InfoTech Spectrum: Iraqi Journal of Data Science (IJDS)
ISSN: 3007-5467, Vol. 3 (No.1), 14-28 (2026)
DOI: https://doi.org/10.51173/ijds.v3i1.50
https://ijds.mtu.edu.iq

# A PROPOSED MODEL CREDIT CARD FRAUD DETECTION MODEL USING MACHINE LEARNING TECHNIQUE

**Harith Safwan Ezzulddin** [1*]

[1] *Department of Computer Sciences, University of Altinbas, Istanbul, Turkey*

** Corresponding author E-mail: Hareth.safwan@gmail.com (Harith Safwan Ezzulddin)*

RESEARCH ARTICLE

**ARTICLE INFORMATION**

***SUBMISSION HISTORY:***
*Received: 25 August 2025*
*Revised: 20 October 2025*
*Accepted: 11 November 2025*
*Published: 30 January 2026*

***KEYWORDS:***
*Credit Card Fraud Detection;*
*Logistic Regression;*
*Random Forest;*
*ATM Fraud Activity Detection;*

**ABSTRACT**

The online payment system is at high risk due to the increasing rates of credit card theft. The primary objective is to identify cases of credit card theft by analyzing the purchase history of cardholders and categorizing them accordingly. These include an increase in the slope of logistics, steep slope, and scattered woodlands. The proposed model utilizes tools such as logistic regression and random forest as machine learning techniques. Additionally, a set of preprocessing techniques is employed, including data balancing using SMOTE. After being trained on a large dataset of credit card transactions, the model is used to detect trends and anomalies that may indicate fraudulent activity, taking into account factors such as transaction amount, location, and time of day. We have used artificial minority oversampling to put the data set into proper perspective. The two algorithms were applied, yielding 97.34% accuracy for Logistic Regression and 99.99% accuracy for Random Forest. The accuracy metric is used for performance evaluation. The results indicate a promising performance that can enhance credit card security, potentially helping to reduce financial losses to victims of fraud.

## 1. INTRODUCTION

Credit cards are plastic tools that contain the name, address, and telephone number of a customer, and sometimes a photograph or signature of the customer to facilitate identification. It is deducted frequently from the account holder's account to cover the cost of goods and services. Modern card readers are ubiquitous and can be found in ATMs, point-of-sale devices, retail terminals, online transactions, and banking institutions [1]. The security of the card depends on the physical integrity and confidentiality of the cardholder's financial information. Credit card companies typically focus their marketing activities on the cardholders. It allows the consumer to pay later, on a later date, and pay again in the next invoice [2]. Different types of devices are used to collect card data, including point-of-sale (POS) terminals and automated teller machines (ATMs). Online and offline purchases are made through the use of credit cards, a common practice that occurs daily. They allow for paying without the use of cash, both in-store and online, as well as a buy now, pay later option. As the use of credit cards continues to rise, there has been a corresponding increase in the rate of fraud associated with these methods [3]. Before applying for a credit card, ensure you understand all its features and their functions, Fig. 1.


**Figure 1.** Credit card symbols and numbers.


DOI: https://doi.org/10.51173/ijds.v3i1.50

Credit card fraud is characterized as illicit financial activity conducted by an individual who is not the legitimate account holder. The credit card fraud detection technology analyses every card transaction to detect fraudulent activities amidst thousands of legitimate purchases [4], [5]. The fraud occurs in many areas, including financial institutions, commercial systems and their computerized applications, insurance systems, and healthcare systems. Contemporary payment technology has led to an increase in online purchases. In 2018, London had an estimated loss of 844.8 million USD attributable to credit card theft. To mitigate such losses, fraud should either be avoided or identified. Many algorithms are used in detecting fraud, and the most effective is the artificial neural network. Fraudsters and fraudulent businesses must continually evolve to circumvent modern innovations in security systems. Fraud prediction and detection must be conducted using accurate methods. An effective approach should be able to detect new fraud before it is carried out [6]. To identify instances of credit card fraud, it is necessary to analyze every transaction and detect any abnormal activity. It is a significant challenge to identify identity theft among numerous genuine transactions. To develop a system that detects fraud, it is necessary to understand the rates at which individuals commit fraud. The primary obstacle is to create an accurate and dependable methodology [7].

In most cases, identity thieves obtain the name of the cardholder, as well as their mailing address. With the information bar, duplication of the same would require the cardholder to know or cooperate with the bank staff [8]. Internet transactions only require the card number; therefore, the burglar does not need the physical card or the cardholder's consent. The fraudster is making great efforts to complete all the dealings before the irregularities are realized. Through various tools and techniques, fraudsters can access important and sensitive information, posing a potential threat, as shown in Fig. 2.


**Figure 2.** Common potential threats [9]

Below are some of the credit card frauds:

- Skimming: This method retrieves the credit card's magnetic strip information by inserting the card into a machine.
- Phishing: Email "traps" pose as legitimate financial institutions and request sensitive card information.
- Card Not Present (CNP) Fraud: This method allows you to make purchases over the phone or online without physically presenting your card.
- Account Takeover: This occurs when an unauthorized individual gains access to private information and then uses it for their own benefit.
- Cards captured during shipment: This occurs when a freshly issued credit card falls into the wrong hands.
- Losing the card: Misplaced cards could put private information at risk if they fall into the wrong hands.
- Fake Websites: It can entice visitors to buy things online after they have established trust in the site's legitimacy and provided their credit card information.

Theft of credit card information is a real threat. These avenues do not involve the systems used to complete credit card transactions. A robust Fraud Prevention System (FPS) would be able to maintain control through the use of proactive network security strategies, such as security devices and firewalls, as well as non-computational strategies like social awareness. However, prevention is not an ultimate guarantee of security since there are several levels of threats [10]. State that the second line of defense is the timely detection of these machinations. In the past, a fraud detection system (FDS) relied on human auditors to review a sample of transactions to identify the presence of fraud. It is impossible to develop and make the system effective. To improve these performance measures, algorithms were used to generate automated FDSs. This system shall aim to examine every transaction in the event of fraud, regardless of the controls in place. The initial automated Financial Decision Systems was based on stipulated basic criteria set by financial gurus. They demonstrated an extremely fast detection time when applied to historical data. The need to detect credit card fraud operations within a very short time has increased with the number of credit card transactions; otherwise, it would result in financial harm to the cardholder, the card issuer, and the merchant [11].

An advanced calculation system is used to identify suspicious behavior. The level of success of anti-fraud measures of a system is directly dependent on the number of criteria, the method of classifying them chosen and the type of data input. Consequently, it is impossible to distinguish between the genuine and dubious aspects of a transaction. The likelihood of either can be calculated through a thorough analysis of previous transactions and fraudulent transactions, as well as the patterns they follow. Various methods exist for identifying fraudulent activities, each with its advantages and applications. There can be no successful fraud detection strategy without all three of these elements:

- The accuracy of identifying fraudulent transactions should be high, and incorrect ratings should be minimal.
- No legitimate business dealings should ever be labelled as fraudulent.
- Quick response times and rapid device output are also essential.

However, remember that, with the advancement of technology, methods of identifying fraud are also enhanced. When this trend is maintained, one day, cases of fraud will be detected at an extremely sophisticated level. Several methods have evolved, including Machine Learning, Soft Computing, Artificial Intelligence, and Data Mining [12].

Credit card fraud is a serious issue that causes significant losses for financial institutions and individuals. The growing trend of credit card usage on the Internet, particularly in physical stores, is a clear indication of the demand for effective systems to detect fraudulent transactions in real-time. The main purpose of this paper is to develop a machine learning model that can accurately predict fraudulent and genuine transactions among a large number of credit card transactions. The models also have the capability of identifying outliers or exceptions in the data, which is crucial for predicting whether a transaction may be authentic. If overall we have way more true transactions than there are phones, our model should handle skewed data properly, regardless of the distribution of positives and negatives within the week. Motivation and contributions: This work has clear motivations. - Decrease costs over credit card fraud in dollars or reputation (and stimulate them) to mitigate? -Predict a potential adoption (or at least make the appearance of what to adopt) of practical measures which consequently increase trustworthiness on credit card transactions. Among their predominant causes, the following ones need to be highlighted: hackers' attacks are becoming increasingly frequent and employing very advanced techniques; even if not sophisticated, they are based on statistical observations.

## 2. LITERATURE REVIEW

Backpropagation and artificial neural networks were employed to identify fraudulent credit card transactions. Stored data includes purchase times, purchase IDs, and client names. It used 20% of the samples as the test set and the remaining 80% as the training set [13]. The method developed for real-time data fraud detection achieved a remarkable detection accuracy of 99.96% in extensive testing [14]. Seven hybrid machine learning models were recommended to detect fraudulent

activity, including Logistic Regression, Decision Tree, Support Vector Machine, Naïve Bayes, and Extreme Gradient Boosting. Since conventional machine learning methods have their limitations in fraud prediction for credit cards, many hybrid approaches that leverage modern machine learning technologies are employed. Adaboost with LGBM yields a very good score of 0.82. However, it is worth noting that aggregating different machine learning models does not always result in a performance gain. The best model that the data can use to explore this question is the Adaboost + LGBM hybrid, with an accuracy of 0.97, as it employs various methods, including Naïve Bayes, logistic Regression, multilayer perceptron, and radial basis Function Networks [15].

A variety of fraudulent credit card transaction detection methods is evaluated using data from European credit cardholders in the Kaggle repository. The results revealed that Random Forest (RF) outperformed other typical choices. Since Personal data is volatile, studying and developing algorithms with real-life data are challenging. Use the LR, KNN, and NB classifiers to predict repeat offenders for credit card fraud. The information pertains to approximately 284,807 individual transactions made by cards across Europe. All approaches were applied to all types, whether processed or raw data. From experiments, we realized that three classifiers, Naïve Bayes (97.92%), K-Nearest Neighbors (97.69%), and Logistic Regression (54.8%), all performed well for the same model decisions [16].

A new method was designed to enhance the effectiveness of the SVM algorithm in card recognition. The success of the SVM method greatly depends on input parameters and training data. A combination of ensemble methods and the Least Squares Support Vector Machine (LS-SVM) has been applied to forecast cards with monthly delayed payments [17]. The process was used to analyze financial data collected from the UCI Taiwanese dataset. The Logistic Regression (LR), Naïve Bayes (NB), and the eleventh version of K-Nearest Neighbors (KNN) are examples of machine learning approaches. The resultant number is the outcome of the two databases. The credit card recognition algorithm is based on data obtained from the European Kaggle repository. The sample consisted of 284,807 cases. Compared to Naïve Bayes classifiers, the accuracy of the Logistic Regression classifiers was 97.09%.

The authors' work was largely focused on intrusion detection in cybersecurity [18]. They were trained on a very sensitive data set. The classifications of this database could be beneficial to hackers. Their skills in detecting cyberattacks were enhanced with the use of a random forest technique. Scholars have utilized machine learning algorithms, including Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF), to combat fraud. The paper utilized records of credit card transactions compiled by financial institutions across Europe. The dataset is a Kaggle dataset comprising 284,808 transactions of credit cards, encompassing a diverse range of both online and in-store purchases. The accuracy of the LR classifier was 90%, while the DT classifier achieved 94.33% and the RF classifier achieved 95.53%. Compared to LR and DT, the RF classifier is better.

## 3. METHODOLOGY

### 3.1. Machine Learning

The term "Machine Learning" was first coined in 1959 by Arthur Samuel. Pattern classification was the main machine learning application in the 1960s [16]. Machine learning (ML) is a subfield of computer science and artificial intelligence that focuses on creating models and algorithms enabling computers to learn autonomously or make predictions or decisions without human intervention. The activities of machine learning encompass the preparation and collection of data, the selection and application of techniques, and the development of models that can be used for forecasting or decision-making. Most of the attention is concentrated on a given task. Estimates, forecasts, recommendations, and similar concepts can be essential to this work or issue. The level of an individual's experience determines their ability to foresee and solve future problems by applying the lessons learned from past experiences. The performance of a machine determines its ability to respond to issues or tasks related to machine learning and produce the best results [18]. Machine learning encompasses both supervised and unsupervised learning principles. Such tactics employ mathematical and statistical methods to extract information from vast amounts of data. Some of the key machine learning concepts include feature engineering, data preparation, validation,

regularization, and hyperparameter optimization. Machine learning (ML) Involves Supervised learning, where labels or datasets are used to train algorithms to classify data accurately or predict results. The primary objective is to develop a model that can accurately forecast output based on the given input data. The acquired information may be used for testing or training purposes. The training data can also be used to build a function that can then be applied to the test data to perform classification and prediction. Classification in machine learning primarily aims to determine the class label of an input sample given a set of training samples whose labels are known. The classification aim is to have a decision boundary that defines the classes of the input feature space [19]. Many fields rely on classification, including the social sciences, marketing, medicine, and finance. Some tasks that involve classification include sentiment analysis, customer churn prediction, credit risk assessment, and disease diagnosis [19].

### 3.2. Artificial Intelligence (AI)

Research on artificial intelligence aims to develop systems and algorithms that can simulate human thought processes. The sphere of artificial intelligence encompasses computers that can think logically, as well as those that can think and act like humans by using language. The primary aim of artificial intelligence studies is to develop a machine that can perform cognitive and decision-making processes commonly associated with human beings. Even though AI is an effective tool for analyzing data and making corporate decisions, it will never completely replace people. Artificial intelligence refers to highly automated systems that can reason and respond to environmental stimuli independently. One such case is a robot that can sense its surroundings and execute motion instructions. Data analysis and scenario analysis are utilized in various fields to inform decisions. In both cases, artificial intelligence is discovering new applications due to advancements in science and technology. Intelligence is defined across numerous distinct categories. Linguistic intelligence pertains to the utilization of language, encompassing proficiency in multiple languages and comprehension of linguistic symbols. Social intelligence involves navigating complex social scenarios. General intelligence relates to cognitive capabilities and problem-solving skills. Broad intelligence refers to the ability to maintain one's surroundings in an orderly manner. Intelligence is characterized by the ability to acquire new information in response to environmental stimuli, innovative problem-solving, and creative thinking.

### 3.3. Random Forest.

To circumvent the drawbacks of a single decision tree, random forest employs a training dataset and selects features at random from each tree [8]. Credit card fraud committed both online and in person has been uncovered using this technique. Random forest had good precision. Data inconsistency is a significant issue for classifiers that have been examined and addressed from various perspectives. Trifecta of problems: oversampling, under sampling, and artificially created minorities. A System of excessive sampling: if we discard papers that could help our classifier build an accurate model, we might lose a lot of useful information. Interpolating between existing instances of the same class is proposed as a method for generating new outliers. These methods have also been successfully applied to the fight against credit card fraud; Combining random forest (RF) and rough set theory (RST) is effective in fraud detection. To improve classification results, RF can be used to rank attributes in order of importance. It is a powerful ensemble learning method for random forest classification, Regression, and other machine learning tasks. It works by combining multiple decision trees, with each tree being trained on a randomly selected subset of features and data samples [20].

### 3.4. Logistic Regression

It is a popular statistical tool for scenarios that require a binary yes-or-no response. To ascertain the likelihood of a binary response variable contingent upon one or more predictor variables, logistic Regression, a generalized linear model, can be employed:

- A linear combination of predictor variables is computed, weighted by a set of coefficients.
- The linear combination is transformed using a logistic function to produce a predicted probability of the binary response variable.

- The predicted probabilities are thresholder to make binary predictions.

The logistic function is defined as.[21].

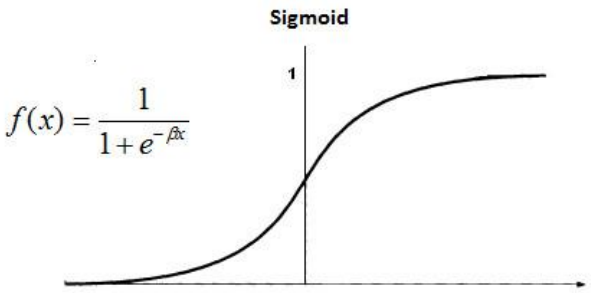


**Figure 3.** Logistic Regression: Sigmoid Function

Where z is the linear function of predictor variables and coefficients, the logistic regression structure [22] is presented in Fig. 4.

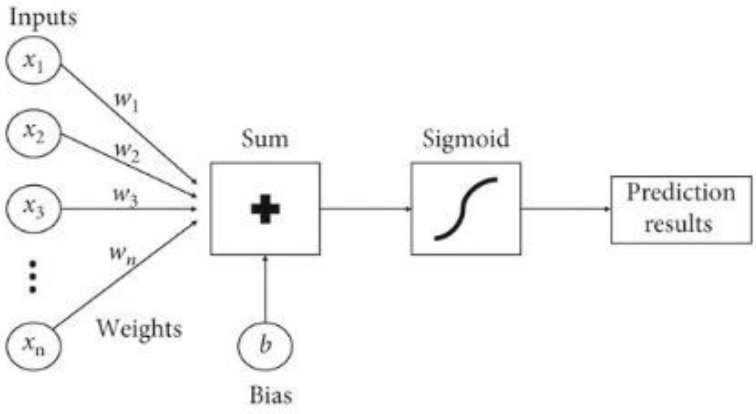


**Figure 4.** Logistic regression structure.

Logistic Regression is constructed by finding coefficients that reduce the distance between the model-generated probabilities and the true binary response values in the training data. Common methods for accomplishing this include maximum likelihood estimation and gradient descent optimization [23]. Various applications of logistic Regression are found in medicine, finance, and social sciences. It is widely used in risk analysis, decision-making and predictive modelling [24], [25].

## 4. RESEARCH DESIGN

To protect credit card users against financial losses and maintain financial system stability, it is crucial to detect credit card fraud promptly and effectively. We offer a sophisticated system based on Python and a variety of methods, including Logistic Regression (LR) and Random Forest (RF), to identify fraudulent credit card transactions. The proposed system comprises four main functions: data collection, data processing, algorithmic classification, and transaction fraud detection.

When credit card transactions are being processed, several attributes are recorded, including amount, date, merchant, and user information. Data preparation includes imputing missing data and sanitizing the data after it was collected. We have analyzed the format of the data, and it is ready to be processed further. The support vector machine, random forest, logistic Regression, and other machine learning methods are used to classify. Their skill in recognizing patterns and exceptions in large and potentially complex data sets influenced their choice. All the algorithms are trained to detect trends in processed data, including valid and suspicious activity. One way to determine whether a transaction is legitimate is through the use of trained models to classify it as either

fraudulent or not. To categorize the transaction features, we contrast them with the patterns discovered during training. The suggested strategy uses a rigorous evaluation of the transaction to determine the possibility of fraud. To estimate the effectiveness of the proposed approach, we will utilize datasets of real credit card transactions. To evaluate the detection efficacy of each method, performance measures will include F1-score, recall, accuracy, and precision. By comparing the numerous algorithms, we can determine the most effective one that has been tested to be the best in detecting credit card fraud cases, Fig. 5.

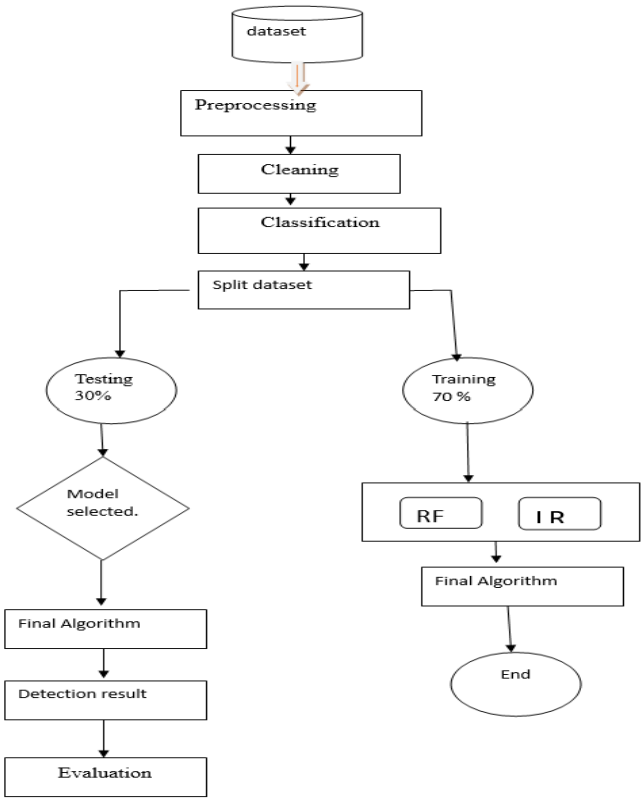


**Figure 5.** Proposed credit card fraud system structure

### 4.1. Dataset Description

The dataset can be downloaded from the Kaggle website. A significant difference appears in the dataset. The components of Payment Card Assurance (PCA) are only derived from conversion-based numerical input variables. We are unable to offer the original functionality due to privacy concerns. Twenty-eight of the thirty features were derived by principal component analysis (PCA), which facilitated dimensionality reduction and the consolidation of several original variables into a more compact set of feature variants. Features V1, V2,..., V28 are the principal components derived from principal component analysis (PCA); the sole attributes that remain untransformed by PCA are time and quantity. The time feature encompasses the duration in seconds that has transpired from the initial transaction in the dataset to each following transaction. A characteristic termed 'amount' is employed in dependent cost-sensitive learning to denote the value of a transaction. In the absence of fraud, the categorical feature response variable is assigned a value of 0; nevertheless, upon detection of fraud, it is assigned a value of 1. Data preprocessing is the first step after collecting data and is performed according to the following steps.

### 4.2. Data cleaning

One critical step when working on credit card data analysis and modelling is data cleaning. It covers Missing Values, Outlier treatment, Duplicate entries, and data consistency. Here are some of the normal activities I would need to perform on a credit card dataset. Data cleaning is an essential step in the analysis and modelling of credit card data. It includes treatment for missing values, outliers, duplicates, and ensuring data consistency.

### 4.3. Data splitting

It will utilize a traditional data-splitting schema, in which 70% of the available credit card data is designated as training data and 30% as testing data. This also ensures that a large sample size is allocated for model learning, and another one is reserved for model testing on out-of-sample data. Model training and validation need to be balanced, so this approach of data splitting is also acceptable.

### 4.4. Balancing data

To address the problem of class imbalance in the credit card data, the proposed solution incorporates the Synthetic Minority Over-sampling Technique (SMOTE). Class imbalance occurs when the number of instances of different classes is significantly disproportionate. The cases of the prevailing group correspond to legal transactions, whereas the cases of the minor group represent instances of fraud. Algorithms in machine learning can be biased when they are not able to give due consideration to the minority classes, because of the imbalance in classes. SMOTE is a useful mechanism for dealing with class imbalance in resampling. SMOTE aims to increase the representation of the minority group in the dataset by artificially creating additional instances [26]. The method will enable a more balanced sample allocation across classes, thereby increasing the ability of machine learning algorithms to recognize patterns in minority groups within classes and enhancing the discriminative potential of the model. The SMOTE algorithm has several steps. A minority case is selected randomly to initiate the development of fake cases. The closest neighbors of an instance are determined by a user-defined parameter, which specifies the number of neighbors to be examined. The synthetic samples are created by interpolating the feature values between the chosen instance and its closest neighbors. With the adoption of this method, SMOTE has the potential to mitigate the negative effects of class imbalance and enhance the representation of the minority class. SMOTE does not require the entire dataset; it only needs to be applied to the training sets. This division facilitates a critical evaluation of the model in terms of its generalizability to other untested data, as the test samples in this case are not contaminated with false positive results.

Moreover, to avoid overfitting, it is better to create random samples rather than simply duplicating them. The precision and reliability of fraud detection will also be improved with the proposed approach, which includes SMOTE to address the issue of unbalanced classes in the credit card study. One technique that can be used to improve the reporting of fraudulent transactions is resampling, which more effectively balances the minority class in the sample. This is to neutralize bias towards the majority class in general learning. The classification of the datasets before and after applying SMOTE is presented in Table 1.

**Table 1.** The comparison between the dataset before and after using SMOTE

| Class | Out | Number of records before SMOT | Number of records after SMOT |
|---|---|---|---|
| 0 | valid | 284807 | 284315 |
| 1 | Fraud | 492 | 284315 |

### 4.5. Classification

The suggested method identifies fraudulent transactions by integrating a classification strategy with various machine learning methods. Four algorithms—Random Forest, XGBoost, Logistic Regression, and Decision Tree—are employed for their proven efficacy in fraud detection applications. Each algorithm is instructed to utilize the preprocessed dataset to identify the characteristics and patterns that differentiate genuine transactions from fraudulent ones. To assess the legitimacy of new, unexamined credit card transactions, the system utilizes trained categorization models. The models utilize established patterns and characteristics to assign probability for classifying transactions as legal or fraudulent. A threshold can be established to ascertain the requisite level of certainty for reporting a transaction as fraudulent.

## 5. RESULTS AND DISCUSSIONS

The correlation of data in the proposed system refers to the study of the correlations between different variables or attributes in the data. With its help, it is better to understand the direction and the strength of a linear relationship between two variables. Knowing data correlation is critical in the selection of features, the identification of redundant variables, and understanding the linkages that enable the detection of fraud. One of the tools that is commonly used in the proposed system to examine the correlation between two variables is the correlation coefficient in the form of a table about the correlation coefficient between the two variables; the correlation coefficients of -1, 1, and 0 indicate very strong negative, very strong positive and no correlation between the two variables, respectively, Fig. 6.

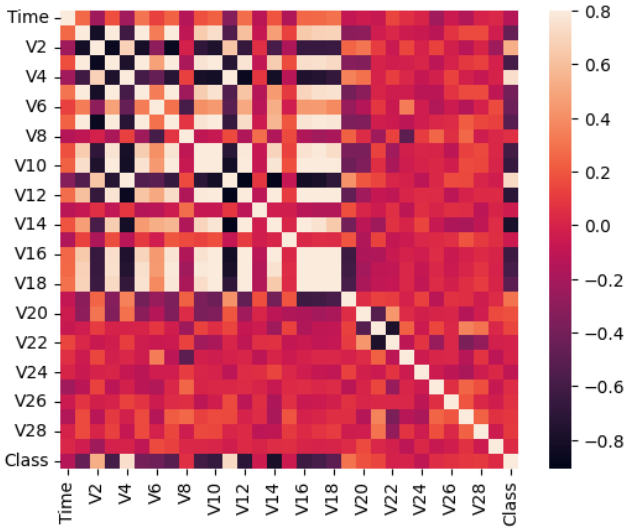


**Figure 6.** The correlation matrix for the proposed system features.

Analyzing the correlation matrix provides insights into the interrelationships among various elements in the dataset. Variables with high positive or negative correlations may indicate strong associations or dependencies. These correlations can help identify prospective traits that are significantly associated with fraudulent transactions or variables that exhibit high correlation with one another.

### 5.1. Dataset Collection and Preprocessing Results

The dataset after collection is described in Fig. 7.

| | O | P | Q | R | S | T | U | V | W | X | Y | Z | AA | AB | AC | AD | AE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 96334 | 0.269661 | 1.349107 | -2.34039 | 0.644954 | 0.394897 | 0.175915 | -0.14283 | -0.65602 | -1.40705 | 0.086743 | 0.020357 | -0.09256 | 0.351887 | -0.14732 | -0.15283 | 68.8 | 0 |
| 96335 | -0.83797 | 0.723859 | 0.445492 | 1.24756 | 0.825234 | 0.575256 | 0.22929 | 0.079409 | -0.17564 | -0.24133 | -0.76926 | 0.059454 | 0.476446 | -0.09208 | -0.18721 | 0.76 | 0 |
| 96336 | -1.70743 | 0.900083 | -0.01175 | 0.008094 | -0.39817 | -0.91337 | 3.565081 | -1.2194 | -1.04497 | 0.316481 | -0.4786 | 1.230784 | 0.021728 | 1.039631 | -0.171 | 13.48 | 0 |
| 96337 | -0.01814 | 0.393223 | 0.338497 | -0.55341 | -0.27076 | -0.54687 | 0.00767 | 0.171216 | 0.419877 | -0.25153 | 0.474195 | 0.047654 | 0.97053 | -0.01711 | 0.092606 | 0.67 | 0 |
| 96338 | 0.170575 | 0.800536 | 0.421126 | -0.46564 | -0.12744 | -0.50281 | -0.06279 | 0.057481 | 0.16957 | 0.099984 | 0.247774 | 0.047542 | 0.24742 | 0.010909 | 0.018824 | 19.58 | 0 |
| 96339 | -0.11717 | -0.85237 | 0.024862 | -0.24647 | -0.42765 | 0.463033 | -0.04992 | -0.2983 | -0.71615 | 0.150882 | 0.563002 | 0.155718 | 0.158571 | -0.0149 | 0.014117 | 10 | 0 |
| 96340 | -1.18187 | -1.87223 | 0.06686 | -0.50167 | 0.675336 | -0.10317 | 0.327476 | 0.240835 | 0.67759 | 0.057862 | 1.034255 | 0.256103 | -0.28656 | 0.157785 | 0.157197 | 159.66 | 0 |
| 96341 | -0.59441 | -1.13284 | 0.025646 | -0.17647 | -0.28467 | 0.730214 | 0.124695 | -0.13491 | -0.22757 | 0.004609 | 0.453094 | 0.128637 | 0.914835 | -0.03975 | 0.016415 | 63.22 | 0 |
| 96342 | 0.176226 | 0.946583 | -1.064 | -0.30101 | 1.194307 | -0.46459 | -0.31964 | -0.40116 | -0.84288 | -0.12 | -0.42979 | 0.350621 | 1.08582 | -0.08236 | 0.008277 | 81.25 | 0 |
| 96343 | -1.11274 | -0.54085 | 1.533411 | 0.831443 | -0.47335 | 1.190121 | 0.273799 | -0.02605 | -0.29525 | -0.18046 | -0.43654 | 0.494649 | -0.28374 | -0.00113 | 0.035075 | 98.01 | 1 |
| 96344 | -0.38978 | -0.66594 | -0.3928 | -0.13724 | -0.62785 | -0.0933 | -0.07064 | 0.052663 | 0.386938 | -0.04318 | -0.71208 | -0.69252 | 0.242657 | 0.190439 | 0.193908 | 28.75 | 0 |
| 96345 | -0.25677 | 1.347209 | 0.186607 | 0.273777 | -0.74665 | -0.55636 | -0.11517 | -0.26068 | -0.71615 | 0.159804 | 0.033516 | 0.14317 | 0.124259 | -0.00651 | 0.027226 | 0.89 | 0 |
| 96346 | 0.181879 | -0.24094 | -0.91738 | 0.397321 | -0.1764 | 1.105127 | 0.246361 | -0.02022 | -0.09951 | -0.04133 | 0.020561 | 0.408418 | -0.26129 | 0.012647 | -0.0091 | 51.99 | 0 |
| 96347 | 0.238537 | -0.75733 | -0.47595 | 0.24669 | -0.46512 | 0.049168 | -0.21876 | -0.0088 | 0.179993 | -0.03183 | 0.271544 | 0.424293 | 0.441027 | -0.00879 | -0.00315 | 3.79 | 0 |
| 96348 | 0.433785 | 0.41801 | 0.751408 | -0.9606 | 0.25194 | 0.484062 | -0.07259 | -0.26339 | -0.80333 | 0.025519 | -0.50883 | 0.289892 | 0.129299 | -0.03535 | -0.00021 | 0.89 | 0 |
| 96349 | 0.365493 | 0.121277 | 1.345309 | -0.00538 | -0.62525 | 0.931302 | 0.069613 | 0.245722 | 0.483132 | -0.1775 | 0.01788 | 0.591537 | -0.11297 | -0.02407 | -0.00044 | 40.85 | 0 |
| 96350 | 0.469418 | -0.7029 | -0.54001 | 0.142963 | -0.42942 | -0.15247 | -0.22492 | -0.04076 | 0.009708 | -0.04831 | 0.206387 | 0.568696 | -0.34028 | 0.027959 | 0.007717 | 19.8 | 0 |
| 96351 | 0.307313 | 1.019997 | 0.26594 | -0.65083 | 0.118528 | -0.21725 | -0.13015 | 0.03396 | 0.159754 | -0.17549 | -0.13645 | 0.751563 | -0.26317 | 0.019137 | 0.016052 | 1.99 | 0 |
| 96352 | 0.393004 | 0.338512 | 0.118925 | 0.027721 | 0.170146 | 1.484484 | 0.886848 | -1.60819 | -0.64601 | 0.079534 | 0.953688 | 0.038508 | 0.008425 | 0.072022 | 0.123583 | 0.76 | 0 |
| 96353 | -2.56867 | 1.314812 | 0.924888 | 3.750004 | 1.175215 | -0.96673 | 0.364927 | -0.13551 | -0.25961 | 0.429638 | -0.31357 | -0.13707 | -0.33185 | 0.400734 | 0.127043 | 0.99 | 0 |
| 96354 | -0.50445 | 0.727453 | -0.19396 | -0.71361 | 1.782057 | -1.61499 | -0.32168 | -0.05243 | 0.040576 | -0.19692 | -0.08208 | 0.132391 | 1.03783 | -0.11183 | -0.04075 | 36.49 | 0 |
| 96355 | 0.514873 | 1.62438 | -0.43683 | 0.581366 | -1.21356 | -1.69672 | -0.27486 | 1.058734 | 0.666593 | -0.08269 | -1.01884 | 0.46366 | 0.412642 | 0.151321 | 0.077757 | 10.59 | 0 |

**Figure 7.** Sample of the dataset.

The suggested approach uses the Synthetic Minority Over-sampling Technique (SMOTE) to balance the unequal distribution of the classes in the credit card data. Inequality between socioeconomic classes arises when there is a substantial dissimilarity in the occurrence of events. Here, the minority group is an epitome of unethical conduct, while the majority serves as a model of integrity. If the machine learning algorithms fail to identify the minority classes effectively due to class imbalance, bias may be introduced. The second step of phase 2 involves correcting missing values and removing outliers from the data. In the Synthetic Minority Over-sampling Technique

(SMOTE), a 30% portion of the data is used for testing, and 70% is used for training. The inclusion of SMOTE in the proposed methodology represents a significant modification to the dataset distribution, particularly when distinguishing between genuine and fraudulent information. To redistribute the data, SMOTE uses the available data to overrepresent the cases of the underrepresented group (fraudulent transactions). The data in Fig. 8 and Fig. 9 illustrate the data distribution before and after applying the SMOTE technique, respectively.

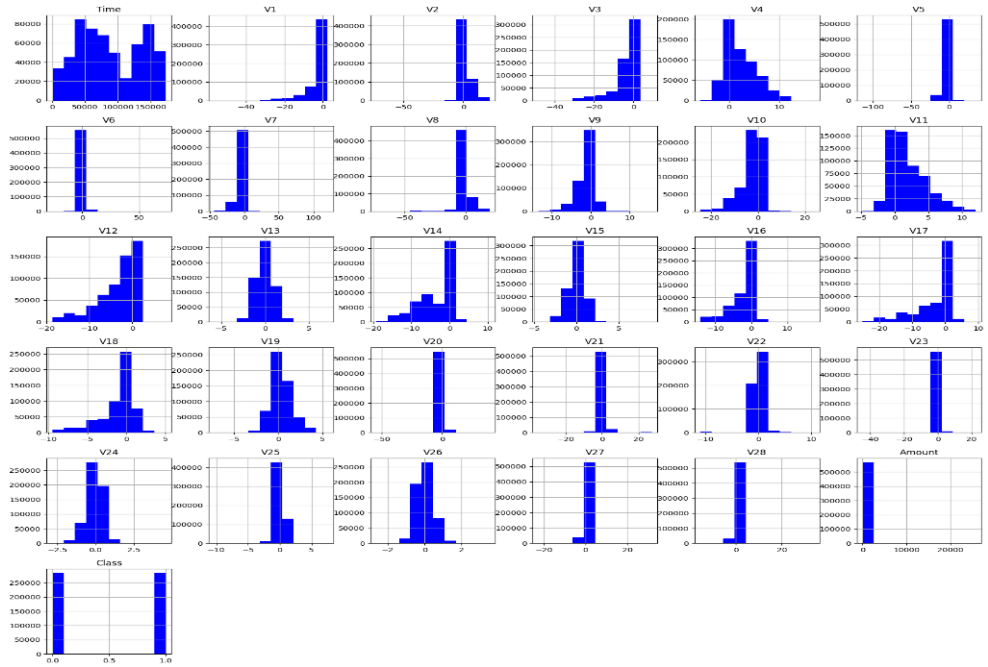


**Figure 8.** The data distribution before using SMOTE

Before the use of the SMOTE dataset, the dataset could have been characterized by a very high level of class imbalance, with only a limited number of fraudulent transactions against the valid transactions. This imbalance may have adverse effects on the performance of classification algorithms, as they can be biased towards the majority class. Nevertheless, with the implementation of SMOTE, the data will become more balanced, and the rate of fraudulent transactions is expected to decrease.

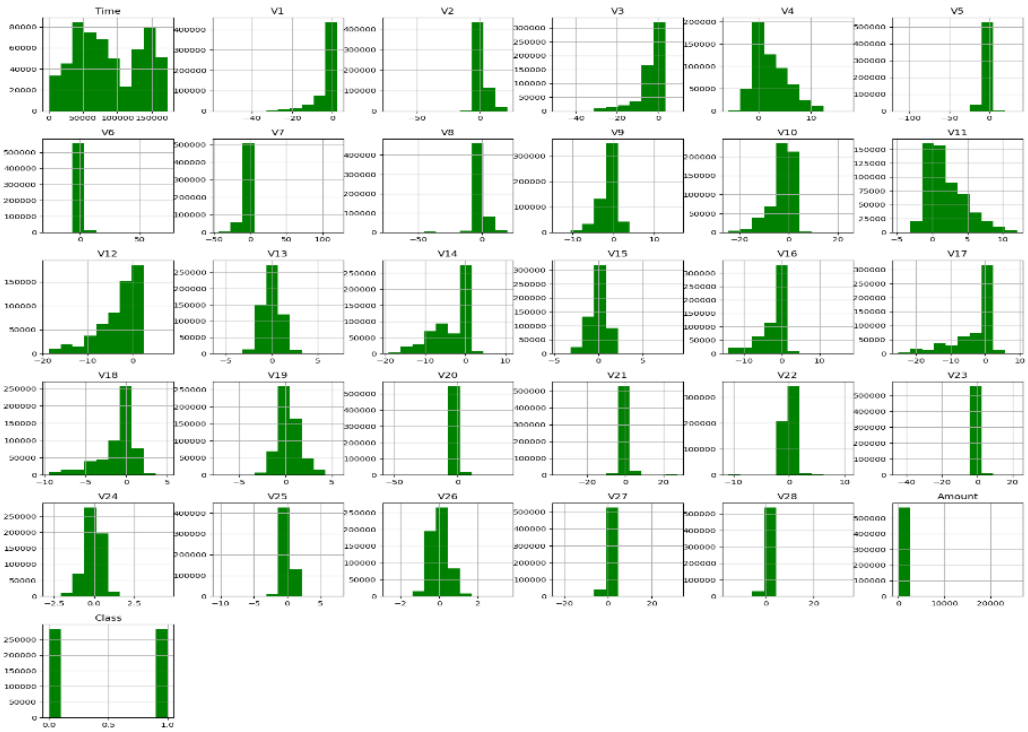


**Figure 9.** The data distribution after STMOE

The results of the distribution of fraud after balancing are shown in Fig. 10, and the distribution of valid data is shown in Fig. 11.

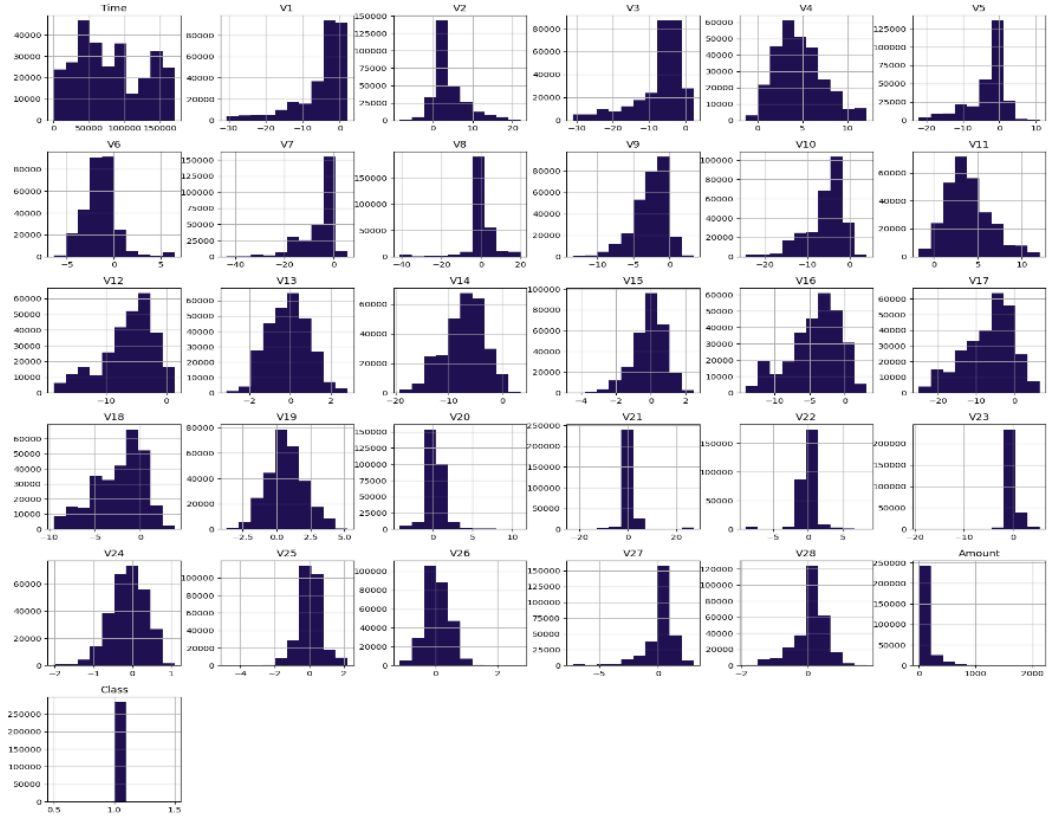


**Figure 10.** The results of the distribution of fraud data

The SMOTE technique generates synthetic instances of the minority class by interpolating between neighboring cases. As a result, the distribution of fraud data is expanded, with new cases being added to represent different fraud patterns. This augmentation helps improve the learning capacity of the classification algorithms and enhances their ability to identify various types of fraudulent transactions.

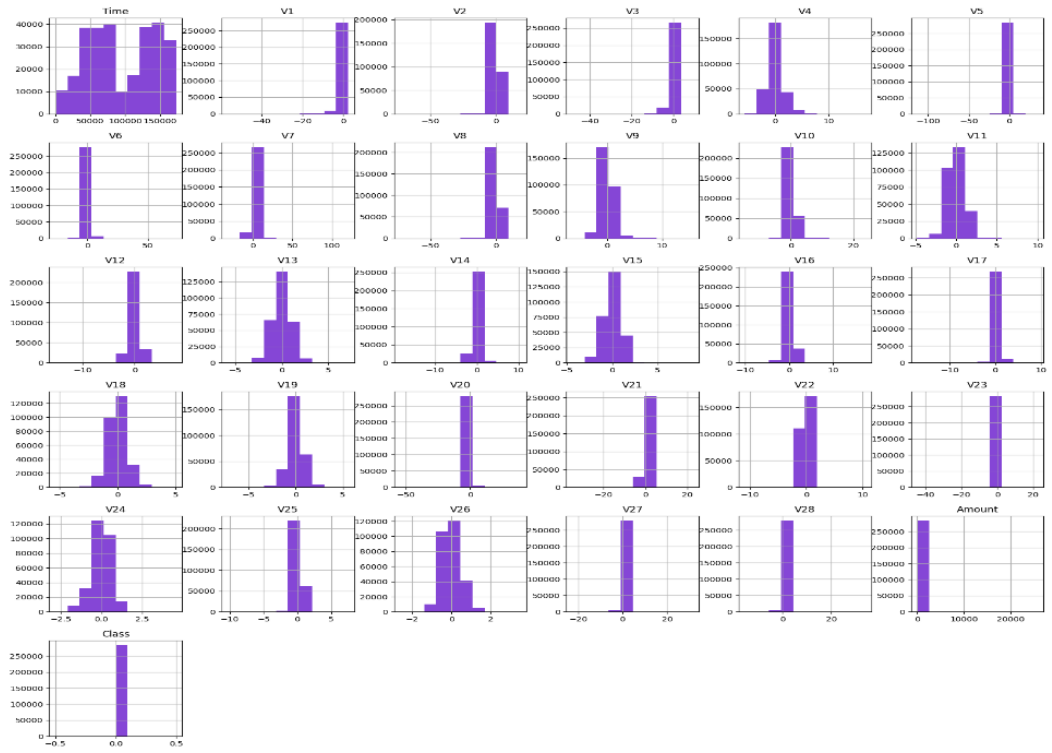


**Figure 11.** Distribution of valid data

Although the primary purpose of SMOTE is to maintain a balance for the minority group, it can also influence the distribution of valid information. The process of interpolation may produce minute variations in the distribution of the majority class instances. The impact on valid data distribution, however, is not as significant as the drastic changes that occur in the fraudulent data distribution. The rebalanced dataset, with a more equal representation of fraud and legitimate data, enables classification models to learn from a larger set of examples and enhance their ability to distinguish between fraudulent and legitimate transactions. This, consequently, improves the precision and efficiency of the suggested system in detecting fraud within credit card systems.

### 5.2. The Classification Results

The outcomes of the algorithms' operation in the proposed system for identifying credit card fraud provide valuable insights into the efficiency of the specified approach. To establish the quality of the system for identifying legitimate and fraudulent transactions, several outcomes are necessary. The outcomes of implementing the techniques suggested by the system are as follows.

One of the most notable features of the Random Forest algorithm is its ability to work with complex data and effectively uncover correlations that cannot be discovered through linear methods. The results indicate that the RF algorithm was an effective classifier of credit card transactions, exhibiting low recall, high precision, and an F1 score. The confusion table for the random forest method is represented in Fig. 12 and Fig. 13.

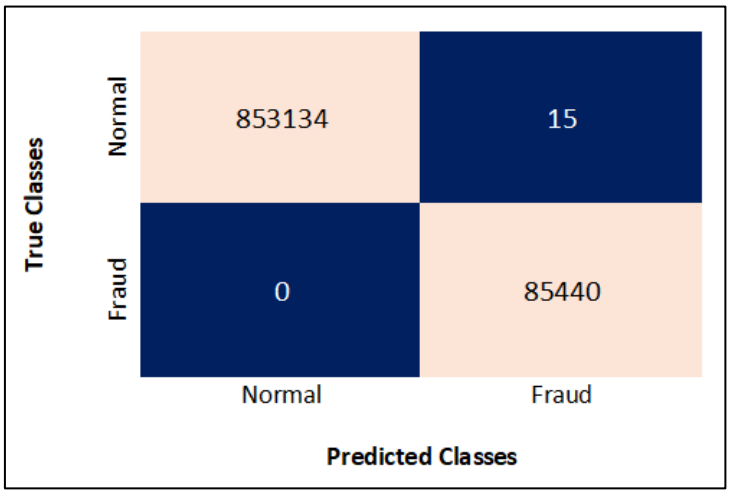


**Figure 12.** The confusion matrix of using the RF algorithm for the proposed system

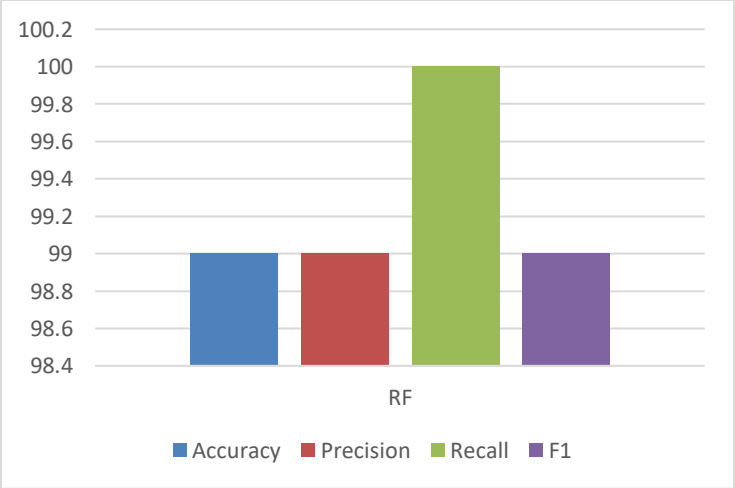


**Figure 13.** The evaluation results for RF

Logistic Regression is a widely used linear model that estimates the probability of a binary outcome. The results indicate that LR achieved satisfactory performance in classifying credit card transactions, with reasonably high accuracy, precision, recall, and F1 score. While LR may not capture complex, nonlinear relationships as effectively as other algorithms, it provides

interpretable coefficients and is computationally efficient, Fig. 14 and Fig. 15.

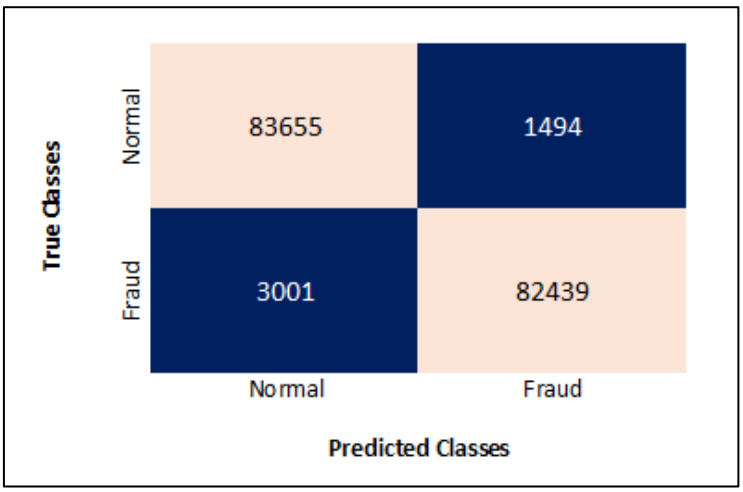


**Figure 14.** The confusion matrix of using the Logistic Regression algorithm for the proposed system

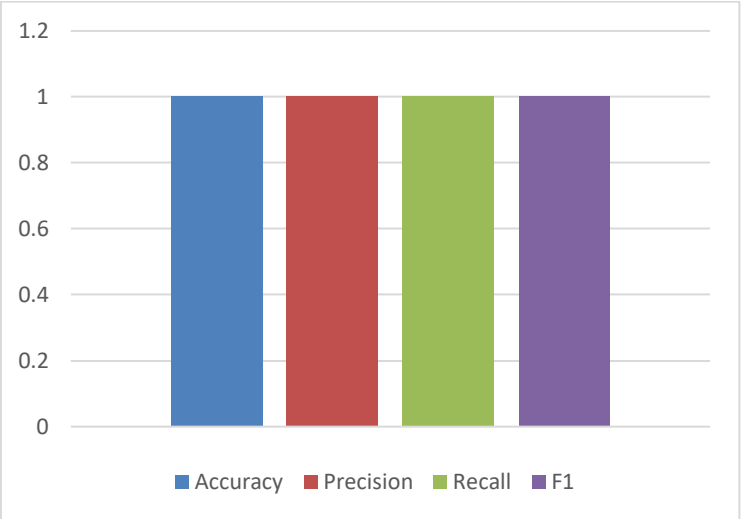


**Figure 15.** The evaluation results for Logistic Regression

## 6. CONCLUSION

Our initiative, titled "Designing and Building a Credit Card Fraud Detection Model Using Machine Learning," aimed to develop techniques for identifying fraudulent activities and mitigating financial losses. Numerous methods utilizing supervised learning are employed; two examples are Logistic Regression (with 97.34% accuracy) and Random Forest (with 99.99% accuracy). To assess different methodologies, we adhere to the original dataset. We have utilized sampling and oversampling methods due to the significant imbalance in our dataset. Finally, if the RF system could identify online fraud more effectively and correctly than any alternative, forecasts would be significantly enhanced. Findings indicate that bigger sample sizes and a more even distribution of workers enhance the efficacy of our model training. Oversampling may be the most suitable sampling strategy when conventional information needs to be preserved in a real-world environment.

## CONFLICT OF INTEREST

The authors declare that there is *no* conflict *of interest* regarding the publication of this paper.

## REFERENCES

[1] G. Sandhya, M. Abishek, S. Gunal Kumar, and R. S. Jisenthira Kumar, "Credit card fraud detection using machine learning algorithms," ICT Systems and Sustainability: Proceedings of ICT4SD 2022, 2022, Springer, vol. 516, pp. 313–320, 2023, doi: 10.1007/978-981-19-5221-0_30.

[2] P. K. Sadineni, "Detection of fraudulent transactions in credit card using machine learning Algorithms," Proceedings of the 4th International Conference on IoT in Social, Mobile, Analytics and Cloud, ISMAC 2020, pp. 659–660, Oct. 2020, doi: 10.1109/I-SMAC49090.2020.9243545.

[3] H. John and S. Naaz, "Credit Card Fraud Detection using Local Outlier Factor and Isolation Forest," International Journal of Computer Sciences and Engineering, vol. 7, no. 4, pp. 1060–1064, Apr. 2019, doi: 10.26438/ijcse/v7i4.10601064..

[4] K. Murugan, A. Felicia, B. Gomathy, P. T. Saravanakumar, S. M. Ramesh, and E. Sakthivel, "A Credit Card Fraud Identification Technique Using Support Vector Machine," International Conference on Applied Intelligence and Sustainable Computing, ICAISC 2023, 2023, doi: 10.1109/ICAISC58445.2023.10199684.

[5] M. M. Koska, C. Aktürk, and T. Talan, "Detection of Credit Card Fraud with Machine Learning Methods," ICHORA 2025 - 2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings, 2025, doi: 10.1109/ICHORA65333.2025.11017135.

[6] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system: A survey," Journal of Network and Computer Applications, vol. 68, pp. 90–113, Apr. 2016, doi: 10.1016/j.jnca.2016.04.007.

[7] A. A. Gunjal, "Application of demolished construction waste for manufacturing of paver block to analyze the result of partial replacement of demolished construction waste in Paver block," International Journal for Research in Applied Science and Engineering Technology, vol. 8, no. 7, pp. 1775–1779, Jul. 2020, doi: 10.22214/ijraset.2020.29614.

[8] J. O. Awoyemi, A. O. Adetunmbi and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," 2017 International Conference on Computing Networking and Informatics (ICCNI), Lagos, Nigeria, 2017, pp. 1-9, doi: 10.1109/ICCNI.2017.8123782.

[9] P. K. Sadineni, "Detection of fraudulent transactions in credit card using machine learning Algorithms," Proceedings of the 4th International Conference on IoT in Social, Mobile, Analytics and Cloud, ISMAC 2020, pp. 659–660, Oct. 2020, doi: 10.1109/I-SMAC49090.2020.9243545.

[10] E. F. Malik, K. W. Khaw, B. Belaton, W. P. Wong, and X. Chew, "Credit Card Fraud Detection Using a New Hybrid Machine Learning Architecture," Mathematics 2022, Vol. 10, Page 1480, vol. 10, no. 9, p. 1480, Apr. 2022, doi: 10.3390/MATH10091480.

[11] A. Kumar, S. Dutta, and P. Pranav, "Analysis of SQL injection attacks in the cloud and in WEB applications," Security and Privacy, vol. 7, no. 3, p. e370, May 2024, doi: 10.1002/SPY2.370.

[12] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," EURASIP Journal on Advances in Signal Processing 2016 2016:1, vol. 2016, no. 1, pp. 1–16, May 2016, doi: 10.1186/S13634-016-0355-X.

[13] P. Chatterjee, D. Das, D. Rawat, and D. B. Rawat, "Securing Financial Transactions: Exploring the Role of Federated Learning and Blockchain in Credit Card Fraud Detection," Authorea Preprints, Oct. 2023, doi: 10.36227/TECHRXIV.22683403.V1.

[14] S. Bhatia, M. Sharma, and K. K. Bhatia, "Sentiment Analysis and Mining of Opinions," Studies in Big Data, vol. 30, pp. 503–523, 2018, doi: 10.1007/978-3-319-60435-0_20.

[15] H. Abdel-Jaber, D. Devassy, A. Al Salam, L. Hidaytallah, and M. El-Amir, "A Review of Deep Learning Algorithms and Their Applications in Healthcare," Algorithms, vol. 15, no. 2, Feb. 2022, doi: 10.3390/A15020071.

[16] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: a review of classification and combining techniques," Artificial Intelligence Review, vol. 26, no. 3, pp. 159–190, Nov. 2006, doi: 10.1007/s10462-007-9052-3.

[17] F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan and M. Ahmed, "Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms," in IEEE Access, vol. 10, pp. 39700-39715, 2022, doi: 10.1109/ACCESS.2022.3166891.

[18] W. Deng, Z. Huang, J. Zhang, and J. Xu, "A Data Mining Based System for Transaction Fraud Detection," 2021 IEEE International Conference on Consumer Electronics and Computer

Engineering, ICCECE 2021, pp. 542–545, Jan. 2021, doi: 10.1109/ICCECE51280.2021.9342376.

[19] I. D. Mienye and N. Jere, "Deep Learning for Credit Card Fraud Detection: A Review of Algorithms, Challenges, and Solutions," IEEE Access, vol. 12, pp. 96893–96910, 2024, doi: 10.1109/ACCESS.2024.3426955.

[20] Y. Chen, C. Zhao, Y. Xu, C. Nie, and Y. Zhang, "Deep Learning in Financial Fraud Detection: Innovations, Challenges, and Applications," Data Science and Management, Aug. 2025, doi: 10.1016/J.DSM.2025.08.002.

[21] M. Grzelak, P. Owczarek, R. M. Stoica, D. Voicu, and R. Vilău, "Application of Logistic Regression to Analyze The Economic Efficiency of Vehicle Operation in Terms of the Financial Security of Enterprises," Logistics 2024, Vol. 8, Page 46, vol. 8, no. 2, p. 46, May 2024, doi: 10.3390/LOGISTICS8020046.

[22] A. Rasool, F. Rehman, N. Sarfaraz, H. Sharif, R. Khan, and A. M. Khan, "Machine Learning Based Impostor Detection by Invariant Features from Nir Finger Vein Imaging," 3rd International Conference on Innovations in Computer Science and Software Engineering, ICONICS 2022, 2022, doi: 10.1109/ICONICS56716.2022.10100461.

[23] A. RB and S. K. KR, "Credit card fraud detection using artificial neural network," Global Transitions Proceedings, vol. 2, no. 1, pp. 35–41, Jun. 2021, doi: 10.1016/J.GLTP.2021.01.006.

[24] M. Grzenda, S. Kaźmierczak, M. Luckner, G. Borowik, and J. Mańdziuk, "Evaluation of machine learning methods for impostor detection in web applications," Expert Syst Appl, vol. 231, p. 120736, Nov. 2023, doi: 10.1016/J.ESWA.2023.120736.

[25] A. Issa, Y. Ali, and T. Rashid, "An efficient hybrid classification approach for COVID-19 based on Harris Hawks Optimization and Salp Swarm Optimization," International journal of online and biomedical engineering, vol. 18, no. 13, pp. 113–130, Dec. 2022, doi: 10.3991/ijoe.v18i13.33195.

[26] C. Meng, L. Zhou, and B. Liu, "A Case Study in Credit Fraud Detection With SMOTE and XGBoost," J Phys Conf Ser, vol. 1601, no. 5, p. 052016, Aug. 2020, doi: 10.1088/1742-6596/1601/5/052016.