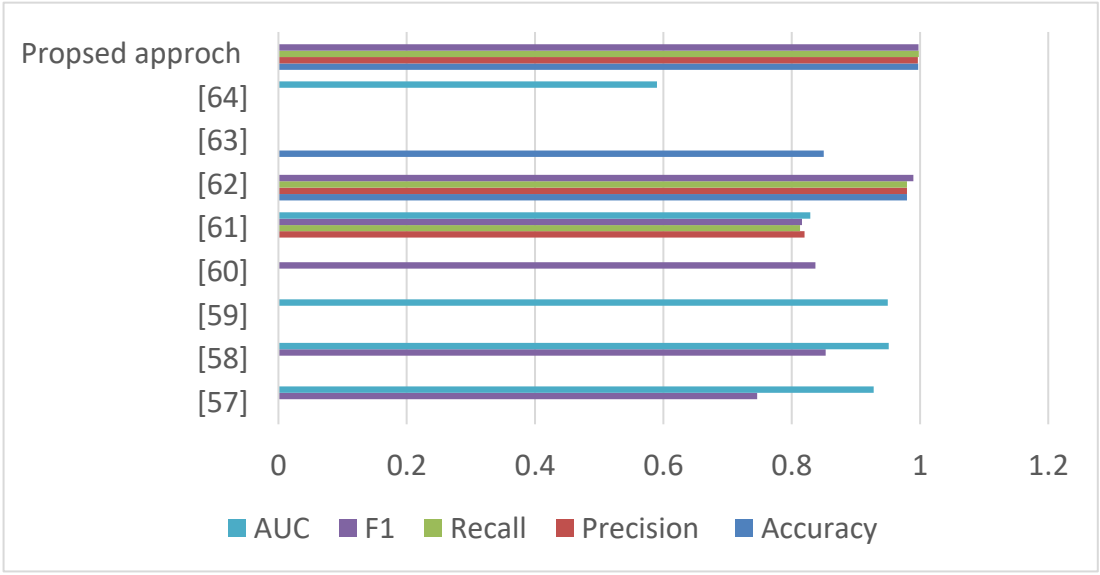

Figure 12 Comparison between our approach and other work on the same dataset

Table 8 Comparison between proposed framework and other models

paper	Accuracy	Precision	Recall	F1	AUC
[57]	N/A	N/A	N/A	81.5%	91.5%
	N/A	N/A	N/A	73.3%	94.0%
	N/A	N/A	N/A	81.8%	91.3%

	N/A	N/A	N/A	74.6%	92.8%
[58]	N/A	N/A	N/A	84.89%	94.37%
	N/A	N/A	N/A	85.29%	95.15%
[59]	N/A	N/A	N/A	N/A	94.96%
	N/A	N/A	N/A	N/A	94.09%
	N/A	N/A	N/A	N/A	90.91%
	N/A	N/A	N/A	N/A	90.84%
[60]	N/A	N/A	N/A	71.01%	N/A
	N/A	N/A	N/A	81.38%	N/A
	N/A	N/A	N/A	83.72%	N/A
[61]	N/A	82%	81.3%	81.6%	82.9%
[62]	98%	98%	98%	99%	N/A
[63]	85%	N/A	N/A	N/A	N/A
[64]	N/A	N/A	N/A	N/A	59%
Our approach	99.74%	99.68%	99.81%	99.75%	N/A

6. Conclusion and Future work

We achieved state of the are performance on the credicard fraud dataset benchmark dataset, as we achieved state of art performance that was able to achieve up to accuracy of 99.74%. precision of 99.68%, recall 99.81% and finally f1-measure of 99.75%. as we have provided a complete solutions for all problems that may face any model in the field of the credit card fraud detection or any other classification problem. As we have provided a hybrid approach for sampling combining the benefits of over-sampling and under-sampling, then we utilized the features selections using SVM and finally using logistic regression for classification approach.

We consider in the future work the utilization of the framework with other fraud datasets, not only this but apply this solution to other datasets to provide the proof that the proposed framework is able to improve the performance of the classification not only on the fraud dataset but on the wider field of classification, especially the tabular datasets. We also consider utilization of other approaches in every step in the framework. For example, we can use the random search instead of the grid search in the hyper parameter optimization.

References

1. Büşra AĞ. The impact of COVID-19 pandemic process on digital payment system: The case of Turkey. Avrasya Sosyal ve Ekonomi Araştırmaları Dergisi. 2020;7(7):229-40.
2. Khan F, Ateeq S, Ali M, Butt N. Impact of COVID-19 on the drivers of cash-based online transactions and consumer behaviour: evidence from a Muslim market. Journal of Islamic Marketing. 2023 Feb 10;14(3):714-34.
3. Bechlioulis AP, Karamanis D. Consumers' changing financial behavior during the COVID-19 lockdown: the case of Internet banking use in Greece. Journal of Financial Services Marketing.

2023 Sep;28(3):526-43.
4. Karpoff JM. The future of financial fraud. Journal of Corporate Finance. 2021 Feb 1;66:101694.
5. Monteith S, Bauer M, Alda M, Geddes J, Whybrow PC, Glenn T. Increasing cybercrime since the pandemic: Concerns for psychiatry. Current psychiatry reports. 2021 Apr;23:1-9.
6. Kemp S, Miró-Llinares F, Moneva A. The dark figure and the cyber fraud rise in Europe: Evidence from Spain. European Journal on Criminal Policy and Research. 2020 Sep;26(3):293-312.
7. Hassan M, Veena C, Singla A, Joshi A, Lourens M. Fraud Detection in IoT-Based Financial Transactions Using Anomaly Detection Techniques. In2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI) 2024 May 9 (pp. 1-6). IEEE.
8. Varela-Vaca ÁJ, Gasca RM, Iglesias D, Gónzalez-Gutiérrez JM. Automated trusted collaborative processes through blockchain & IoT integration: The fraud detection case. Internet of Things. 2024 Apr 1;25:101106.
9. Soleymanzadeh R, Aljasim M, Qadeer MW, Kashef R. Cyberattack and fraud detection using ensemble stacking. AI. 2022 Jan 18;3(1):22-36.
10. Lu W, Zhao X. Research and improvement of fraud identification model of Chinese A-share listed companies based on M-score. Journal of Financial Crime. 2021 Jun 4;28(2):566-79.
11. MADAH MARZUKI M, NIK ABDUL MAJID WZ, AZIS NK, ROSMAN R, HAJI ABDULATIFF NK. Fraud risk management model: A content analysis approach. The Journal of Asian Finance, Economics and Business. 2020;7(10):717-28.
12. Subash A, Song I. Real-time behavioral biometric information security system for assessment fraud detection. In2021 IEEE international conference on computing (ICOCO) 2021 Nov 17 (pp. 186-191). IEEE.
13. Chigada JM. A qualitative analysis of the feasibility of deploying biometric authentication systems to augment security protocols of bank card transactions. South African Journal of Information Management. 2020;22(1):1-9.
14. Afriyie JK, Tawiah K, Pels WA, Addai-Henne S, Dwamena HA, Owiredu EO, Ayeh SA, Eshun J. A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions. Decision Analytics Journal. 2023 Mar 1;6:100163.
15. Janani S, Sivarathinabala M, Anand R, Ahamad S, Usmani MA, Basha SM. Machine Learning Analysis on Predicting Credit Card Forgery. InInternational Conference On Innovative Computing And Communication 2023 Feb 17 (pp. 137-148). Singapore: Springer Nature Singapore.
16. Chethana C, Pareek PK. Analysis of Credit Card Fraud Data Using Various Machine Learning Methods. InBig Data, Cloud Computing and IoT 2023 Apr 19 (pp. 103-116). Chapman and Hall/CRC.
17. Vanini P, Rossi S, Zvizdic E, Domenig T. Online payment fraud: from anomaly detection to risk management. Financial Innovation. 2023 Mar 13;9(1):66.
18. Jiang S, Dong R, Wang J, Xia M. Credit card fraud detection based on unsupervised attentional anomaly detection network. Systems. 2023 Jun 13;11(6):305.
19. Al-Hashedi KG, Magalingam P. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. Computer Science Review. 2021 May 1;40:100402.
20. Sánchez-Aguayo M, Urquiza-Aguiar L, Estrada-Jiménez J. Fraud detection using the fraud triangle theory and data mining techniques: A literature review. Computers. 2021 Sep 30;10(10):121.
21. Ashtiani MN, Raahemi B. Intelligent fraud detection in financial statements using machine learning and data mining: a systematic literature review. Ieee Access. 2021 Jul 13;10:72504-25.
22. Jiang S, Dong R, Wang J, Xia M. Credit card fraud detection based on unsupervised attentional anomaly detection network. Systems. 2023 Jun 13;11(6):305.

23. Salekshahrezaee Z, Leevy JL, Khoshgoftaar TM. The effect of feature extraction and data sampling on credit card fraud detection. Journal of Big Data. 2023 Jan 17;10(1):6.
24. Sisodia D, Sisodia DS. A hybrid data-level sampling approach in learning from skewed user-click data for click fraud detection in online advertising. Expert Systems. 2023 Feb;40(2):e13147.
25. Bakhtiari S, Nasiri Z, Vahidi J. Credit card fraud detection using ensemble data mining methods. Multimedia Tools and Applications. 2023 Aug;82(19):29057-75.
26. Scott BA, Johnstone MN, Szewczyk P, Richardson S. Matrix Profile data mining for BGP anomaly detection. Computer Networks. 2024 Apr 1;242:110257.
27. Massi MC, Ieva F, Lettieri E. Data mining application to healthcare fraud detection: a two-step unsupervised clustering method for outlier detection with administrative databases. BMC medical informatics and decision making. 2020 Dec;20:1-1.
28. Deng W, Huang Z, Zhang J, Xu J. A data mining based system for transaction fraud detection. In2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE) 2021 Jan 15 (pp. 542-545). IEEE.
29. Aftabi SZ, Ahmadi A, Farzi S. Fraud detection in financial statements using data mining and GAN models. Expert Systems with Applications. 2023 Oct 1;227:120144.
30. Sahu A, Harshvardhan GM, Gourisaria MK. A dual approach for credit card fraud detection using neural network and data mining techniques. In2020 IEEE 17th India council international conference (INDICON) 2020 Dec 10 (pp. 1-7). IEEE.
31. Ali SH, Raslan AT. Using Data Mining Techniques for Fraud Detection in The Non-banking Sector. Journal of Computing and Communication. 2024 Jan 31;3(1):132-42.
32. Settipalli L, Gangadharan GR. WMTDBC: An unsupervised multivariate analysis model for fraud detection in health insurance claims. Expert Systems with Applications. 2023 Apr 1;215:119259.
33. Seera M, Lim CP, Kumar A, Dhamotharan L, Tan KH. An intelligent payment card fraud detection system. Annals of operations research. 2024 Mar;334(1):445-67.
34. Fanai H, Abbasimehr H. A novel combined approach based on deep Autoencoder and deep classifiers for credit card fraud detection. Expert Systems with Applications. 2023 May 1;217:119562.
35. Almarshad FA, Gashgari GA, Alzahrani AI. Generative adversarial networks-based novel approach for fraud detection for the european cardholders 2013 dataset. IEEE Access. 2023 Sep 27.
36. Chu YB, Lim ZM, Keane B, Kong PH, Elkilany AR, Abusetta OH. Credit Card Fraud Detection on Original European Credit Card Holder Dataset Using Ensemble Machine Learning Technique. Journal of Cyber Security. 2023;5:33-46.
37. Damoun F, Seba H, Hilger J. Graph-Level Heterogeneous Information Network Embeddings for Cardholder Transaction Analysis.
38. Jemai J, Zarrad A, Daud A. Identifying Fraudulent Credit Card Transactions using Ensemble Learning. IEEE Access. 2024 Mar 22.
39. Alshutayri A. Fraud Prediction in Movie Theater Credit Card Transactions using Machine Learning. Engineering, Technology & Applied Science Research. 2023 Jun 2;13(3):10941-5.
40. Dang TK, Ha T. A Comprehensive Fraud Detection for Credit Card Transactions in Federated Averaging. SN Computer Science. 2024 May 23;5(5):578.
41. Feng X, Kim SK. Novel Machine Learning Based Credit Card Fraud Detection Systems. Mathematics. 2024 Jan;12(12):1869.
42. Abdullahi MJ. Leveraging Machine Learning in Classifying Fraudulent and Legitimate Transactions in Banking Sector. Ilorin Journal of Computer Science and Information Technology. 2024 Jun 30;7(1):40-64.
43. Rahmaninia M, Moradi P. OSFSMI: online stream feature selection method based on mutual

information. Appl Soft Comput. 2017
44. Sanz H, Valim C, Vegas E, Oller JM, Reverter F. SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. BMC bioinformatics. 2018 Dec;19:1-8.
45. Xiao J, Cao H, Jiang X, Gu X, Xie L. GMDH-based semi-supervised feature selection for customer classification. Knowl-Based Syst. 2017.
46. Goswami S, Das AK, Chakrabarti A, Chakraborty B. A feature cluster taxonomy based feature selection technique. Expert Systems with Applications. 2017 Aug 15;79:76-89.
47. Wu Y, Liu Y, Wang Y, Shi Y, Zhao X. JCDSA: a joint covariate detection tool for survival analysis on tumor expression profiles. BMC bioinformatics. 2018 Dec;19:1-8.
48. Yang C, Fridgeirsson EA, Kors JA, Reps JM, Rijnbeek PR. Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. Journal of big data. 2024 Jan 3;11(1):7.
49. Hussin Adam Khatir AA, Bee M. Machine learning models and data-balancing techniques for credit scoring: What is the best combination?. Risks. 2022 Aug 24;10(9):169.
50. Bischl B, Binder M, Lang M, Pielok T, Richter J, Coors S, Thomas J, Ullmann T, Becker M, Boulesteix AL, Deng D. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2023 Mar;13(2):e1484.]
51. Zhao Y, Zhang W, Liu X. Grid search with a weighted error function: Hyper-parameter optimization for financial time series forecasting. Applied Soft Computing. 2024 Mar 1;154:111362.
52. Vo HT, Hoang TN, Quach LD. An approach to hyperparameter tuning in transfer learning for driver drowsiness detection based on bayesian optimization and random search. International Journal of Advanced Computer Science and Applications. 2023;14(4).0
53. LaValley MP. Logistic regression. Circulation. 2008 May 6;117(18):2395-9.
54. Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. InInternational conference on intelligent computing 2005 Aug 23 (pp. 878-887). Berlin, Heidelberg: Springer Berlin Heidelberg.
55. Cortes C. Support-Vector Networks. Machine Learning. 1995.
56. Liashchynskyi P, Liashchynskyi P. Grid search, random search, genetic algorithm: a big comparison for NAS. arXiv preprint arXiv:1912.06059. 2019 Dec 12.
57. Salekshahrezaee Z, Leevy JL, Khoshgoftaar TM. The effect of feature extraction and data sampling on credit card fraud detection. Journal of Big Data. 2023 Jan 17;10(1):6.
58. Jiang S, Dong R, Wang J, Xia M. Credit card fraud detection based on unsupervised attentional anomaly detection network. Systems. 2023 Jun 13;11(6):305.
59. Leevy JL, Hancock J, Khoshgoftaar TM. Comparative analysis of binary and one-class classification techniques for credit card fraud data. Journal of Big Data. 2023 Jul 17;10(1):118.
60. Al Balawi S, Aljohani N. Credit-card fraud detection system using neural networks. Int. Arab J. Inf. Technol.. 2023 Mar 1;20(2):234-41.
61. Zhao C, Sun X, Wu M, Kang L. Advancing financial fraud detection: Self-attention generative adversarial networks for precise and effective identification. Finance Research Letters. 2024 Feb 1;60:104843.
62. El Hlouli FZ, Riffi J, Mahraz MA, Yahyaouy A, El Fazazy K, Tairi H. Credit Card Fraud Detection: Addressing Imbalanced Datasets with a Multi-phase Approach. SN Computer Science. 2024 Jan 9;5(1):173.
63. Maithili K, Kumar TS, Subha R, Murthy PS, Sharath MN, Gupta KG, Ravuri P, Madhuri TN, Verma V. Development of an efficient machine learning algorithm for reliable credit card fraud identification and protection systems. InMATEC Web of Conferences 2024 (Vol. 392, p. 01116). EDP Sciences.
64. Zhu H, Zhou M, Xie Y, Albeshri A. A self-adapting and efficient dandelion algorithm and its

application to feature selection for credit card fraud detection. IEEE/CAA Journal of Automatica Sinica. 2024 Jan 29;11(2):377-90.

65. Hazem M. El-Bakry "Fast Iris Detection for Personal Verification Using Modular Neural Networks," Proc. of the 7th Fuzzy Days International Conference, Dortmund, Germany, October 1-3, 2001, pp. 269-283.
66. Hazem M. El-Bakry, "Fast Virus Detection by using High Speed Time Delay Neural Networks," Journal of Computer Virology, vol.6, no.2, 2010, pp.115-122.
67. Hazem El-Bakry, "Face Detection Using Neural Networks and Image Decomposition," Proc. of INNS-IEEE International Joint Conference on Neural Networks, 12-17 May, 2002, Honolulu, Hawaii, USA.
68. Hazem M. El-Bakry, M. A. Abo-elsoud, and M. S. Kamel, "Fast Modular Neural Networks for Human Face Detection," Proc. of IEEE-INNS-ENNS International Joint Conference on Neural Networks, Como, Italy, Vol. III, pp. 320-324, 24-27 July, 2000.
69. Hazem El-Bakry, "Face Detection Using Neural Networks and Image Decomposition," Proc. of INNS-IEEE International Joint Conference on Neural Networks, 12-17 May, 2002, Honolulu, Hawaii, USA.
70. Hazem M. El-Bakry, and Nikos Mastorakis, "Realization of E-University for Distance Learning," WSEAS Transactions on Computers, vol. 8, issue 1, Jan. 2009, pp. 48-62.
71. Hazem El-Bakry: "Comments on Using MLP and FFT for Fast Object/Face Detection," Proc. of IEEE IJCNN'03, Portland, Oregon, pp. 1284-1288, July, 20-24, 2003.
72. Hazem M. El-Bakry, and Qiangfu Zhao, "Speeding-up Normalized Neural Networks For Face/Object Detection," Machine Graphics & Vision Journal (MG&V), vol. 14, No.1, 2005, pp. 29-59.
73. Hazem M. El-Bakry, "New Fast Time Delay Neural Networks Using Cross Correlation Performed in the Frequency Domain," Neurocomputing Journal, vol. 69, October 2006, pp. 2360-2363.
74. Hazem M. El-Bakry and Mohamed Hamada, "A New Implementation for High Speed Neural Networks in Frequency Space," Lecture Notes in Artificial Intelligence, Springer, KES 2008, Part I, LNAI 5177, pp. 33-40.
75. Hazem M. El-Bakry, and Qiangfu Zhao, "Fast Time Delay Neural Networks," International Journal of Neural Systems, vol. 15, no.6, December 2005, pp.445-455.
76. Hazem M. El-Bakry, "Human Iris Detection Using Fast Cooperative Modular Neural Nets," Proc. of INNS-IEEE International Joint Conference on Neural Networks, pp. 577-582, 14-19 July, 2001, Washington, DC, USA.
77. Hazem M. El-Bakry, "New Fast Principal Component Analysis for Face Detection," Journal of Advanced Computational Intelligence and Intelligent Informatics, vol.11, No.2, 2007, pp. 195-201.
78. Hazem M. El-Bakry, and Qiangfu Zhao, "Fast Normalized Neural Processors For Pattern Detection Based on Cross Correlation Implemented in the Frequency Domain," Journal of Research and Practice in Information Technology, Vol. 38, No.2, May 2006, pp. 151-170.
79. Menna Elkhateeb, Abdulaziz Shehab, and Hazem El-bakry, "Mobile Learning System for Egyptian Higher Education Using Agile-Based Approach," Education Research International, Volume 2019, Article ID 7531980, 13 pages.
80. Hazem M. El-Bakry, "A Novel High Speed Neural Model for Fast Pattern Recognition," Soft Computing Journal, vol. 14, no. 6, 2010, pp. 647-666.
81. Hazem El-Bakry, "Fast Face Detection Using Neural Networks and Image Decomposition," Proc. of the 6th International Computer Science Conference, AMT 2001, Hong Kong, China, December 18-20, 2001, pp.205-215.
82. Hazem M. El-Bakry, "An Efficient Algorithm for Pattern Detection using Combined Classifiers and Data Fusion," Information Fusion Journal, vol. 11, issue 2, April 2010, pp. 133-148.