Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Credit Card Fraud Detector Based on Machine Learning Techniques

Omar Rajab Mohsen[1], Ghalia Nassreddine[2] ✉ and Mazen Massoud[3]
[123]Business department, Jinan University, Tripoli, Lebanon
Corresponding Author: Ghalia Nassredddine, **E-mail**: ghalia.nasseredine@jinan.edu.lb

| ABSTRACT
The massive development of technology has affected commerce and given rise to e-commerce and online shopping. Nowadays, consumers prioritize e-shopping over the brick and motor stores due to numerous benefits, including time and transport convenience. However, this progressive upsurge in online payment increases the number of credit card frauds. Therefore, defending against fraudsters' activity is obligatory and can be achieved by securing credit card transactions. The objective of this paper is to build a model for credit card fraud detection using Machine learning techniques. An innovative approach to credit card fraud detection grounded on machine learning is proposed in this study. Machine learning (ML) is an artificial intelligence subfield comprising learning techniques from experience and completing tasks without being explicitly programmed. Three ML techniques have been used: Support vector machine, logistic regression, Random Forest, and Artificial Neural network. First, the most significant features that affect the type of transaction (fraud or not fraud) have been selected. After that, the ML model was applied. The performance of the proposed approach is tested using a confusion matrix, recall, precision, f-measure, and accuracy. The proposed method is tested using accurate data that consists of 284807 transactions. The result shows the efficiency of the proposed approach.

| KEYWORDS

Credit Card, Fraud Detection, Machine Learning, Support Vector Machine, Logistic Regression, Artificial Neural Network, Random Forest.

| ARTICLE INFORMATION

ACCEPTED: 25 June 2023 **PUBLISHED:** 30 June 2023 **DOI:** 10.32996/jcsts.2023.5.2.2

1. Introduction

Electronic commerce, or e-commerce, represents buying and selling products/services online. It embraces transferring accounts or data using the internet or any other electronic network. Usually, these transactions may occur in several methods, including (Raziei, 2020):

- Business-to-business (B2B): includes types of transactions among businesses, for instance, a factory and wholesaler.
- Business-to-consumer (B2C): represents direct transactions between businesses and consumers, for instance, online shops.
- Consumer-to-consumer (C2C): represents the transaction involving the company and consumers.
- Consumer-to-business (C2B): consumers deliver products or services to businesses in exchange for payment or other advantages.

The use of e-commerce has gained high popularity among organizations and consumers worldwide. In 2010, e-commerce represented 5% of total retail sales in the USA. However, e-commerce increased to 16% of retail sales during the Corona pandemic. This progress affected the financial sector, and digitalization of its operation is now an essential prerequisite (Rajora et al., 2018). Recently, digitalization processes are prevalent due to their consistency and simplicity. Digital transformation encompasses conveniently effective and satisfying payment methods. Generally, consumers are adopting e-shopping for many reasons. E-

shopping saves time, reduces transportation expenditures, and offers shoppers wider shopping choices. Consumers can compare products online in their homes (Popat& Chaudhary, 2018). However, the massive demand for online shopping increased credit card fraud cases. Credit card fraud is an identity theft involving transactions using another's credit card information to pay purchases or withdraw funds.

Consequently, deterring fraudster's activity becomes mandatory, leveraging the need to secure credit card transactions. Many techniques have been used to create secure e-banking (Dhankhad et al., 2018). Fraud prevention procedures are not limited to these two main approaches (Rajora et al., 2018). First, fraud prevention covers preclusion activities seeking to reduce or eliminate the possibility of fraud before its occurrence. This technique is mainly used for ATMs or websites (Balagolla et al., 2021). Second, the fraud detection process covers adequately identifying a fraudulent transaction after its manifestation (Chen et al., 2021). The process of fraud detection can be applied using two methods:

- The Expert-Driven approach indicates the position of the actual transaction. This approach is based on a set of straightforward rules created by professional experts. However, it is very challenging for human brains to think in more than three-dimensional variables and generate all possible pattern combinations (Malviya& Yadav, 2018).
- The Data-Driven approach is a model generated using Machine Learning techniques. The primary purpose of this model is to find the defrauding patterns in available data. Its main advantage is to detect a new form of fraud even if it has not been executed. Regardless, this method cannot demonstrate the generated alerts (Tran et al., 2018).

Machine learning (ML) is an extension of artificial intelligence representing machine technical capabilities to perform tasks. Recently, ML has been widely used in many sectors, such as finance, banking, healthcare, and education (Zhou, 2021). The latter does not rely on complicated programs.

This study aims to build a model for credit card fraud detection using ML techniques. This model belongs to the data-driven approach. The main contribution of this study is to test the accuracy of existing ML techniques in credit-card fraud-detection problems. In this study, two ML techniques will be studied:

- Classification Techniques: Logistic Regression, Artificial Neural Networks, and Support vector machines will be tested in detecting credit card fraud problems.
- Ensemble learning Techniques: The accuracy of the random forest method will be tested.

This paper is organized as follows: First, Machine learning will be presented in Section 2. The credit card fraud problem will be illustrated in Section 3. The proposed approach will be described in Section 4. The used dataset and the result will be shown in Section 6. This chapter ends with a conclusion in Section 6.

2. Machine Learning

During the last decade, Information technology tools have witnessed an enormous evolution and development. The human brain is conducted to develop different systems that make human life more comfortable. Humans have created many tools **to** achieve several daily tasks. They are able individuals to complete additional life requirements such as traveling, healthcare, and banking.

Machine learning (ML) represents an advanced science offering computer-based systems. It denotes learning and developing using historical information without human intervention. It is used to train machines to manage data efficiently. However, data cannot be interpreted based on current information due to altered circumstances. Difficulties in data interpretation raised technical encounters in information extraction. Therefore, ML techniques are mandatory to be applied. Consequently, organizations adopting advanced technology implement machine learning to extract appropriate data. The increased awareness of the ML technique and the benefits derived from its use motivated critical economic sectors to embrace it.

ML systems demonstrate intelligence just like a human. Machine learning aims to learn from an available dataset to predict appropriate outcomes. The studies (Ni et al., 2021) have endorsed this advantage. Waring et al. (2020) verified that ML allows organizational systems to acquire knowledge without explicit programs and systems. Thus, ML will enable systems to think, feel and act like humans. Therefore, current ML models generate innovative tasks, including classifying mail as spam or no-spam, fixing grammar and spelling mistakes, recognizing objects from images, detecting fake news in social media, and understanding written and spoken words (Cioffi et al., 2020). The following section embraces the application process of ML steps.

2.1. Process

Importing intelligence into systems appears to be a complex process. However, this amalgamation process is straightforward. It can be divided into seven successive tasks, as illustrated in Figure 1.

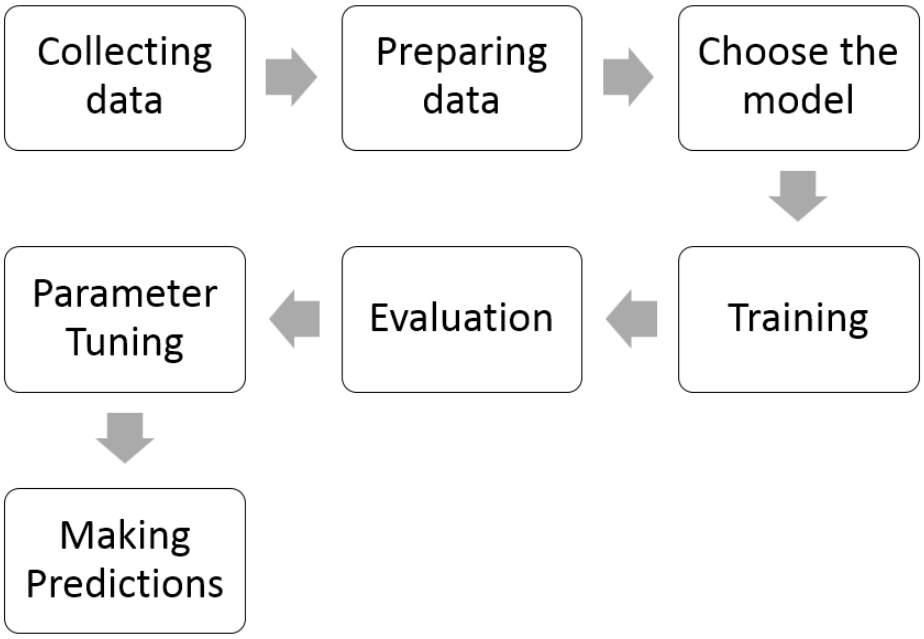

Figure 1: Main steps of the ML system (Hüffmeier et al., 2020)

2.1.1. Collecting data

In the first step of data collection, the system should extract its knowledge base from accessible data. Therefore, collecting data is an essential task of the ML process. Data quality provided to the system will determine the accuracy level of the ML model. If the data is incorrect or imprecise, the output of the ML system may be irrelevant. Therefore, the quality of incorporated data should possess four specified properties (Lee & Shin, 2020). First, the data should be obtained from reliable sources. It should be error-free, or at least it can incorporate a few missing cases. The used data should be precise with significant values to the project.

2.1.2. Preparing Data

Data preparedness should be constructed on the following steps:

- Data randomization and arrangement should ensure equal distribution among collected information. Data arrangement should not influence the learning process significantly.
- Data cleaning data should be proceeded by eliminating inappropriate and repeated figures. This step can comprise data restructuring and adjustment. For instance, data could be altered according to needed rows, columns, and indexes.
- Data should be displayed as graphs or histograms to provide a statistical comprehension of its fundamental structure. These statistical outputs determine the relationship between different variables.
- Dividing cleaned data into two sets: training to regulate the model and testing to compute the accuracy of a model after training

2.1.3. Choosing a Model

An appropriate ML technique is selected after carefully understanding an accurate data structure. The chosen model should be highly compatible with a pre-determined problem. Over the last decade, countless algorithms have been generated according to established needs, such as face recognition, speech recognition, and forecasting. The model should be selected according to problem type and data to get the highest accurate outcomes.

2.1.4. Training

It is one of the main steps in the ML process in which the training set is used to find a pattern and produce a prediction. As a result, the model learns how to complete the task set. The model improves at predicting over time as it is trained.

2.1.5. Evaluation

This step evaluates the model's performance after training. The evaluation is performed based on the testing set. The testing set should be different from the training set. If not, the result of the assessment will not be accurate.

2.1.6. Parameter Tuning

In this step, the model's accuracy and performance can be improved by tuning the model parameters. Established parameters are determined by the programmer as variables and incorporated into the model. The accuracy will be maximum through given *parameter values. These values are referred to as "tuning parameters."*

2.1.7. Generating Predictions

In this step, the model will be used on new data to predict accurately.

2.2. Benefits and Limitations of Machine Learning

ML is a very effective artificial intelligence tool. ML handles current technology evolution and assists humans in performing tasks efficiently. It is a potent tool holding transformational and revolutionizing potential. As every coin has two faces, machine learning has limitations and benefits.

2.2.1. Benefits of ML Tools

ML tools are widely used for their various benefits, including (Jordan & Mitchell, 2015):

- Easily recognized patterns: Machine Learning can examine a vast data set and uncover precise structures and practices. For example, ML tools are deployed to grasp browsing behaviors and purchase records. For instance, it can dissect the browsing history of Amazon consumers. This analysis helps Amazon detects consumers' needs and matches the right products with the right prospect.
- Automation: ML-based systems do not require human intervention. Indeed, ML allows the methods of automatic learning and programmed prediction. In addition, algorithms can be improved by themselves using previous experience. For instance, as new threats are identified in antivirus software, they learn to filter them. ML is adept at detecting spam.
- Continuous Improvement: decision making accuracy and performance improves through ML algorithms' accumulated experience. It produces restored predictions. For example, in a weather forecast model, the volume of recorded data increases continuously. Therefore, ML algorithms can be trained based on these data and make more precise forecasts faster.
- Manage multi-dimensional data: Machine Learning algorithms are virtuous at managing multi-dimensional and multi-variety data. ML can achieve this target in active or insecure environments.
- Massive Applications: ML techniques are widely used in leading sectors (Wang et al., 2009). ML could be applied to traffic prediction, text recognition, product recommendation, fraud detection, email filtering, self-driving cars, navigation systems, language translation, medical diagnosis, robotics, finance, and banking.

2.2.2. Limitations of ML Techniques

Along with its benefits, three critical elements limit ML application (Malik, 2020):

- Data collection: ML requires massive data sets. These data sets should be inclusive/unbiased and of high quality. Therefore, new data generation is mandatory at this stage.
- Time and Resources: Machine learning requires time for algorithms training and data generation. Sufficient data is needed to fulfill the purpose. This data should be accurate and relevant. For example, it could imply that more computing power is required. It involves scarifying the high number of resources.
- High error sensitivity: ML algorithms are self-contained but extremely sensitive to errors. For instance, an algorithm training with too small and unrepresentative data; therefore, biased predictions result from a small training set. Customers are exposed to irrelevant advertisements as a result. In the case of machine learning, such errors can start a chain reaction that can go undetected for long periods. It takes a long time to identify and correct the problem's source when they are discovered.

2.3. Machine Learning Techniques

Problems can be classified into one of n available categories based on the similarity index of its characteristic. Algorithms classification can be trained to execute complex tasks without human intervention. ML-based systems are deployed in businesses to decrease operational costs, improve efficiency, and raise speed. Numerous algorithms are arrayed to classify problems. This study underlines only three essential techniques:

- **Logistic Regression (LR):** Logit models are commonly used in classification and predictive analytics. Logistic regression calculates the likelihood of an event occurring based on independent variables, such as voting or not voting. The dependent variable ranges from 0 to 1 because the outcome is a probability. In logistic regression, the odds—the likelihood of success divided by the probability of failure—are transformed using a logit transformation. This logistic function is also known as the log odds or the natural logarithm of odds, and the following formulas represent it:

$$Logit(pi) = 1/(1 + exp(-pi))$$

$$ln(pi/(1-pi)) = Beta\_0 + Beta\_1 * X\_1 + \ldots + B\_k * K\_k$$

- **Support Vector Machine (SVM):** An N-dimensional hyperplane (N — the number of features) distinguishes data points. Several hyperplanes could be used to separate the two data points classes. For instance, Increasing the margin distance helps to find a plane with the greatest significant margin or space between data points from both categories. It adds reinforcement by allowing future data points to be classified accurately.
- **Artificial Neural Network (ANN):** An "artificial neural network" (ANN) is a biologically inspired artificial intelligence subfield. It is modeled after the brain. An artificial neural network (ANN) is a computational network based on biological neural networks taking the form of the human brain structure. Neurons in artificial neural networks are linked in various layers. It has a high similarity with the human brain neurons interconnection model.
- **Random Forest:** The Random Forest (RF) machine learning algorithm, developed by Leo Breiman and Adele Cutler, combines the output of multiple decision trees to produce a single result. Its ease of use and flexibility have fueled its adoption because it handles classification and regression problems.

Table 1 shows a comparison between these algorithms.

Table 1: Comparison between LR, SVM, ANN, and RF

Algorithm	Advantage	Disadvantage
LR	-It is easier to implement, interpret, and train. -It is quick to classify unknown records. -It is accurate for many simple data sets and performs well when the dataset is linearly separable.	The assumption of linearity between the dependent and independent variables is required, as is average or no multi-collinearity between independent variables.
SVM	-It is effective in high-dimensional spaces and works well with unstructured and semi-structured data such as text and images. -It is based on the geometrical properties of the data.	- SVM is ineffective with large data sets. -SVM performs poorly when there is more noise in the data set -SVM underperforms for an incredibly massive number of features concerning the number of training data samples. -This classification lacks a probabilistic explanation. The support vector classifier places data points above and below the classifying hyperplane.
ANN	-A neural network can perform tasks that a linear program cannot. -When a neural network item declines, it can continue without some of its similar features. -A neural network determines without the need for reprogramming. -It can be used in any application.	-To function, the neural network required training. -Because the structure of a neural network differs from that of microprocessors, it must be emulated. -It took a long time to process large neural networks.
RF	-It improves accuracy by reducing overfitting in decision trees. -It is applied to both classification and regression problems.	-It requires a lot of computational power and resources because it builds many trees and then combines their outputs.

	-It is effective with both categorical and continuous values. It automates missing values in data. It does not require data normalization because it uses a rule-based approach.	-It also takes a long time to train because it combines many decision trees to determine the class. -It also lacks interpretability due to the ensemble of decision trees and failure to determine each variable's significance.

2.4 Credit Card Fraud

Credit card fraud can denote cash withdrawal from another person's credit card using an Automated Teller Machine (ATM). It embraces paying for a purchase without the credit card owner's permission or knowledge. These actions can be accomplished by getting a credit card via physical theft. Another technique of credit card fraud is to steal the credit card number with the necessary personal information to produce illegal transactions (Varmedja et al., 2019). Credit card fraud detection is different from identity theft. Indeed, Credit card stealing is the most typical form of identity theft. The latter can be restricted as stealing the personal information of another individual. However, credit card fraud is using another person's credit card information to get money or commit illegal purchases (Al Smadi & Min, 2020). Five credit card fraud types are classified:

2.4.1 Card-Not Present (Mekterović, et al., 2021):

Card-not-present fraud is the most significant type of widespread credit card fraud. Card-not-present transactions are one of the most popular types. As in many purchases, actual cards are not used physically. Simply customers should enter credit card details to complete a purchase. Online shopping is an example of this type.

There are different techniques to protect from this type of fraud. One of these methods is 3D security which other banks apply to protect the credit card owner from fraud. It is a security element that is available on a debit card. Certain e-shopping websites participating in the 3D secure procedure use this method to demand a one-time password (OTP) for each payment. The OTP will send to the registered mobile phone of the card owner. However, to protect the credit card information from any thread, it is better to only enter the card details on secure and reputable e-shopping websites. It is also better to check the website's security certificate to confirm that it can be trusted. In addition, be sure that on the webpage on which one enters the card information, the web address begins with 'HTTPS,' and the browser displays a padlock.

2.4.2 Counterfeit and Skimming (Shetty & Murthy, 2022; Singh & Jain, 2020):

Card skimming theft is credit card fraud involving credit card usage at ATMs, shopping centers, or restaurants. This type of fraud is based on using a skimmer - a device integrated into card readers to collect credit card information. The thieves used the collected data to make fraudulent purchases. Skimmers can usually be detected by making quick inspections before inserting a card. This type of fraud is primarily used in ATMs (see Figure 2).

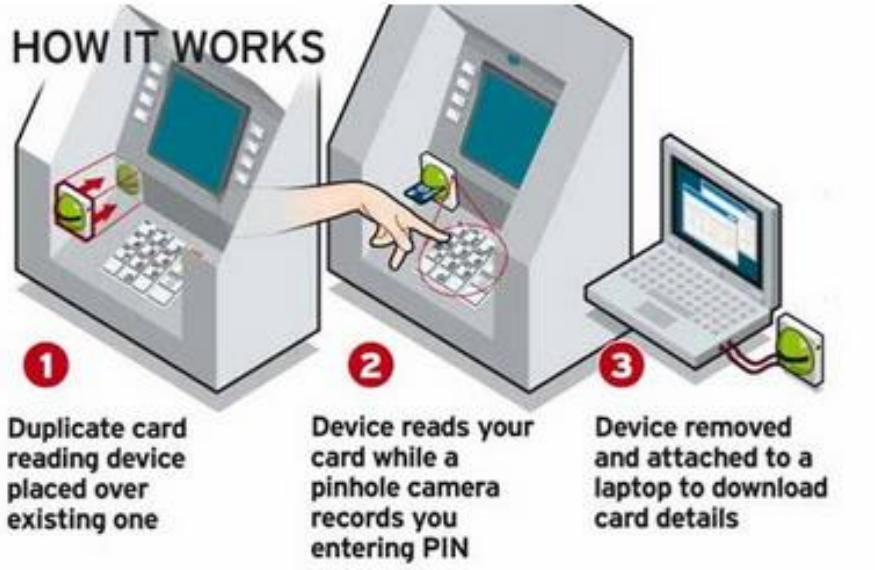

Figure 2: Skimmer card (United Financial, 2019).

A credit card PIN is a set of four-digit identification numbers to confirm the cardholder's identity. It is obligatory to use the PIN to achieve a transaction. In addition, many techniques can be used to steal the PIN. One of the stealing PIN techniques is to use a small camera implanted near card reader devices to record the PIN. Another technique is using fake keypads over an ATM keypad (see Figure 3).

Figure 3: Fake keyboard for card skimmer (Krebs, 2010).

Practical tips are beneficial to avoid being skimmed (Lee & Scott, 2017):

- Memorize the PIN and do not write it down or share it with anyone.
- Before using any ATM, perform a quick scan. Examine it to ensure that it has not been tampered with.
- Do not insert or swipe the card if the card reader appears loose, crooked, or damaged, the graphics are not aligned, or if part of the machine is a different color.
- If another machine is nearby (for example, two ATMs next to each other), compare them to see if there are any apparent differences.
- Try not to use a non-bank ATM.
- Check for a possibly fake ATM keyboard.
- Cover the keypad with one's hands when entering the PIN if a camera records the number.
- Use ATMs in public view with security monitoring: Indeed, these machines are less likely to have been tampered with.
- Check the account regularly; if anything is wrong, contact the bank and stop the card immediately.

2.4.3 Lost and Stolen Card (Lakshmi & Kavilla, 2018):

If the card has been lost or stolen, the thief can use it until it is canceled. Therefore, the first step the credit card user should take is to call the bank to cancel the stolen or lost card. Some banks allow the user to cancel the credit card with a simple button in their mobile application.

2.4.4 Card-Never Arrived (Lakshmi & Kavilla, 2018):

When a user requests a credit card, the card will be sent via mail in many countries. Card-never-arrived fraud is the credit card *fraud type in which the card is either blocked before it arrives or the card thief steals it from the user's inbox.*

2.4.5 False Application (Lakshmi & Kavilla, 2018):

According to this classification, another person may request a substitute credit card for another beneficiary or apply for a card in a different name by ensuring a connection to a separate bank account. This person may run up thousands of dollars on a credit card or completely use one's credit score before recognizing that the account has been stolen. Therefore, the stolen client gets stuck with repayments checks.

2.5 Credit Card Fraud Detector

A credit card payment is a straightforward technique. This system sends a few numbers to the bank to verify an account and authorize the transaction. However, this operation has become unsafe recently. Indeed, it is very complicated to apply strict data

security on a few simple numbers that must be shared with the parties to which the transaction will apply. For this reason, detecting fraudulent transactions becomes essential for merchants and e-shopping.

Credit card fraud detection identifies and rejects fraudulent transactions instead of processing them. There are many known tools and methods used for detecting fraud. Many businesses use a combination of these methods (Bin Sulaiman et al., 2022).

Figure 4 displays the credit-card fraud-detection system and its structure. A credit card check will be achieved when a credit card transaction occurs. After that, the transaction will be rejected if the credit is blocked or the balance is insufficient.

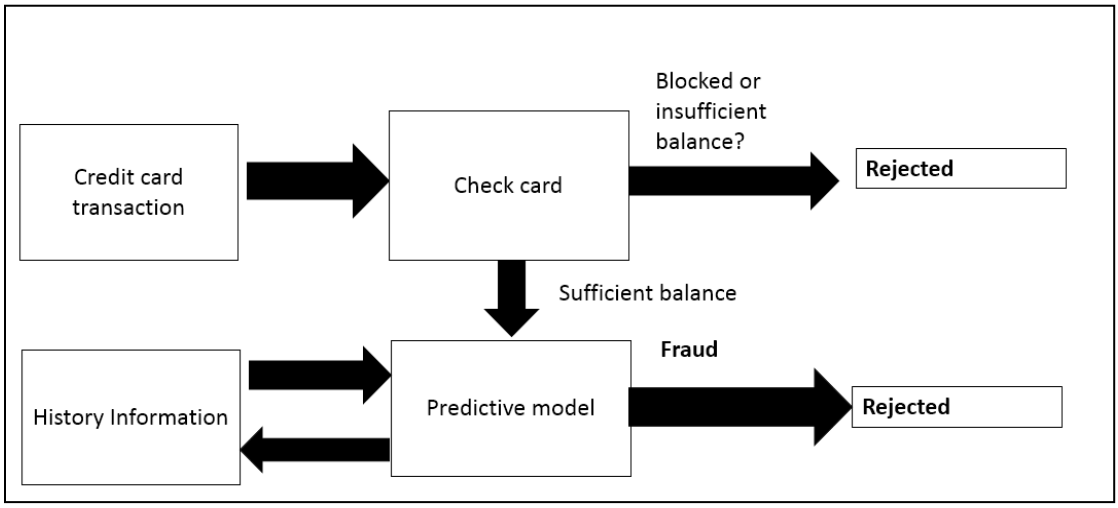

Figure 4: A credit card fraud detection framework (Cheng et al., 2020).

A predictive model is applied when the balance is sufficient to detect fraudulent transactions. If it is a fraud, it will be rejected. The predictive model uses historical information to detect fraudulent transactions (Cheng et al., 2020).

3. Machine Learning Based-System for Fraud Detection

Because of the rapid growth in the cashless or digital transactions field, credit cards are widely used worldwide. Customers receive thousands of credit cards from credit card companies. According to providers, all credit card users must be genuine and authentic. Any mistake in issuing a card can result in a financial crisis. The possibility of fraudulent transactions is increasing due to the rapid growth of cashless transactions. Analyzing various credit card customer behaviors from previous transaction history datasets assists in identifying fraudulent transactions. Any deviation from recognized spending patterns indicates a fraudulent transaction. Credit card fraud is commonly detected using data mining and machine learning techniques. Figure 2 shows the main steps of this proposed method. The confusion matrix (CM) – is used to determine the performance metrics for each algorithm.

3.1 Confusion Matrix

CM visually represents the data classification. This matrix divides outputs into two or more categories (see Table 2).

Table 2: Classifications based on the confusion matrix (actual vs. predicted classes)

	Predicted True	**Predicted False**
Actual True	True Positive (TP)	False Negative (FN)
Actual False	False Positive (FP)	True Negative (TN)

This classification, represented in Table (2), is explained as follows:

- The actual class denotes the correct type of any element (relevant vs. non-relevant candidate). The predicted class, on the other hand, represents the class indicated by an ML model.
- A relevant candidate has a positive value of 1, whereas an irrelevant candidate has a negative value of 0.
- The number of relevant predicted elements that should be relevant is represented by True Positive (TP). In contrast, the number of irrelevant predicted factors that should be irrelevant is represented by True Negative (TN).
- The number of relevant predicted elements that should be irrelevant is represented by False Positive (FP). In contrast, the number of unrelated predicted factors that should be relevant is represented by False Negative (FN).

3.2 Performance Metrics

- The precision is the value of TP elements over TP and FP, calculated using CM as follows:

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

- The recall is the value of TP elements over TP and FN, calculated using CM as follows:

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

- The f-measure, based on CM, balances the precision and recall values as follows:

$$\boldsymbol{F - measure} = \mathbf{2} \times \frac{\boldsymbol{Recall} \times \boldsymbol{Precision}}{\boldsymbol{Recall} + \boldsymbol{Precision}}$$

- The accuracy, based on CM, is the ratio of correct predictions to the sample size as follows:

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

4. Result

In this section, the proposed method will be tested on a real from Kaggle.com. First, the dataset will be presented. Then, the most significant attributes will be selected. Finally, the results will be illustrated and discussed.

4.1 Dataset

The dataset was obtained from "kaggle.com." It contains credit card transactions made by European cardholders in September 2013. This dataset includes 492 frauds out of 284,807 transactions over two days. The dataset is severely unbalanced. Positive (fraud) transactions account for 0.172% of total transactions.

We cannot provide the data's original features and additional background information due to confidentiality concerns. PCA's primary components are features V1, V2,..., and V28. PCA does not transform the elements of 'Time' and 'Amount.' It only accepts numerical input variables derived from a PCA transformation. The 'Time' feature specifies the number of seconds possessing an elapse between each transaction and the first transaction in the dataset. The 'Amount' feature represents the transaction amount. This feature is helpful for cost-sensitive learning based on examples. If there is fraud, the response variable 'Class' is one. Otherwise, it is zero.

This dataset contains 31 columns representing the main attributes of the data (see Figure 6). It illustrates that the last column (called class) has two possible values (0 or 1). This column represents the associated class of each transaction:

- Zero represents a not_fraud class.
- One depicts a Fraud class.

Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V
0	-1.3598	-0.0728	2.5363	1.3782	-0.3383	0.4624	0.2396	0.0987	0.3638	0.0908	-0.5516	-(
0	1.1919	0.2662	0.1665	0.4482	0.06	-0.0824	-0.0788	0.0851	-0.2554	-0.167	1.6127	:
1	-1.3584	-1.3402	1.7732	0.3798	-0.5032	1.8005	0.7915	0.2477	-1.5147	0.2076	0.6245	(
1	-0.9663	-0.1852	1.793	-0.8633	-0.0103	1.2472	0.2376	0.3774	-1.387	-0.055	-0.2265	(
2	-1.1582	0.8777	1.5487	0.403	-0.4072	0.0959	0.5929	-0.2705	0.8177	0.7531	-0.8228	(
2	-0.426	0.9605	1.1411	-0.1683	0.421	-0.0297	0.4762	0.2603	-0.5687	-0.3714	1.3413	(

	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24
16	-0.6178	-0.9914	-0.3112	1.4682	-0.4704	0.208	0.0258	0.404	0.2514	-0.0183	0.2778	-0.1105	0.0669
27	1.0652	0.4891	-0.1438	0.6356	0.4639	-0.1148	-0.1834	-0.1458	-0.0691	-0.2258	-0.6387	0.1013	-0.3398
45	0.0661	0.7173	-0.1659	2.3459	-2.8901	1.11	-0.1214	-2.2619	0.525	0.248	0.7717	0.9094	-0.6893
65	0.1782	0.5078	-0.2879	-0.6314	-1.0596	-0.6841	1.9658	-1.2326	-0.208	-0.1083	0.0053	-0.1903	-1.1756
28	0.5382	1.3459	-1.1197	0.1751	-0.4514	-0.237	-0.0382	0.8035	0.4085	-0.0094	0.7983	-0.1375	0.1413
13	0.3599	-0.3581	-0.1371	0.5176	0.4017	-0.0581	0.0687	-0.0332	0.085	-0.2083	-0.5598	-0.0264	-0.3714

V26	V27	V28	Amount	Class
-0.1891	0.1336	-0.0211	149.62	0
0.1259	-0.009	0.0147	2.69	0
-0.1391	-0.0554	-0.0598	378.66	0
-0.2219	0.0627	0.0615	123.5	0
0.5023	0.2194	0.2152	69.99	0
0.1059	0.2538	0.0811	3.67	0

Figure 6: columns of the dataset.

Each column (choose one) represents data associated with one transaction. Moreover, the data contains 284807 transactions. Figure 7 shows the histogram of the attribute class. It illustrates that only 492 transactions from 284807 transactions are fraud (0.17% of all transactions). This percentage is very low, therefore, tough to detect.

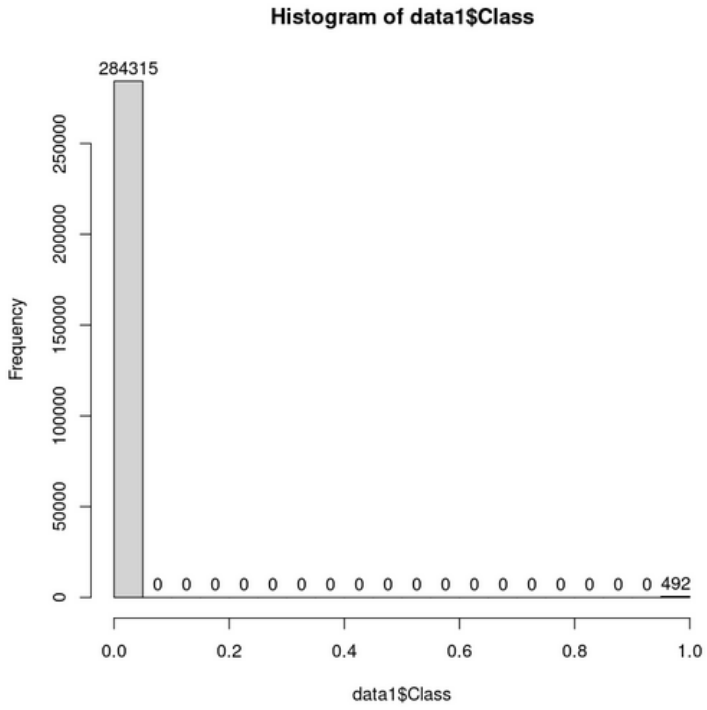

Figure 7: Histogram of attribute Class

4.2 Significant Attributes

The dataset used in this study is large, with 31 attributes. Therefore, a test should be done to select the most significant attributes using the glm() function. The result of this function shows that only attributes V4, V8, V9, V10, V13, V14, V16, V20, V21, V22, V27 and V28 are significant. Therefore, we can remove 19 attributes from 31 as they are insignificant. After removing all the attributes that are not significant, the dimension of the dataset contains 284807 rows and 12 columns.

4.3 Logistic regression results

Table The result is shown in Table 3.

Table 3: Confusion Matrix for LR on the testing set

		Predicted	
		Not Fraud	Fraud
Actual	Not Fraud	59842	9
	Fraud	36	73

As illustrated in Table 3, the following values can be concluded:

- TN is equal to 59842. Thus, 59842 negative transactions are classified as unfavorable (negative value) from 59851 (99%)
- FP is equal to 9. Thus, nine transactions are classified as positives; however, they are negatives. Therefore, only nine negative transactions are missing.
- TP is equal to 73. Therefore, 73 transactions are classified as fraud from 109 (66%)
- FN is equal to 36. Thus, only 36 transactions from 109 fraud transactions are missing.

4.4 Support vector machine results

The confusion matrix is illustrated in Table 4. From this Table, the following values can be concluded:

Table 4: Confusion Matrix for SVM on the testing set

		Predicted	
		Not Fraud	Fraud
Actual	Not Fraud	59848	9
	Fraud	26	77

According to Table 4, the following points can be concluded:

- TN is equal to 59848. Thus, 59848 negative transactions are classified as negative from 59857 (99%)
- FP is equal to 26. Thus, twenty-six transactions are classified as positives; however, they are negatives. Therefore, only twenty-six negative transactions are missing.
- TP is equal to 77. Therefore, 77 transactions are classified as a fraud from 86 (89.5%)
- FN is equal to 9. Thus, only nine transactions from 86 fraud transactions are missing.

4.5 Random Forest Results

The result of the confusion matrix is illustrated in Table 5.

Table 5: The confusion matrix of the RF model on the testing set

		Predicted	
		Not Fraud	Fraud
Actual	Not Fraud	59840	17
	Fraud	46	57

According to Table 5, the following points can be concluded:

- True Negative (TN) is equal to 59840. Thus, 59840negative transactions are classified as negative from 59857 (99%)
- False Positive (FP) is equal to 17. Thus, seventeen transactions are classified as positives; however, they are negatives. Therefore, only seventeen negative transactions are missing.
- True positive (TP) is equal to 57. Therefore, 57 transactions are classified as a fraud from 103 (55.3%)
- False negative (FN) is equal to 46. Thus, only 46 transactions from 103 fraud transactions are missing.

4.6 Artificial Neural Networks Results

The performance is computed using the confusion matrix and other metrics (see Table 6).

Table 6: The confusion matrix of the ANN model on the testing set.

		Predicted	
		Not Fraud	Fraud
Actual	Not Fraud	59838	19
	Fraud	36	67

According to Table 6, the following points can be concluded:

- True Negative (TN) is equal to 59838. Thus, 59838 negative transactions are classified as negative from 59857 (99%)
- False Positive (FP) is equal to 19. Thus, nineteen transactions are classified as positives; however, they are negatives. Therefore, only nineteen negative transactions are missing.
- True positive (TP) is equal to 67. Therefore, 67 transactions are classified as a fraud from 103 (65%)
- False negative (FN) is equal to 36. Thus, only 36 transactions from 103 fraud transactions are missing.

4.7 Comparison

The four techniques applied in the proposed method concerning credit card fraud detection will be compared. Table 7 shows a comparison according to recall, precision, f-measure, and accuracy.

Table 7: Comparison between LR, SVM, RF, and ANN

Algorithm	Recall (%)	Precision (%)	F-measure (%)	Accuracy (%)
LR	82	59	69	99
SVM	89.5	74.7	81.5	99
RF	55.3	77	64.3	99
ANN	65	78	71.4	99

According to Table 5.7, the following points can be concluded:

- The accuracy of all models is very high (99%). This high percentage is due to the high number of elements in the used dataset. Indeed, the dataset contains 284807 elements, and only 492 transactions are fraudulent (0.14%). Therefore, the comparison cannot be made according to the accuracy value.
- The recall value is highest in SVM. As a result, the SVM correctly identifies the highest percentage of data samples as belonging to a positive class out of the total samples for that class. Therefore, The SVM has an actual positive rate.
- The precision is highest in the ANN. As a result, the model ANN produces the highest quality optimistic prediction.

F-measure will be used for comparison based on these points. The F-measure, also known as the F1 score, is a machine-learning evaluation metric that measures the accuracy of a model. It combines a model's precision and recall scores. As a result, SVM is the best model because it has the highest F-measure value.

5. Conclusion

Digitalization has gained traction due to its richness, clarity, and ease of use in e-commerce. Regardless, the considerable development of e-commerce and online payment raises the number of credit card frauds. Indeed, Credit card fraud represents one type of identity theft involving transactions employing another's credit card details to deliver purchases or withdraw funds. Hence, there is a vital requirement to deter the fraudster's activity. Therefore, the need for secure credit card transactions grows. For this reason, many techniques have been tested. Machine learning (ML) is a part of artificial intelligence incorporating innovative methods. The main objective of this study was to examine the efficiency of ML techniques in detecting credit card fraud transactions. Four ML techniques are used in this study: Logistic Regression, Support vector Machine, Random Forest, and Artificial Neural Network. First, the data is collected and cleaned. After that, the most significant features are selected based on the p-value. Thus, all not significant features were eliminated. The data is split now on the training and testing set. The ML model is trained using a training set. The model's performance is calculated based on Recall, Precision, F-measure, and Accuracy. The proposed approach is tested using a real dataset collected from kaggle.com. The dataset is enormous. It is composed of 31 features and 284807 rows. After testing the significance of the feature, only 18 elements were significant. Thus, the not substantial columns are removed. LR SVM, RF, and ANN results show that all algorithms have an accuracy equal to 99%. These outcomes are considered acceptable and expected as the percentage of fraudulent transactions in the dataset is very low (0.14%). Thus, accuracy cannot be used for comparison. The comparison is made according to the F1 score. The SVM has the highest value of f-measure (81.5%). The significance of this study can be resumed by:

- The step of selecting of most significant features before applying the ML algorithm increased its practical efficiency. This step is instrumental, especially for problems related to big data.
- Approve the efficiency of a simple ML technique such as LR in the fraud detection problem.
- The comparison is made according to the f-measure value and not the accuracy.

However, the efficiency of using a p-value test to select the significant feature should be tested with more complex ML models like random Forests, Artificial Neural networks, and others. The result shows the ML algorithm's efficacy for such a problem. After performing this study, the authors suggest the following:

- Banks should educate users about credit card fraud transactions and how thieves can do it.
- Banks should educate users on how to avoid the theft of their credit card data.
- Additional research should be done on ML tools and their methods for detecting fraudulent transactions.
- E-commerce should create a highly protected interface for online payment in online shopping.
- Banks should create advanced tracking mechanisms for possible fraudulent transactions with a direct notification system to the credit card holder.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Alzubi, J., Nayyar, A., & Kumar, A. (2018, November). Machine learning from theory to algorithms: an overview. In *Journal of Physics: Conference Series* (Vol. 1142, No. 1, p. 012012). IOP Publishing.
[2] Al Smadi, B., & Min, M. (2020, October). A critical review of credit card fraud detection techniques. In *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (pp. 0732-0736). IEEE.
[3] Al-Sahaf, H., Bi, Y., Chen, Q., Lensen, A., Mei, Y., Sun, Y., ... & Zhang, M. (2019). A survey on evolutionary machine learning. *Journal of the Royal Society of New Zealand, 49*(2), 205-228.
[4] Balagolla, E. M. S. W., Fernando, W. P. C., Rathnayake, R. M. N. S., Wijesekera, M. J. M. R. P., Senarathne, A. N., &Abeywardhana, K. Y. (2021, April). Credit card fraud prevention using blockchain. In *2021 6th International Conference for Convergence in Technology (I2CT)* (pp. 1-8). IEEE.
[5] Berry, M. W., Mohamed, A., & Yap, B. W. (Eds.). (2019). Supervised and unsupervised learning for data science. *Springer Nature*.
[6] Bin Sulaiman, R., Schetinin, V., &Sant, P. (2022). Review of Machine Learning Approach on Credit Card Fraud Detection. *Human-Centric Intelligent Systems*, 1-14.
[7] Bloomenthal, A. (2021, May). Credit Card: What It Is, How It Works, and How to Get One. Available Online: https://www.investopedia.com/terms/c/creditcard.asp. Last Accessed: December 27, 2022.
[8] Bursztyn, L., Ferman, B., Fiorin, S., Kanz, M., & Rao, G. (2018). Status goods: experimental evidence from platinum credit cards. *The Quarterly Journal of Economics, 133*(3), 1561-1595.
[9] Chen, J. I. Z., & Lai, K. L. (2021). Deep convolution neural network model for credit-card fraud detection and alert. *Journal of Artificial Intelligence*, *3*(02), 101-112.
[10] Cheng, D., Xiang, S., Shang, C., Zhang, Y., Yang, F., & Zhang, L. (2020, April). Spatio-temporal attention-based neural network for credit card fraud detection. In *Proceedings of the AAAI conference on Artificial intelligence* (Vol. 34, No. 01, pp. 362-369.
[11] Cioffi, R., Travaglioni, M., Piscitelli, G., Petrillo, A., & De Felice, F. (2020). Artificial intelligence and machine learning applications in smart production: Progress, trends, and directions. Sustainability, 12(2), 492.
[12] Connelly, L. (2020). Logistic regression. *Medsurg Nursing, 29*(5), 353-354.
[13] Dhankhad, S., Mohammed, E., & Far, B. (2018, July). Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study. In *2018 IEEE international conference on information reuse and Integration (IRI)* (pp. 122-125). IEEE.
[14] Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science, 14*, 241-258.
[15] Dornadula, V. N., &Geetha, S. (2019). Credit card fraud detection using machine learning algorithms. *Procedia computer science*, *165*, 631-641.
[16] Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence, 110*, 104743.
[17] Forough, J., &Momtazi, S. (2021). Ensemble of deep sequential models for credit card fraud detection. *Applied Soft Computing, 99*, 106883.
[18] Hüffmeier, J. & Lundman, J. & Elern, F. (2020). TRIM AND BALLAST OPTIMISATION FOR A TANKER BASED ON MACHINE LEARNING. 10.13140/RG.2.2.13476.50569.
[19] Itoo, F., & Singh, S. (2021). Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. *International Journal of Information Technology*, *13*(4), 1503-1511.
[20] Jain, V., Agrawal, M., & Kumar, A. (2020, June). Performance analysis of machine learning algorithms in credit card fraud detection. In *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)* (pp. 86-88). IEEE.
[21] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255-260.
[22] Krebs, B. (2010). All about skimmers. Available online: https://krebsonsecurity.com/all-about-skimmers/. Last Accessed: December 27, 2022.
[23] Lakshmi, S. V. S. S., & Kavilla, S. D. (2018). Machine learning for credit card fraud detection system. International Journal of Applied Engineering Research, 13(24), 16819-16824.
[24] Lee, J. G., & Scott, G. G. (2017). *Preventing credit card fraud: A complete guide for everyone from merchants to consumers*. Rowman & Littlefield.
[25] Lee, I., & Shin, Y. J. (2020). Machine learning for enterprises: Applications, algorithm selection, and challenges. *Business Horizons, 63*(2), 157-170.
[26] Lucas, Y., &Jurgovsky, J. (2020). Credit card fraud detection using machine learning: A survey. *arXiv preprint arXiv:2010.06479*.
[27] Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)*, *9*, 381-386.
[28] Malik, M. M. (2020). A hierarchy of limitations in machine learning. arXiv preprint arXiv:2002.05193.
[29] Malviya, A., & Yadav, H. (2020). An Assessment on Credit Card Fraud Detection: Survey. *International Journal of Advanced Computer Technology*, *9*(5), 12-15.
[30] Maniraj, S. P., Saini, A., Ahmed, S., & Sarkar, S. (2019). Credit card fraud detection using machine learning and data science. *International Journal of Engineering Research*, *8*(9), 110-115.
[31] Mekterović, I., Karan, M., Pintar, D., &Brkić, L. (2021). Credit card fraud detection in card-not-present transactions: Where to invest? *Applied Sciences*, *11*(15), 6766.

[32] Nasr, M., Hamdy, M., Hegazy, D., & Bahnasy, K. (2021). An efficient algorithm for unique class association rule mining. Expert Systems with Applications, 164, 113978.
[33] Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons. 4*, 51-62.
[34] Ni, D., Xiao, Z., & Lim, M. K. (2021). Machine learning in recycling business: an investigation of its practicality, benefits and future trends. *Soft Computing*, *25*, 7907-7927.
[35] Ozer, M. E., Sarica, P. O., & Arga, K. Y. (2020). New machine learning applications to accelerate personalized Medicine in breast cancer: Rise of the support vector machines. *Omics: a Journal of Integrative Biology, 24*(5), 241-246.
[36] Popat, R. R., & Chaudhary, J. (2018, May). A survey on credit card fraud detection using machine learning. In 2018 2nd international conference on Trends in Electronics and Informatics (ICOEI) (pp. 1120-1125). IEEE.
[37] Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of naive Bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing, 5*(2), 221.
[38] Rajora, S., Li, D. L., Jha, C., Bharill, N., Patel, O. P., Joshi, S., ... & Prasad, M. (2018, November). A comparative study of machine learning techniques for credit card fraud detection based on time variance. In *2018 IEEE Symposium Series on Computational* Intelligence (SSCI) (pp. 1958-1963). IEEE.
[39] Raziei, S. (2020). Evaluation of the Implementation of C2B, B2C, B2B, A2C and A2B Models of e-Commerce in Excellence Education System of Academic Institutions Using AHP and Fuzzy AHP. *International Journal of E-Education, e-Business, e-Management and e-Learning*, *10*(1), 13-24.
[40] Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(4), e1249.
[41] Sailusha, R., Gnaneswar, V., Ramesh, R., & Rao, G. R. (2020, May). Credit card fraud detection using machine learning. In *2020 4th international conference on intelligent computing and control systems (ICICCS)* (pp. 1264-1270). IEEE.
[42] Shetty, A. A., & Murthy, K. V. (2022). Investigation of Card Skimming Cases: An Indian Perspective. *Journal of Applied Security Research*, 1-14.
[43] Singh, A., & Jain, A. (2019). Adaptive credit card fraud detection techniques based on feature selection method. In *Advances in computer communication and computational sciences* (pp. 167-178). Springer, Singapore.
[44] Singh, A., & Jain, A. (2020). An empirical study of AML approach for credit card fraud detection–financial transactions. *International Journal of Computers Communications & Control*, *14*(6), 670-690.
[45] Singh, A., Thakur, N., & Sharma, A. (2016, March). A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development* (INDIACom) (pp. 1310-1315). IEEE.
[46] Shirgave, S., Awati, C., More, R., &Patil, S. (2019). A Review On Credit Card Fraud Detection Using Machine Learning. *International Journal of Scientific & technology research*, *8*(10), 1217-1220.
[47] Somvanshi, M., Chavan, P., Tambade, S., & Shinde, S. V. (2016, August). A review of machine learning techniques using decision tree and support vector machine. In *2016 International Conference On Computing Communication Control And Automation* (ICCUBEA) (pp. 1-7). IEEE.
[48] Sun, S. (2013). A survey of multi-view machine learning. *Neural Computing and Applications, 23*(7), 2031-2038.
[49] Taha, A. A., &Malebary, S. J. (2020). An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. *IEEE Access*, *8*, 25579-25587.
[50] Tanouz, D., Subramanian, R. R., Eswar, D., Reddy, G. P., Kumar, A. R., &Praneeth, C. V. (2021, May). Credit card fraud detection using machine learning. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 967-972). IEEE.
[51] Thennakoon, A., Bhagyani, C., Premadasa, S., Mihiranga, S., &Kuruwitaarachchi, N. (2019, January). Real-time credit card fraud detection using machine learning. In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 488-493). IEEE.
[52] Tran, P. H., Tran, K. P., Huong, T. T., Heuchenne, C., HienTran, P., & Le, T. M. H. (2018, February). Real-time data-driven approaches for credit card fraud detection. In *Proceedings of the 2018 international conference on e-business and applications* (pp. 6-9).
[53] United Financial (2019). Card Skimmers: What Are They? Available Online: https://www.unitedfinancialcu.org/card-skimmers-what-are-they/. Last Accessed: December 27, 2022.
[54] Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., &Anderla, A. (2019, March). Credit card fraud detection-machine learning methods. In *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)* (pp. 1-5). IEEE.
[55] Verma, S., Dickerson, J., & Hines, K. (2020). Counterfactual explanations for machine learning: A review. arXiv preprint arXiv:2010.10596.
[56] Wang, H., Ma, C., & Zhou, L. (2009, December). A brief review of machine learning and its application. In *2009 International Conference on Information Engineering and Computer Science* (pp. 1-4). IEEE.
[57] Waring, J., Lindvall, C., & Umeton, R. (2020). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial intelligence in Medicine, 104*, 101822.
[58] Xie, Y., Li, A., Gao, L., & Liu, Z. (2021). A heterogeneous ensemble learning model based on data distribution for credit card fraud detection. *Wireless Communications and Mobile Computing*.
[59] Zhou, Z. H. (2021). *Machine Learning*. Springer Nature.