Journal of Theoretical and Applied Information Technology
15th November 2024. Vol.102. No. 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ADVANCED CREDIT CARD FRAUD DETECTION: A NOVEL APPROACH INTEGRATING ADVERSARIAL-GUIDED OVERSAMPLING WITH MACHINE LEARNING

[1] ABDEL RAHMAN AMR, [2] WAEL HASSAN GOMAA*, [3] FARID ALI MOUSA

[1] Computer Science Department, Faculty of Computer Science, October University for Modern Science and Arts, Egypt.
[2]Information System Department, Faculty of Computers and Artificial Intelligence, Beni-Suef University, Egypt.
[3]Information Technology Department, Faculty of Computers and Artificial Intelligence, Beni-Suef University, Egypt
Email: [1] abdelrahman.amr3@msa.edu.eg, *[2] wael.goma@gmail.com*,[3] *farid.cs@gmail.com*

ABSTRACT

Fraud detection in credit card transactions presents a significant challenge due to the highly imbalanced nature of the data, where fraudulent transactions make up only a small fraction of the total. In this paper, we introduce a novel approach to address this issue by integrating adversarial-guided oversampling with machine learning techniques. Our method enhances the detection of fraudulent transactions by focusing on the minority class, using decision trees and neural networks to guide the generation of synthetic data samples. These samples are created through adversarial processes and validated by a neural network trained to distinguish between real and synthetic transactions. The proposed framework significantly improves the performance of traditional machine learning models, achieving remarkable accuracy, precision, recall, and F1 scores. Specifically, our method yields an accuracy of 0.9968, with precision, recall, and F1 scores all exceeding 0.995. This superior performance is a result of effectively addressing the class imbalance in the dataset, leveraging advanced sampling techniques, and employing robust machine learning classifiers. By enhancing the identification of fraudulent activities, our approach provides a substantial improvement in fraud detection systems for credit card transactions, ultimately offering a more reliable and efficient solution to this critical problem.

Keywords *Credit Card, Fraud Classification, Fraud Detection Techniques*

1. INTRODUCTION

Over the past decade, internet usage has grown significantly, leading to the rapid expansion and increased popularity of services such as e-commerce, tap-and-pay systems, and online bill payments. Consequently, credit card fraud has surged, with cybercriminals becoming more aggressive in their attacks on transactions. Various safeguards, including credit card data encryption and tokenization, have been implemented to protect these transactions [1].

Machine learning (ML), a subset of artificial intelligence, allows computers to learn from historical data and improve their predictive capabilities without needing explicit programming. In this work, we leverage ML techniques to detect credit card fraud. Credit card fraud refers to unauthorized transactions made using a credit or debit card. The Federal Trade Commission (FTC) reported 1,579 data breaches, affecting 179 million records, with credit card fraud being the most prevalent. Therefore, it is crucial to develop a robust fraud detection system to prevent financial losses [2].

Fraud detection involves monitoring a cardholder's transaction patterns to determine whether a transaction is legitimate. If deemed suspicious, the transaction is flagged as fraudulent. Researchers have explored various approaches for detecting fraudulent credit card transactions by developing models based on artificial intelligence, data mining, fuzzy logic, and machine learning. While detecting credit card fraud remains a challenging task, machine learning has made significant progress in this field. In our proposed system, we use ML techniques to build a fraud detection system. Fraud detection during online transactions relies on analyzing extensive amounts of data, which produces a binary outcome: genuine or fraudulent. Fraudulent datasets include features

like the country of origin of the card, customer age, and account balance. These features, along with hundreds of others, contribute to determining the likelihood of fraud [3].

One of the key challenges in using ML techniques for fraud detection is the difficulty of replicating most results due to the confidentiality of credit card transactions. Consequently, the datasets used to train ML models for fraud detection often contain anonymized features [3].

Building a fraud detection system is more complex than it appears. Practitioners must decide which learning strategy (e.g., supervised or unsupervised learning), algorithms, and features to use, and how to address the class imbalance problem (since fraudulent cases are far less frequent than legitimate ones). In addition to class imbalance, another challenge is the overlap between genuine and fraudulent transactions due to incomplete transaction data, which often causes machine learning algorithms to perform poorly.

In real-world applications, a fraud detection algorithm predicts whether a transaction is genuine or fraudulent, alerting investigators to suspicious activity. However, due to the potential burden this creates, only a small number of transactions are validated by investigators, leading to limited feedback for the system, which in turn may result in less accurate models. Lastly, acquiring real-world financial datasets is extremely difficult, as financial institutions are reluctant to disclose sensitive customer information due to confidentiality concerns—this presents a major challenge for research in this area [4].

The objective of this research is to develop a machine learning algorithm that detects credit card fraud, securing personal data and preventing financial losses. The aim is to demonstrate the effectiveness of machine learning in minimizing fraud. The primary goal of this thesis is to apply machine learning techniques to perform predictive analysis on a dataset of credit card transactions, identifying fraudulent transactions. To tackle the issue of class imbalance, various sampling techniques and machine learning algorithms will be employed [4].

Types of Fraud

1. Bankruptcy Fraud

Bankruptcy fraud occurs when individuals engage in credit-based transactions despite lacking the financial capacity to repay their debts. This fraudulent activity results in significant losses for financial institutions as they are unable to collect the owed amounts. To prevent this, institutions often conduct pre-emptive assessments of creditworthiness, using data from credit bureaus to evaluate an individual's financial history. By identifying patterns that indicate potential insolvency, financial institutions can reduce the risk of extending credit to individuals who are likely to default [5].

2. Theft Fraud/Counterfeit Fraud

Theft fraud refers to the illegal acquisition and use of another person's credit card for unauthorized purchases. Counterfeit fraud, on the other hand, involves the creation and use of fraudulent credit card details, particularly in e-commerce transactions. Both types of fraud lead to financial losses for banks and merchants. To combat this, institutions implement rapid detection and response systems that are triggered when unauthorized transactions are detected. Additionally, stronger authentication protocols and transaction monitoring systems help protect against these types of fraud in digital transactions [6].

3. Application Fraud

Application fraud involves the submission of false or misleading information by individuals attempting to obtain credit cards under fraudulent pretences. This undermines the credit application process and exposes financial institutions to higher risks of default and financial loss. To counteract this, institutions implement thorough verification systems that scrutinize the accuracy of the information provided by applicants. By cross-checking data from various sources and conducting in-depth due diligence, financial institutions can reduce the risk of fraudulent applications, ultimately safeguarding themselves from potential losses. [7]

4. Behavioural Fraud

Behavioural fraud involves the use of valid credit card details to make unauthorized transactions, often in environments where the physical card is not required, such as online or over-the-phone purchases. To mitigate this, financial institutions deploy advanced fraud detection algorithms designed to identify abnormal transaction patterns that may signal fraud. By analysing historical transaction data with machine learning algorithms, they aim to differentiate between legitimate and fraudulent activities. However, the effectiveness of these systems depends on continuously improving

the algorithms to reduce false positives and enhance fraud detection accuracy [8].

This paper is structured as follows: Section 2 covers related work, Section 3 outlines the proposed approach and discusses the findings, and Section 4 presents the conclusion.
finally addressed.

2. PREVIOUS WORK

Several machine learning (ML) algorithms, including logistic regression (LR), decision tree (DT), support vector machine (SVM), and random forest (RF), were employed by the authors of a study to develop a credit card fraud detection system. These classifiers were evaluated using a 2013 dataset of European cardholders, which exhibited a highly imbalanced distribution between legitimate and fraudulent transactions. The researchers assessed the performance of each ML technique based on classification accuracy. The results showed accuracy scores of 97.70% for LR, 95.50% for DT, 97.50% for SVM, and 98.60% for RF. Despite these strong results, the authors suggested that advanced pre-processing methods could further enhance classifier performance [9].

Varmedja et al. proposed a machine learning approach for detecting credit card fraud, using a Kaggle dataset that contains transactions made by European cardholders over two days. To address the class imbalance in the dataset, they applied the Synthetic Minority Oversampling Technique (SMOTE). The researchers tested several ML techniques, including RF, Naive Bayes (NB), and multilayer perceptron (MLP). Experimental results indicated that the RF algorithm performed best, achieving a fraud detection accuracy of 99.96%. The NB and MLP approaches followed closely, with accuracy rates of 99.23% and 99.93%, respectively. The authors recommended further research to develop a feature selection method that could improve the precision of ML models [10].

In another study, Khatri et al. investigated the performance of various ML methods for credit card fraud detection. The techniques considered in their study included DT, k-Nearest Neighbor (KNN), LR, RF, and NB. They utilized an unbalanced dataset collected from European cardholders and evaluated the precision of each method as a key performance metric. The experimental results showed that DT achieved 85.11% precision, KNN 91.11%, LR 87.5%, RF 89.77%, and NB only 6.52% [11].

A related study compared several machine learning classifiers, such as Random Forest, Naive Bayes, XGBoost, and Logistic Regression, in fraud detection. Using metrics like precision, accuracy, F1 score, recall, and MCC, the study addressed the challenge of imbalanced datasets by applying the SMOTE technique. Random Forest was the best performer, with 99.96% accuracy [12].

Another investigation evaluated the performance of six ML models, including Random Forest, SVM, K-Nearest Neighbor, Logistic Regression, Classification and Regression Trees, and XGBoost, on real-world transaction data. Exploratory Data Analysis helped identify key features, such as transaction amount and time, to improve fraud detection accuracy [13].

Additionally, a study compared ML techniques like Decision Trees, Random Forest, K-Nearest Neighbor, and SVM on a highly imbalanced fraud dataset. SVM demonstrated superior performance, highlighting its effectiveness in fraud detection [14].

Moreover, other research explored the use of alternative algorithms, such as SVM, Naive Bayes, and neural networks, for fraud detection. Performance was evaluated using metrics such as accuracy, precision, recall, F-measure, transaction intervention rate, and customer coverage rate [3, 15, 16].
A variety of machine learning (ML) algorithms are used in the evaluated research to detect credit card fraud, each of which makes use of distinct preprocessing methods and performance indicators. Numerous research use techniques like SMOTE to address the problem of class imbalance, which is prevalent in fraud detection datasets. These techniques have been successful in improving performance metrics for models like Random Forest (RF) and Support Vector Machine (SVM). Across experiments, RF regularly delivers great accuracy, sometimes exceeding 99.96%, demonstrating its resilience when dealing with complicated data structures and imbalanced datasets. However, because high accuracy can conceal subpar performance in minority (fraud) class detection, measurements like accuracy alone could not fully reflect a model's effectiveness in unbalanced circumstances.
As a result, research employing a variety of criteria, including as precision, recall, and F1 score, offers a more thorough evaluation of performance. Simpler models like Decision Trees (DT) and Logistic

Regression (LR) may not have the complexity required to adequately detect fraud trends, even with their high accuracy. Also, research that only uses accuracy measurements ignores the precision-recall trade-off that is necessary to reduce false positives and false negatives in fraud detection. Overall, even though sophisticated machine learning algorithms and preprocessing methods increase detection rates, more study into feature selection and the incorporation of different metrics, like the Matthews Correlation Coefficient (MCC), may offer a better comprehension of classifier performance and dependability in practical settings.

3 DATASETS, EXPERIMENTS, AND RESULTS

3.1 Data set

The dataset consists of credit card transactions made by European cardholders over a two-day period in September 2013. Out of 284,807 transactions, only 492 are labeled as fraudulent, reflecting a highly imbalanced dataset, with fraudulent transactions accounting for just 0.172% of the total. This distribution is typical for fraud detection tasks, where anomalies tend to make up a small percentage of overall transactions.

The dataset features financial data that has been anonymized using Principal Component Analysis (PCA), ensuring the confidentiality of the cardholders. Each transaction is represented by numerical values without any contextual information. The dataset contains 28 numerical features, along with an "Amount" column that indicates the transaction amount. The final column, labeled "Class," identifies whether a transaction is fraudulent (1) or legitimate (0).

In the analysis, the initial column representing time in seconds was excluded. The primary objective is to predict the "Class" column by training a model capable of accurately detecting fraudulent transactions based on the numerical features provided.

3.2 Classifiers Evaluation Criteria

This section introduces traditional evaluation methods commonly found in the literature for binary classifications, including precision, recall, accuracy, sensitivity, specificity, and F-measure [18].

3.2.1 Precision:

Precision measures the proportion of correctly predicted positive instances out of all instances classified as positive. It is calculated as:[18]

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \quad \textit{(1)}$$

3.2.2 Recall (Sensitivity)

Recall, also known as sensitivity, indicates the proportion of positive examples correctly identified from all actual positive instances. It is computed as [18]:

$$\boldsymbol{Recall\ (Sensitivity)} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \quad \textbf{(2)}$$

3.2.3Accuracy

Accuracy assesses the overall correctness of the model's predictions by considering both true positive and true negative predictions in relation to the total number of instances. It is defined as [19]:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \quad \textit{(3)}$$

3.2.4 F-measure

The F-measure, also known as the F1 score, is the harmonic mean of precision and recall, providing a single measurement that balances both metrics. It ranges from 0 to 1, with 1 indicating a classifier perfectly capturing precision and recall. The F-measure is calculated using the following [18]:

$$\boldsymbol{F-measure} = \frac{\boldsymbol{2(Precision)(Sensitivity)}}{\boldsymbol{(Precision)+Sensitivity}} \quad \textit{(4)}$$

Where,

TN is True Negative: case was negative and predicted negative,

TP is True Positive: case was positive and predicted positive,

FN is False Negative: case was positive but predicted negative,

FP is False Positive: case was negative but predicted positive.

These evaluation metrics offer comprehensive insights into the performance of binary classification models, enabling researchers to assess their effectiveness in various scenarios

3.3 Fraud Classification

3.3.1 Support vector machine (SVM)

Support Vector Machine (SVM) is a highly utilized and powerful classifier in fraud detection.

Scientifically, SVM works by constructing a separating hyperplane, labeled as H, to distinguish between fraudulent and non-fraudulent transactions. The fundamental concept of SVM is to maximize the margin between two parallel hyperplanes, ensuring that no data points fall within this margin. This margin represents the space where the classifier can confidently separate fraudulent from legitimate transactions.

The primary objective of SVM is to classify instances using a linear feature function, although it can also manage non-linear classification through kernel functions. These kernel functions transform the input data into higher-dimensional spaces, enabling linear separation in cases where it is not feasible in the original space.

SVM operates on labeled data, searching for the hyperplane that maximizes the margin, while selecting specific data points, called support vectors, that define the hyperplane. These support vectors are crucial, as they represent key instances that determine the decision boundary between fraudulent and legitimate transactions.

In fraud detection, SVM seeks to find the optimal hyperplane that separates fraudulent transactions from legitimate ones, with the goal of maximizing the margin between the two classes. As shown in Figure 1, a wider margin improves generalization and reduces the chance of misclassifying transactions, thereby enhancing the fraud detection system's performance.

This study employed three commonly used SVM kernel functions: the polynomial kernel, the linear kernel, and the Gaussian kernel, including the Radial Basis Function (RBF) [20].

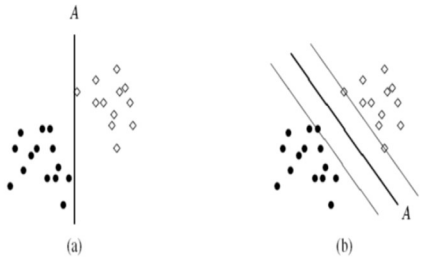

Figure 1: SVM classifier

Table 1 presents a summary of the fraud detection results from the Support Vector Machine (SVM) model. With an accuracy of 0.9992, the SVM demonstrated impressive performance, effectively classifying most transactions. Precision was notably high at 0.94, meaning that a large proportion of transactions flagged as fraudulent were indeed fraudulent. However, the recall score was considerably lower at 0.635, reflecting the percentage of actual fraudulent transactions that were correctly identified. This indicates that while the SVM model excels at identifying legitimate transactions, a significant number of fraudulent cases might go undetected.

The model's overall performance is also reflected in its F1 score of 0.758, which balances precision and recall. The confusion matrix in Figure 2 provides additional insight into the model's behavior. It shows that the SVM model made only a small number of errors by incorrectly classifying non-fraudulent transactions as fraudulent (False Positives). However, it also failed to identify some fraudulent transactions (False Negatives), despite correctly classifying a large portion of them (True Positives).

In conclusion, while the SVM model shows strong potential for fraud detection, there is still room for improvement, particularly in enhancing recall to capture more fraudulent transactions.

Table 1: SVM performance metrics

Accuracy	Precision	Recall	F1 Score
0.9992	0.94	0.635	0.758

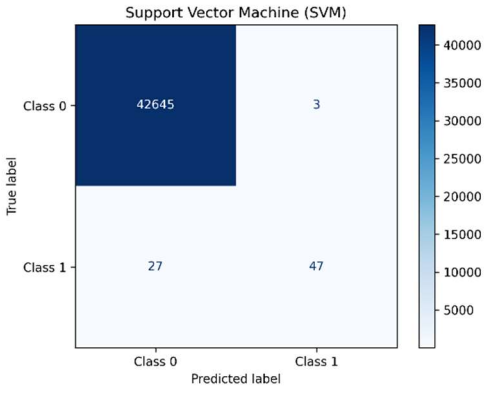

Figure 2: Support vector Machine Confusion Matrix

3.3.2 The K nearest neighbour (KNN)

KNN is a type of learning that is supervised. Using data attributes as a guide, this approach attempts to

classify the set of input patterns (in this instance, photos). A test picture is classified by Knn to the class with the greatest number of nearest K patterns. Thus, in this instance, the majority voting method is employed . Knn does the classification by using a variety of distance functions while maintaining training patterns [21]. Many measures for measuring distance were proposed in [22];

1. City block distance

The distance between two points can be calculated using city block distance if you follow a direction that resembles a grid. The total of the Manhattan distance and the block distance between two things is the same. As 'q' and 'r' are defined in an n-dimensional vector space, city block distance is a vector-based method. The block or L1 distance is calculated by adding the borders. The space left by the city block is below.

$$L_1(q,r) = \sum_y |q(y) - r(y)| \quad (5)$$

Essentially, the distance in the grids is the number of edges that must be crossed to move from point 'q' to point "r." As a result, the point sets are discretely represented in two dimensions.

2. Cosine similarity

By comparing the cosines of the angles of two vectors in an inner product area, one may determine how similar they are. This is known as cosine similarity. The normalization point product of the two texts will be determined by using the attribute as a vector in the Cosine similarity metric. The customer tries to find the cosine of the angle between the two things by determining the cosine resemblance. Because the objects are at a 90-degree angle and the cosines add up to zero, the sample does not exchange attributes. The expression for it in a mathematical equation is as follows.

$$\cos(\theta) = \frac{X.Y}{|X| * |Y|} \quad (6)$$

3. Euclidean distance

The regular distance that can be measured with a ruler between two dots is known as the Euclidean distance. Since the axes on a graph represent the common characteristics for data objects, this is the simplest method for calculating similarities. The data items in the final map are said to be diagrammed in the preferable space.

Finding the relationship between two objects can be done mathematically as follow:

$$E(X,Y) = \sqrt{\sum_{i=0}^{n}(X_i - Y_i)^2}$$ **Error! No text of specified style in document.** (7)

4. Jaccard Coefficient

By calculating the overlap's scale and proportions, two sets can be compared for similarity. For two sets, it is defined as the cardinality of its union divided by the cardinality of its crossroads. Calculated to seem as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (8)$$

5. Minkowski Distance

As can be seen below, the distance from Minkowski is a generalization of the Euclidean distance.

$$dist = \left(\sum_{k=1}^{n} |p_k - q_k|^r\right)^{\frac{1}{r}} \quad (9)$$

In this case, n stands for the dimensions, r for the parameters, k for the kth component, and p and q for the data items.

Table 2 provides specifics on the K-Nearest Neighbors (KNN) model's efficacy in fraud detection, where it produced noteworthy outcomes. With an astonishingly high accuracy of 0.9994, the KNN model demonstrated its capacity to correctly categorize the great majority of transactions. At 0.946, precision—a measure of the percentage of properly recognized fraudulent transactions among all transactions flagged as fraudulent—was again very good. With a recall score of 0.716—a measure of the percentage of successfully identified fraudulent transactions among all actual fraudulent transactions—the KNN model performed admirably. Furthermore, the F1 score—a balanced indicator of recall and precision—was strong at 0.815, highlighting the KNN model's overall efficacy in fraud detection.

The confusion matrix, shown in Figure 3, provides further information about the model's performance. The matrix shows that only a small percentage of transactions were incorrectly categorized as fraudulent (False Positives), with the KNN model properly classifying the vast majority of non-fraudulent transactions (True Negatives). Furthermore, the model correctly classified certain transactions as non-fraudulent (False Negatives) while simultaneously correctly identifying a sizable proportion of fraudulent transactions (True Positives). All things considered, the KNN model shows great promise for fraud detection, especially when it comes to correctly detecting fraudulent transactions while reducing misclassifications.

Table 2: KNN performance metrics

Accuracy	Precision	Recall	F1 Score
0.9994	0.946	0.716	0.815

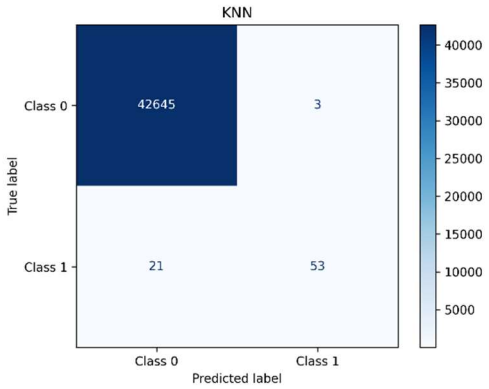

Figure 3: KNN Confusion Matrix

3.3.3 Convolutional Neural Network (1 D)

A Convolutional Neural Network (CNN) is a type of deep neural network commonly employed for analyzing visual data, particularly in tasks like image classification. CNNs excel at learning spatial hierarchies of features, and they have potential applications in fraud detection by extracting relevant patterns from transaction data or other related inputs.

A one-dimensional Convolutional Neural Network (1D CNN) for credit card fraud detection consists of multiple interconnected layers, designed to process transaction data sequentially and systematically. At its core are convolutional layers, responsible for detecting local patterns and features in the transaction sequences. These layers perform element-wise operations, sliding a set of filters over the sequence to capture important patterns. Rectified Linear Units (ReLU) add non-linearity, enabling the network to discover complex relationships in the data. Pooling layers follow, downsampling the feature maps to reduce dimensionality while retaining crucial information.

Next, flattening layers convert the feature maps into one-dimensional vectors, preparing them for input into fully connected (dense) layers, where each neuron connects to all neurons in the following layer. This enables the network to learn high-level representations of the transaction data. For binary classification tasks, the output layer consists of a single neuron with a sigmoid activation function, producing a probability score that indicates the likelihood of fraud. During training, optimization algorithms like Adam or RMSprop adjust the network's parameters iteratively to minimize a loss function (such as binary cross-entropy), improving the model's performance. Collectively, these components enable the 1D CNN to process sequential transaction data and identify critical patterns relevant to accurate credit card fraud detection.

Table 3 presents the performance metrics of the 1D CNN model for fraud detection. The model achieved an impressive accuracy of 0.99948, demonstrating its ability to correctly classify the majority of transactions. Precision was high at 0.933, indicating the percentage of accurately identified fraudulent transactions among all flagged as fraudulent. The recall score, which measures how many actual fraudulent transactions were detected, was 0.7567. The F1 score, a balanced measure of precision and recall, stood at 0.8358, highlighting the model's strong overall performance in fraud detection.

The confusion matrix in Figure 4 further details the model's classification. The matrix shows that the 1D CNN made only a small number of false positive errors, accurately identifying most non-fraudulent transactions (True Negatives). It also successfully flagged a significant number of fraudulent transactions (True Positives), although some fraud cases were misclassified as non-fraudulent (False Negatives). Overall, the 1D CNN model shows significant promise for fraud detection, demonstrating excellent accuracy and the ability to discern fraudulent patterns in credit card transactions.

Table 3: 1 D CNN performance metrics

Accuracy	Precision	Recall	F1 Score
0.99948	0.933	0.7567	0.8358

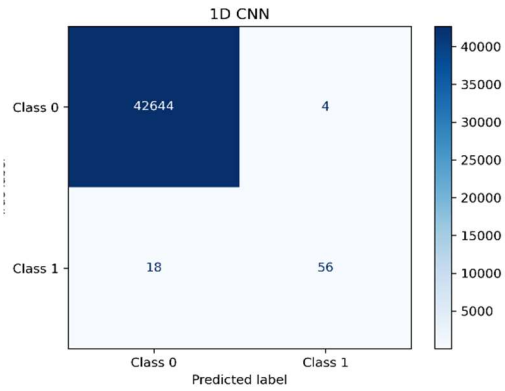

Figure 4: 1D CNN Confusion Matrix

3.3.4 Random Forest

One highly effective method for detecting fraudulent credit card transactions is the Random Forest classifier. This approach involves constructing multiple decision trees on different subsets of the transaction dataset and combining their predictions to arrive at a final outcome. By leveraging this ensemble technique, Random Forest overcomes the limitations of individual decision trees and enhances prediction accuracy, even with complex, large-scale datasets commonly encountered in credit card fraud detection scenarios.

A key advantage of Random Forest in this context is its ability to handle both classification and regression tasks effectively. Its versatility and capacity to process large, multidimensional datasets make it a popular choice for fraud detection. Moreover, Random Forest is resilient to missing or noisy data, ensuring consistent model accuracy and strong generalization performance.

The training phase of the Random Forest algorithm for fraud detection typically involves two steps. First, a set number of decision trees (N) are created, forming a forest. Second, the predictions from each tree are calculated. The ensemble method reduces the risk of overfitting by averaging the predictions across multiple trees. Unlike a single decision tree, Random Forest takes advantage of parallel tree growth, enabling efficient computation even for large datasets, such as those in credit card fraud detection.

Random Forest consists of numerous uncorrelated trees, each trained on different resampled portions of the transaction data. This variation among the trees is crucial for reducing overall model variance, leading to improved robustness and generalizability of the predictions. Through ensemble learning and tree parallelization, the Random Forest model can detect complex patterns and relationships that may indicate fraudulent activity within credit card transactions.

Table 4 summarizes the fraud detection performance of the Random Forest classifier. With an impressive accuracy of 0.999, the model accurately classified most transactions. The precision score was notably high at 0.948, indicating the proportion of correctly identified fraudulent transactions among all transactions flagged as fraudulent. Additionally, the recall score was 0.743, representing the proportion of correctly detected fraudulent transactions out of all actual fraudulent cases. The F1 score, a balanced measure of precision and recall, stood at 0.833, highlighting the model's overall effectiveness in fraud detection.

The confusion matrix, depicted in Figure 5, offers further insights into the model's performance. It shows that only a small number of transactions were mistakenly classified as fraudulent (False Positives), while most non-fraudulent transactions were correctly identified (True Negatives). Additionally, the model correctly identified a significant number of fraudulent transactions (True Positives), though some fraud cases were misclassified as non-fraudulent (False Negatives). Overall, the Random Forest model demonstrates significant promise for fraud detection, exhibiting high accuracy and successfully identifying fraudulent patterns in credit card transaction data.

Table 4: RF performance metrics

Accuracy	Precision	Recall	F1 Score
0.999	0.948	0.743	0.833

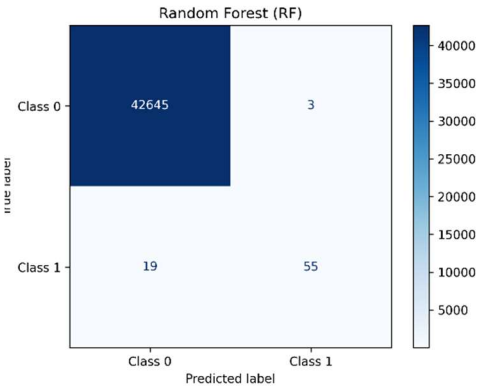

Figure 5: RF Confusion Matrix

3.3.5 Artificial Neural Network (ANN)

An Artificial Neural Network (ANN) is a computational model inspired by the structure and function of biological neural networks. Composed of interconnected nodes (neurons) arranged in layers, ANNs are capable of learning complex patterns from data. In fraud detection, ANNs can be trained on historical transaction data to recognize behaviors indicative of fraudulent activity.

Due to their ability to uncover intricate patterns within vast and complex datasets, ANNs have become highly effective tools for credit card fraud detection. A typical ANN for this purpose consists

of multiple layers of neurons, each performing a specific function on the input data. Through a process of forward propagation and backpropagation, ANNs iteratively adjust their parameters to minimize prediction errors, allowing them to detect subtle fraudulent patterns in otherwise legitimate transactions. By analyzing features such as transaction amount, time, location, and consumer behavior, ANNs can effectively identify anomalous activity that may suggest fraud.

Table 5 highlights the performance of the ANN model in fraud detection, illustrating its ability to accurately classify credit card transactions. With an exceptional accuracy score of 0.9994, the ANN proved highly effective in classifying most transactions correctly. Precision was also impressive at 0.93, indicating the percentage of correctly identified fraudulent transactions among all transactions flagged as fraudulent. The recall score, which measures the proportion of accurately detected fraudulent transactions among all actual fraud cases, was notable at 0.729. With an F1 score of 0.818—a balanced measure of precision and recall—the ANN model demonstrated strong overall performance in fraud detection.

The confusion matrix, shown in Figure 6, provides further insights into the model's results. It reveals that only a small percentage of non-fraudulent transactions (True Negatives) were incorrectly classified as fraudulent (False Positives) by the ANN model. Additionally, the model correctly identified many fraudulent transactions (True Positives) while misclassifying some fraud cases as non-fraudulent (False Negatives). Overall, the ANN model shows considerable promise for fraud detection, offering excellent accuracy and successfully identifying fraudulent patterns in credit card transaction data.

Table 5: ANN performance metrics

Accuracy	Precision	Recall	F1 Score
0.9994	0.93	0.729	0.818

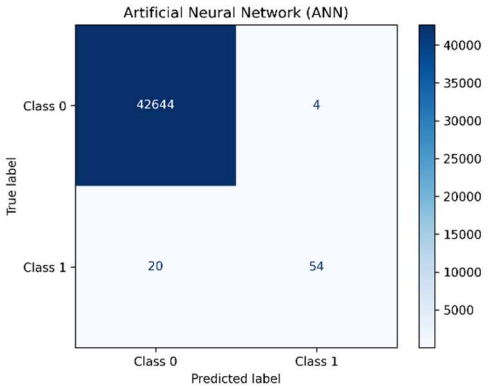

Figure 6: ANN Confusion Matrix

3.3.6 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a statistical technique used for both dimensionality reduction and classification. LDA aims to find linear combinations of features that maximize the separation between two or more classes of data. It is particularly useful when the data classes are well-separated and follow a normal distribution. In the context of fraud detection, LDA can help differentiate between fraudulent and non-fraudulent transactions by projecting the data into a lower-dimensional space that enhances class separability.

Table 6 summarizes the performance of the LDA model in fraud detection, showcasing its effectiveness in classifying credit card transactions. The model achieved an impressive accuracy score of 0.999, demonstrating its ability to accurately classify the vast majority of transactions. Precision, which measures the percentage of correctly identified fraudulent transactions out of all transactions marked as fraudulent, was high at 0.86. Additionally, the recall score was 0.77, reflecting the proportion of correctly identified fraudulent transactions among all actual fraudulent cases. With an F1 score of 0.814—an indicator that balances both precision and recall—the LDA model demonstrated strong overall performance in fraud detection.

The confusion matrix, shown in Figure 6, provides further details on the model's classification results. It reveals that only a small percentage of transactions were incorrectly labelled as fraudulent (False Positives), while most non-fraudulent transactions (True Negatives) were correctly classified. Similarly, the model accurately identified many fraudulent transactions (True Positives) while misclassifying a few fraud cases as non-fraudulent (False Negatives). Overall, the LDA model shows

significant potential for fraud detection, offering excellent accuracy and successfully distinguishing fraudulent patterns within credit card transaction data.

Table 6: LDA performance metrics

Accuracy	Precision	Recall	F1 Score
0.999	0.86	0.77	0.814

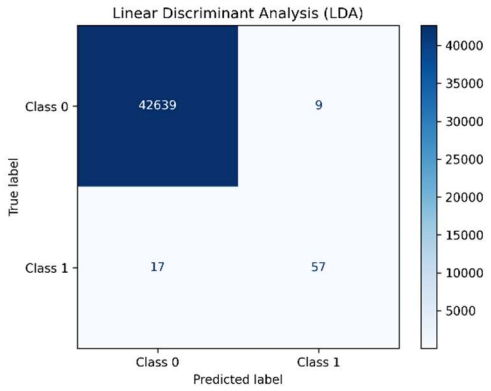

Figure 7: LDA Confusion Matrix

3.3.7 Proposed Model

The proposed approach leverages adversarial-guided oversampling, incorporating a robust double-check mechanism to ensure that generated samples authentically belong to the minority class. This method is driven by rules derived from the minority class, which are learned by training a decision tree on the dataset. After generating samples based on these rules, a well-trained neural network performs a secondary validation to confirm that the new samples indeed represent the minority class.

As depicted in Figure 8, the process is broken down into three key stages: training, generating, and testing, each described in detail below:

1. Training:

In the first stage, two classifiers are trained on the imbalanced dataset: a decision tree and a neural network. The decision tree is specifically trained to extract classification rules for the minority class. These rules will later guide the generation of synthetic samples. Simultaneously, the neural network is trained to classify both minority and majority class data accurately. The neural network will be used during the testing phase to verify the authenticity of generated samples.

2. Generating:

The second stage focuses on generating new samples. This process involves iterating through the attributes of the dataset and generating new values for each attribute. These values are constrained by the upper and lower bounds derived from the minority class rules extracted in the first step. The decision tree, as an explainable classifier, plays a pivotal role in this stage by guiding the generation of samples that strictly adhere to the rules it identified. This ensures that the generated samples align with the minority class characteristics.

3. Testing:

The final stage, known as the testing phase, involves evaluating the generated samples using the trained neural network. The neural network assesses whether each sample belongs to the minority class. Samples that pass this test are retained, while those that do not are discarded. In this context, the neural network functions as an unexplainable classifier, relying on the optimal configuration of neuron weights to classify the data accurately. This dual-classifier system—using both the explainable decision tree for generation and the unexplainable neural network for validation—ensures that the minority class samples generated undergo rigorous validation, increasing the reliability of the synthetic data.

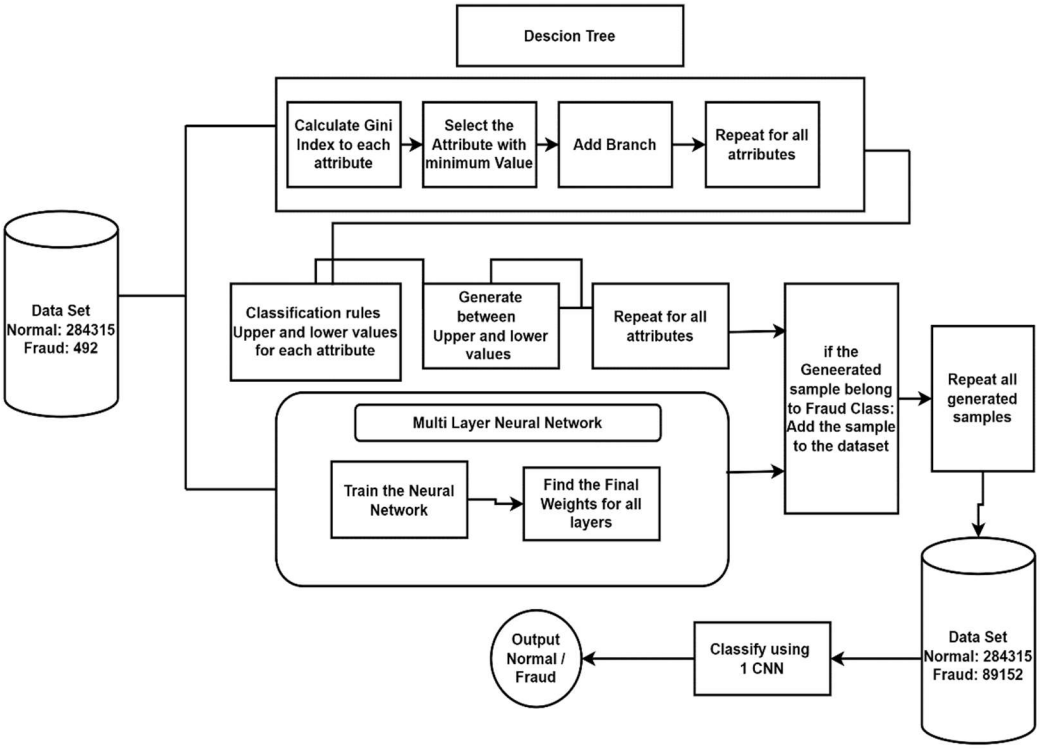

Figure 8: Proposed System Block diagram

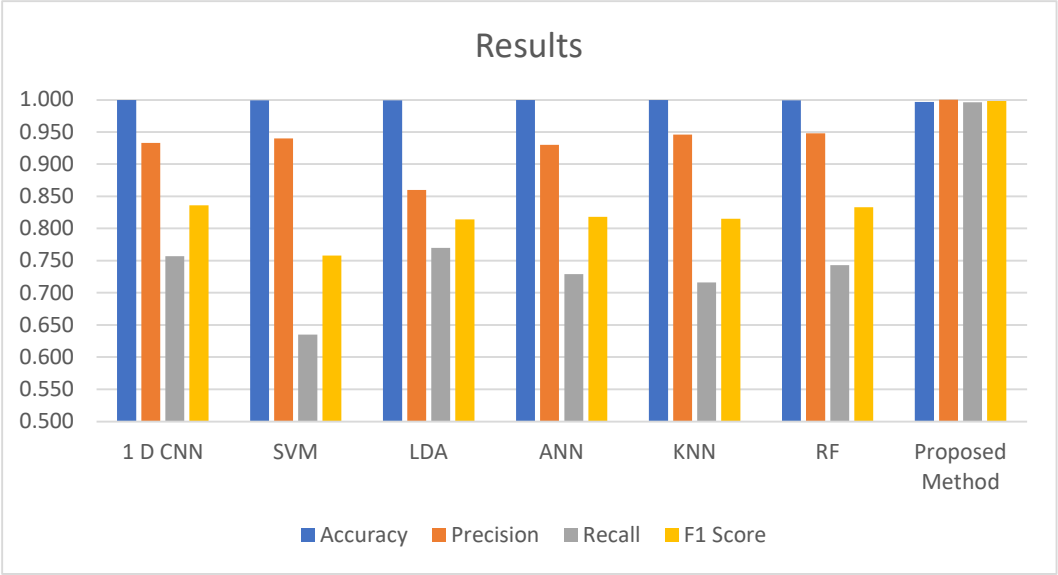

Figure 9: Results Comparison

The new adversarial directed oversampling technique was used by the authors to create synthetic samples, and then they used a 1D Convolutional Neural Network (1D CNN) to categorize the freshly created data. The 1D CNN model was chosen due to its excellent performance in earlier trials and its ability to identify sequential patterns in the transaction data.

Table 7: Proposed System performance metrics

Accuracy	Precision	Recall	F1 Score
0.9968	0.9999	0.9959	0.9979

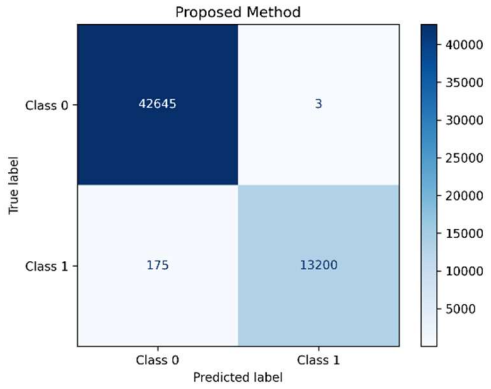

Figure 10: Proposed Method confusion matrix

In the final phase of the fraud detection process, a trained 1D Convolutional Neural Network (1D CNN) classifier was employed to evaluate the synthetic samples generated to represent the minority class (fraudulent transactions). The primary objective of the 1D CNN was to verify the authenticity of these samples, determining if they exhibited patterns commonly associated with fraudulent activity. The model received each synthetic sample and analyzed the sequential properties of the transaction data, focusing on temporal attributes such as time, amount, and transaction type.

During this evaluation, the 1D CNN utilized its architecture and learned parameters to extract relevant features and subtle patterns from the synthetic data. By examining the temporal sequences of transaction attributes, the model was able to detect minute irregularities that are often indicative of fraudulent behavior.

After processing each sample, the 1D CNN assigned a probability score reflecting the likelihood that the sample belonged to the minority class (fraudulent transactions). Samples with high probability scores were retained as valid synthetic samples, while those with lower scores were discarded.

Table 7 presents the exceptional performance of the proposed fraud detection system, which integrates a 1D CNN classifier with an innovative adversarial-guided oversampling technique. The system achieved an impressive accuracy of 0.9968, demonstrating its ability to correctly classify most credit card transactions. The precision score of 0.9999 highlights a remarkably low false positive rate, indicating the system's ability to accurately detect fraudulent transactions without misclassifying legitimate ones. With a recall score of 0.9959, the system showed strong capability in identifying actual fraudulent transactions. Furthermore, the F1 score of 0.9979, which balances precision and recall, underscores the system's overall effectiveness in fraud detection.

The confusion matrix in Figure 7 further illustrates the system's performance. It shows that the model successfully identifies a large portion of fraudulent transactions (True Positives) while accurately classifying most non-fraudulent transactions (True Negatives). These results highlight the proposed method's high precision and accuracy, making it a robust solution for detecting fraudulent credit card transactions.

Figure 8 compares the performance of the proposed method against several other fraud detection techniques, including 1D CNN, Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Artificial Neural Network (ANN), K-Nearest Neighbours (KNN), and Random Forest (RF). Each technique is evaluated based on its F1 score, recall, accuracy, and precision. The proposed method stands out with an exceptional precision score of 1.000, signifying a minimal false positive rate, alongside strong performance across all other metrics. With an F1 score of 0.998, the method demonstrates an optimal balance between precision and recall, outperforming competing methods, which also show high accuracy and precision.

This approach addresses the challenges posed by imbalanced datasets and captures complex fraudulent patterns by combining the power of a 1D CNN with an adversarial-guided oversampling technique. Its scalability and efficiency enable the real-time processing of large volumes of transactions, allowing for swift fraud detection and prevention. Overall, the proposed system represents a promising advancement in the field of credit card fraud detection, offering enhanced capabilities for identifying fraudulent activities in a scientific and practical context.

3. CONCLUSIONS

In conclusion, the proposed approach in this work represents a significant advancement in the field of credit card fraud detection. One of the key challenges in this domain—the imbalance between fraudulent and legitimate transactions—has been effectively addressed through the integration of machine learning techniques and adversarial-guided oversampling. By utilizing decision trees and neural networks to develop classifiers specifically tailored for the minority class, the system generates synthetic samples that closely mimic real fraudulent transactions. These generated samples are then rigorously validated by a trained neural network to ensure their authenticity and relevance to the minority class.

The adversarial-guided oversampling protocol begins by training a decision tree on an imbalanced dataset to extract rules defining the minority class. Simultaneously, a neural network is trained for accurate classification of both classes. During generation, new samples are created for each attribute, constrained by the decision tree's minority-class rules. In testing, the neural network validates the generated samples, retaining only those that authentically represent the minority class. This dual-step validation, combining explainable rule-based generation and neural network verification, ensures high-quality synthetic samples that bolster the minority class.

The results demonstrate the effectiveness of this approach, with an accuracy of 0.9968 and precision, recall, and F1 scores all exceeding 0.995. These metrics underscore the system's ability to detect fraudulent transactions with high reliability while significantly reducing false positives, representing a notable improvement over traditional machine learning method.

This research introduces a method that not only tackles dataset imbalance but also leverages advanced sampling techniques to provide a robust solution for credit card fraud detection. Furthermore, its scalability and efficiency allow for the processing of large transaction volumes in real-time, enabling rapid identification and prevention of fraudulent activities.

In summary, this work makes a substantial contribution to enhancing fraud detection capabilities in credit card transactions. By combining cutting-edge machine learning algorithms with innovative sampling strategies, it presents a reliable and powerful system that has the potential to revolutionize how the financial industry approaches fraud detection.

REFERENCES

[1]. Divadari, S., J. Surya Prasad, and P. Honnavalli. *Managing data protection and privacy on cloud*. in *Proceedings of 3rd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2022*. 2023. Springer.

[2]. Zambrano, D.A., *How litigation imports foreign regulation.* Virginia Law Review, 2021. **107**(6): p. 1165-1254.

[3]. Ileberi, E., Y. Sun, and Z. Wang, *A machine learning based credit card fraud detection using the GA algorithm for feature selection.* Journal of Big Data, 2022. **9**(1): p. 24.

[4]. Trivedi, N.K., et al., *An efficient credit card fraud detection model based on machine learning methods.* International Journal of Advanced Science and Technology, 2020. **29**(5): p. 3414-3424.

[5]. Yadav, R., A. Patil, and R. Sengupta, *An analysis of Satyam case using bankruptcy and fraud detection models.* SocioEconomic Challenges, 2023. **7**(4): p. 24-35.

[6]. Baesens, B., *Fraud analytics: a research.* Journal of Chinese Economic and Business Studies, 2023. **21**(1): p. 137-141.

[7]. Cheliatsidou, A., et al., *The international fraud triangle.* Journal of Money Laundering Control, 2023. **26**(1): p. 106-132.

[8]. Chatterjee, P., D. Das, and D.B. Rawat, *Digital twin for credit card fraud detection: Opportunities, challenges, and fraud detection advancements.* Future Generation Computer Systems, 2024.

[9]. Almarshad, F.A., G.A. Gashgari, and A.I. Alzahrani, *Generative Adversarial Networks-Based Novel Approach for Fraud Detection for the European Cardholders 2013 Dataset.* IEEE Access, 2023.

[10]. Rai, A.K. and R.K. Dwivedi. *Fraud detection in credit card data using machine learning techniques*. in *Machine Learning, Image Processing, Network Security and Data Sciences: Second International Conference, MIND 2020, Silchar, India, July 30-31, 2020, Proceedings, Part II 2*. 2020. Springer.

[11]. Khan, M.Z., et al., *The Performance Analysis of Machine Learning Algorithms for Credit Card Fraud Detection.* Int. J. Online Biomed. Eng., 2023. **19**(3): p. 82-98.

[12]. Faraji, Z., *A review of machine learning applications for credit card fraud detection with a case study.* SEISENSE Journal of Management, 2022. **5**(1): p. 49-59.
[13]. Ghai, V. and S.S. Kang. *Credit card transaction data analysis and performance evaluation of machine learning algorithms for credit card fraud detection*. in *AIP Conference Proceedings*. 2022. AIP Publishing.
[14]. Hazim, L.R., *Four classification methods Naïve Bayesian, support vector machine, K-nearest neighbors and random forest are tested for credit card fraud detection*. 2018, Fen Bilimleri Enstitüsü.
[15]. Mahmud, M.S., P. Meesad, and S. Sodsee. *An evaluation of computational intelligence in credit card fraud detection*. in *2016 International Computer Science and Engineering Conference (ICSEC)*. 2016. IEEE.
[16]. Mytnyk, B., et al., *Application of artificial intelligence for fraudulent banking operations recognition.* Big Data and Cognitive Computing, 2023. **7**(2): p. 93.
[17]. Dhiman, S. and R. Bhatt, *Credit Card Fraud Detection.* 2022.
[18]. Muntean, M. and F.-D. Militaru. *Metrics for evaluating classification algorithms*. in *Education, Research and Business Technologies: Proceedings of 21st International Conference on Informatics in Economy (IE 2022)*. 2023. Springer.
[19]. Aftab, A., et al., *Fraud Detection of Credit Cards Using Supervised Machine Learning.* Pakistan Journal of Emerging Science and Technologies (PJEST, 2023. **4**(3).
[20]. Kannagi, A., et al., *Intelligent mechanical systems and its applications on online fraud detection analysis using pattern recognition K-nearest neighbor algorithm for cloud security applications.* Materials Today: Proceedings, 2023. **81**: p. 745-749.
[21]. Abu Alfeilat, H.A., et al., *Effects of distance measure choice on k-nearest neighbor classifier performance: a review.* Big data, 2019. **7**(4): p. 221-248.
[22]. Fu, K., et al. *Credit card fraud detection using convolutional neural networks*. in *Neural Information Processing: 23rd International Conference, ICONIP 2016, Kyoto, Japan, October 16–21, 2016, Proceedings, Part III 23*. 2016. Springer.
[23]. Aburbeian, A.M. and H.I. Ashqar. *Credit card fraud detection using enhanced random forest classifier for imbalanced data*. in *International Conference on Advances in Computing Research*. 2023. Springer.
[24]. Raphael, B.A., B.G. Adashu, and A.I. Wreford, *Card fraud detection using artificial neural network and multilayer perception algorithm.* International Journal of Algorithms Design and Analysis Review, 2023. **1**(1): p. 21-30.
[25]. Chung, J. and K. Lee, *Credit Card Fraud Detection: An Improved Strategy for High Recall Using KNN, LDA, and Linear Regression.* Sensors, 2023. **23**(18): p. 7788.