Credit Card Fraud Detection using Machine Learning and Data Mining Techniques - a Literature Survey

Devicharan Rai M. [1] & Jagadeesha S. N. [2]

[1] Research Scholar, Institute of Computer Science and Information Science, Srinivas University, Mangalore, India,
Assistant Professor, Department of Computer Science, Vivekanada PU College, Puttur, India,
ORCIDID: 0000-0001-8758-8676; Email ID: devicharanrai2006@gmail.com

[2] Research Professor, Institute of Computer Science and Information Science, Srinivas University, Mangalore, Karnataka, India,
ORCIDID: 0000-0002-5185-2233; Email: jagadeesha2012@gmail.com

Subject Area: Computer Science.
Type of the Paper: Literature Review.
Type of Review: Peer Reviewed as per |C|O|P|E| guidance.
Indexed In: OpenAIRE.
DOI: https://doi.org/10.5281/zenodo.8190094
Google Scholar Citation: IJAEML

How to Cite this Paper:
Rai, D. M., & Jagadeesha, S. N. (2023). Credit Card Fraud Detection using Machine Learning and Data Mining Techniques - a Literature Survey. *International Journal of Applied Engineering and Management Letters (IJAEML)*, *7*(3), 16-35. DOI: https://doi.org/10.5281/zenodo.8190094

International Journal of Applied Engineering and Management Letters (IJAEML)
A Refereed International Journal of Srinivas University, India.

Crossref DOI: https://doi.org/10.47992/IJAEML.2581.7000.0186

Received on: 03/01/2023
Published on: 28/07/2023

Credit Card Fraud Detection using Machine Learning and Data Mining Techniques - a Literature Survey

Devicharan Rai M. [1] & Jagadeesha S. N. [2]

[1] Research Scholar, Institute of Computer Science and Information Science, Srinivas University, Mangalore, India,
Assistant Professor, Department of Computer Science, Vivekanada PU College, Puttur, India,
ORCID-ID: 0000-0001-8758-8676; Email ID: devicharanrai2006@gmail.com

[2] Research Professor, Institute of Computer Science and Information Science, Srinivas University, Mangalore, Karnataka, India,
ORCID-ID: 0000-0002-5185-2233; Email: jagadeesha2012@gmail.com

ABSTRACT

Purpose: *To understand the algorithms used in Credit Card Fraud Detection (CCFD) using Machine Learning (ML) and Data Mining (DM) techniques, Review key findings in the area and come up with research gaps or unresolved problem. To become knowledgeable about the current discussions in the area of ML and DM.*

Design/Methodology/Approach: *The survey on CCFD using ML and DM was conducted based on data from academic papers, web articles, conference proceedings, journals and other sources. Information is reviewed and analysed.*

Results/Findings: *Identification of credit card fraud is essential for protecting a person's or an organization's assets. Even though we have various safeguards in place to prevent fraudulent activity, con artists may develop a method to get around the checkpoints. We must create straightforward and efficient algorithms employing ML and DM to anticipate fraudulent activities in advance.*

Originality/Value: *Study of ML and DM algorithms in CCFD from diverse sources is done. This area needs study due to recent methods by fraudsters in digital crime have developed. The information acquired will be helpful for creating new methodologies or improving the outcomes of current algorithms.*

Type of Paper: *Literature Review.*

Keywords: Machine learning, Data mining, Credit card fraud, legitimate, fraudulent, SWOT analysis

1. INTRODUCTION :

The Internet is a global network that links billions of computers to the World Wide Web and to each other. Internet fraud is an attempt to deceive or mislead someone else online. Typically, this indicates that the victim of a scam loses money to the fraudsters. Computer services like chat rooms, e-mail, message boards, and websites can all be used for internet fraud. Types of internet fraud include Credit card fraud, Phishing, Online shopping frauds, Identity theft, Lottery fraud, Matrimonial frauds, Tax scams, etc.

The process of CCFD [1] involves the identification of fraudulent purchase attempts and the subsequent refusal to accept such attempts in place of the order. Data mining is the process of gleaning useful information from enormous amounts of previously collected data. Finding algorithms that have become better because of experience gained from data is known as machine learning.

Software programmes are able to make more accurate predictions of outcomes without needing to be explicitly told to do so when machine learning (ML) [2], which is a kind of artificial intelligence (AI) [3] is utilised. The utilisation of historical information as a source of information by ML algorithms makes it possible to generate reliable projections of future output values. The process of investigating the data and creating a model that will best reveal and prevent fraudulent transactions is known as CCFD with machine learning.

2. OBJECTIVES OF REVIEW PAPER :

(1) To study content coverage of current theories, research, and information in CCFD.
(2) To explore the utilization of ML to assess CCFD
(3) To evaluate DM for CCFD assessment and discussion.
(4) To perform SWOT analysis of CCFD using ML and DM.

3. METHODOLOGY :

The term "cardholder data" refers to any identifying information about a person that can be used to locate them and is associated with someone who makes use of a credit or debit card (CD). The primary account number (PAN), as well as other information about the cardholder such as their name, expiration date, or service code, is contained inside the cardholder data. A data set is a term that refers to a collection of data (or dataset). In the case of tabular data, a data set is equivalent to one or more database tables.
Credit card secondary data are gathered because they are private information. Data is confidential and cannot be given in public for safety reasons. Principal Component Analysis is a method for reducing the proportions of large data sets, and one of its primary applications is to shorten the length of these sets. This is accomplished by paring down the number of variables used in the analysis while ensuring that the vast majority of the information contained in the original set is preserved.
With the help of principal component analysis, we are able to condense the information that is found in enormous data tables by utilising a reduced number of summary indexes that are both simpler to present and understand,
The accessible datasets are listed in the following table.

Table 1: Datasets for CCFD

S. No.	Website	Link	Dataset
1	Kaggle	https://www.kaggle.com/datasets/kartik2112/fraud-detection/download?datasetVersionNumber=1	European
2	Data.world	https://data.world/vlad/credit-card-fraud-detection/file/CC.csv	Artificial
3	OpenML	https://www.openml.org/d/31	German
4	Github	https://github.com/stedy/Machine-Learning-with-R-datasets/blob/master/credit.csv	Not specified

The transactions made using credit cards by European cardholders in 2013 are included in the dataset provided by Kaggle. During two days, there were a total of 284,807 transactions that took place, the dataset discovered 492 instances of fraudulent activity. The dataset has a lot of inconsistencies, with frauds making up only 0.172% of all transactions.
It only takes numeric values as the input, and those variables have been transformed using PCA. Because of concerns about maintaining the data's confidentiality, the data's original attributes and any further context are not provided. The primary components obtained by PCA include the functions V1, V2, and so on through V28. Time and Amount are the only two aspects that have not been changed in any way. Data set found in data.world is an artificial dataset created for the thesis adaptive machine learning [4] for CCFD.
The German Credit dataset classifies individuals as having either favourable or unfavourable credit risks based on various characteristics. It consists of 20 labels, such as credit history, credit purpose, and age, among others.

4. LITERATURE REVIEW :

The following is a quick review of earlier literature related to CCFD. The table lists some of the scholarly literature found on ML.

Table 2: Scholarly literature on machine Learning

S. No.	Area and Focus of the Research	Contribution	Reference
1	Using ML for CCFD The focus is using SMOTE on imbalanced dataset	Algorithm Random Forest is the most effective.	Varmedja et al., (2019). [5]
2	A Machine Learning Survey on Detecting CCF. The focus is on credit card safety.	Machine learning outperforms previously used prediction, clustering, outlier detection, and other techniques.	Popat et al., (2018). [6]
3	Detection of fraudulent activity on credit cards using ML. Focus is to extract the behavioural patterns	The Random Forest method yields the results of the highest possible quality.	Dornadula., et al., (2019). [7]
4	Detection of fraudulent activity on credit cards in real time using machine learning. Focus is on four basic instances of fraud in real world transactions	Utilising resampling approaches has a significant impact on making a classifier to perform comparably better.	Thennakoon et al., (2019).[8]
5	Detection of credit card fraud using skewed data Focus is on improvements in fraud detection	Metrics recall and accuracy were increased under sampling the data and training the model on it.	Mishra, et al., (2018). [9]
6	Procedure for Oversampling Determined by the Classification. The focus is on SMOTE to generate synthetic samples for each positive sample.	Accuracy, F1-score, AUC, and ROC metrics provide evidence of the high performance of OS-CCD.	Jiang et al., (2021).[10]
7	Algorithms for Intrusion Detection Systems Using Machine Learning. Focus is on performance measurements for false negatives and false positives to increase the intrusion detection system's detection rate	There isn't a single machine learning method that can effectively tackle all the different forms of attacks. The random forest classifier attained the best rate of accuracy.	Almseidin et al., (2017). [11]
8	Detection of CCF using ML techniques. The focus is on the effectiveness of naive bayes, k-nearest neighbour, and logistic regression on skewed data.	Except for accuracy, all measures considered demonstrate substantial performance for KNN.	Awoyemi et al., (2017). [12]
9	Detection of CCF using Bayesian and neural networks. The focus is on comparison of two algorithms.	The training time for Bayesian networks is shorter, and they produce better results, however ANNs have a speedier fraud detection process.	Maes et al., (2002). [13]
10	Algorithms For Supervised Machine Learning Review Focus is on analysis of supervised machine learning methods of different types	Depending on the specific of the issue, the appropriate algorithm should be used.	Kumar et al., (2017). [14]

Varmedja, D., et al. [5] provides a number of different algorithms for classifying transactions as either fraudulent or legal. Oversampling was performed using the SMOTE method since the dataset had a

significant imbalance in its distribution of values. Performance metrics accuracy, precision, recall were used to assess the algorithms.

Popat, R. R., et al.,[6] discuss factors behind usage of credit card and different algorithms used. The protection of credit card transactions is the primary objective, with the secondary goal being the simplicity and confidence with which users can access online banking. Identifying the fraudulent activity on credit cards is accomplished through the utilisation of a wide variety of techniques, such as Deep Learning [15], Logistic Regression, Naive Bayesian, Support Vector Machine (SVM), Neural Network [16], Artificial Immune System [17], K Nearest Neighbor, Data Mining, Decision Tree [18], and so on.

Dornadula, V. N., et al. [7], proposed the design and implementation of a fraud detection technique for streaming transaction data with the intention of analysing the transactional history of clients and extracting behavioural patterns. The transactions performed by cardholders from various groups are then combined using the sliding window approach so that the behaviour pattern [19] of each group may be retrieved separately. After that, the groupings are utilised to train [20] many classifiers on an individual basis in the future. After this, the classifier that received the best rating score can be chosen as one of the most efficient approaches to predict illegal transactions.

Thennakoon, A.,et al. [8] elucidate machine learning models to address fraud, and the most effective approach. Fraudulent transactions [21] can take different forms and fall under several categories [22]. The evaluation serves as a manual for choosing the best algorithm for the particular fraud type [23], and it illustrates the evaluation with the proper performance metric [24]. Real-time CCFD is another crucial topic.

Awoyemi, J. O., et al. [12] discuss the dataset [25] sampling strategy, variable choice, and the kind of fraud detection that is used all have a substantial impact on the level of success that is achieved in detecting fraudulent activity involving credit card transactions. On the skewed data, a hybrid under-sampling [26]/over-sampling technique is used. Accuracy [27], sensitivity [28], specificity [29], precision, Matthews correlation coefficient [30], and balanced classification rate [31] are utilised in order to evaluate the effectiveness of the approaches.

Maes, S., et al.,[13] presents two machine learning algorithms that are used when there is uncertainty. For financial organizations [32], detecting CCF is crucial. Using artificial neural networks and Bayesian belief networks, it is possible to display the findings of issues found in real world financial data. This illustrates how several ML techniques, such as LR, NB, and RF with classifiers utilising the boosting technique, can be implemented on an imbalanced dataset [33]. A comprehensive analysis and assessment of the existing and planned models for detecting fraudulent activity on credit cards, as well as a study comparing these various methods, has been conducted. Therefore, several classification models are applied to the data, and the performance of the models is evaluated using quantitative metrics such as accuracy, precision, recall [34], f1 score [35], and confusion matrix [36]. The findings of the study indicate that the best classifier can be developed by training [37] and testing [38] with the assistance of supervised methods, which results in a better solution [39].

The SMOTE [40] approach was applied for oversampling because the dataset contained imbalances. In addition to this, a procedure for selecting features was carried out, and the dataset was then split into training and test data. During the experiment, the following algorithms were utilised: logistic regression, random forest, naive bayes, and multilayer perceptron [41]. The results demonstrate that each algorithm is highly accurate for detecting credit card fraud. The following table lists some of the scholarly literature found on DM.

Table 3: Scholarly literature data mining

S. No.	Area & Focus of the Research	Contribution	Reference
1	An Efficient Algorithm for an Incremental Mining ie sliding window Focus is on I/O, CPU cost, and memory utilization control	By using the concepts of cumulative filtering, SWF lowers I/O and CPU costs. It manages memory usage using the sliding-window partition technique.	Lee (2001). [42]
2	Data mining-based intrusion detection system	EDADT algorithm minimises the dataset's real size and	Nadiammai et al., (2013). [43]

	The focus is on IDS linked with data mining.	analyses current threats effectively and with a lower false alarm rate.	
3	Imbalanced datasets used in data mining Focus is on The performance metrics and sampling methods for mining imbalanced datasets	When evaluating classifiers learned on imbalance data sets, accuracy is not a useful metric. Replication can make the decision region that determines the classification of the minority class smaller.	Chawla (2009). [44]
4	Detecting CCF using ML As A Data Mining Method. Focus is on combining the DM and ML approaches	Bayesian classifiers generated considerably better results with filtered data.	Yee et al., (2018). [45]
5	Neural Networks and Data Mining Technique The focus is to identify the optimal classifier for the model.	On both fronts, RF performs significantly better than any other classifier than the average	Sahu et al., (2020).[46]
6	Data Mining Application. Focus is to detect fraud in realtime	Identify sudden deviations in established patterns and identify fraud's general usage patterns	Akhilomen (2013). [47]
7	Detecting credit card fraud in online retail that uses data mining. The focus is combining automatic and manual categorisation	Random Forest did better than LR and SSVM.	Carneiro et al. (2017).[48]
8	Examining the different DM and ML Techniques. The focus is on a review of different methods for DM and ML.	Innovative algorithms should be employed instead of conventional data mining techniques to identify fraudulent transactions.	Patil et al., (2018). [49]
9	Data Mining Techniques Applied to an Analysis of Multiple Distributions for Spotting Credit Card FraudFocus is on under sampling, cross validation,	Comparatively, RF outperforms NB, SVM, and KNN.	ATA et al., (2020). [50]
10	The banking industry is using the DM technology to detect fraud in real time. Focus is on patterns that might result in fraud	two clients who just so happen to share the same characteristics will undoubtedly act the same	John et al., (2016). [51]

Lee, C., et al. [42], explain characteristics of sliding window filtering, utilises the ideas of cumulative filtering and scan reduction techniques to lower I/O and CPU costs while also efficiently regulating memory usage through the use of sliding-window partition. An ongoing time-variant transaction database can benefit greatly from effective incremental mining with the SWF algorithm.

Nadiammai, G. V. [43] describe the essentiality of the Intrusion Detection System (IDS) in spotting network irregularities and threats. IDS and the data mining idea are combined to quickly and efficiently find the user-relevant concealed data that will be of interest.

Chawla, N.,[44], analyse using ML methods to solve complex, real-world problems, many of which have data that is skewed. Distribution of the testing data may not match that of the training data, and the real costs of misclassification might not be understood at the time of learning. When the data is unbalanced or the prices of various errors vary significantly, predictive accuracy, metric for measuring classifier performance, may not be acceptable.

Yee, O. S., et al.,[45] explain how Bayesian network classifiers K2 [52], Tree Augmented Naive Bayes (TAN) [53], Naive Bayes, Logistics, and J48 [54] are used in the research to describe supervised based classification. After the dataset was preprocessed using normalization and Principal Component Analysis [55], all classifiers beat the results obtained without preprocessing the dataset by more than 95.0 percent.

Sahu, A., et al.,[46], explain addressing the underlying issue of data imbalance using two separate methods. To enhance the total number of samples coming from the minority class [56], the first technique employs data resampling [57], in contrast, the second technique employs a cost-based strategy, where additional weight [58] is added by the error function for each class. The weights enable the samples of fraudulent transactions to be given greater weight than the regular samples.

Akhilomen [47], proposed to conduct the best categorization of each transaction into its related group in the absence of a prior output, the self-organizing map neural network (SOMNN) [59] technique. In contrast to previous statistical models and the two-stage clusters, the receiver-operating curve (ROC) [60] [61] of the CCFD watch could detect over 95% of fraud instances without producing any false alarms.

Carneiro, N. [48], explain creation and the establishment of a system for detecting fraud in a significant online retailer. Mix of manual and automatic classification, which evaluates various machine learning techniques and provides insights throughout the development process. It assists academics and practitioners in the design and deployment of DM-based solutions for fraud detection and other related challenges. This helped the fraud analysts improve their manual revision procedure, leading to a good result.

Patil, V. et al. [49], express method of examining different credit card users behaviour from historical transaction history databases, fraud transactions can be detected. Any deviation in expenditure patterns from the previously observed patterns may point to the presence of fraudulent activity. Techniques like data mining are frequently employed in the identification of credit card fraud.

Oguz A.T.A. Et al. [50] discuss about confusion matrix and Area Under the Curve ranking measure to compare models in a skewed dataset in order to identify which would be the best model for fraud detection. The findings for the NB, SVM, KNN, and RF classifiers show the highest accuracy, with scores of 97,80%, 97,46%, 98,16%, and 98,23%, respectively. Four-division datasets (75:25, 90:10, 66:34, and 80:20) were used for comparison, and the findings showed that the RF outperforms NB, SVM, and KNN.

John, A., et al. [51], explains regarding using data-mining techniques, bank fraud can be detected by examining client data to spot trends that could indicate fraud. Once the patterns are discovered, banking procedures can be given a greater level of verification and authenticity. Following table lists some of the scholarly literature found on CCFD.

Table 4: Scholarly literature on credit card fraud

S. No.	Area and Focus of the Research	Contribution	Reference
1	Imbalanced Classification Methods for CCFD. Focus is on imbalance of data	When the imbalance is extreme, the methods typically utilised to correct imbalance problems may have undesirable outcomes**.**	Makki et al., (2019). [62]
2	Machine Learning and Data Science for CCFD. Focus is on detecting all fraudulent transactions while reducing erroneous fraud categories	To improve the final result's accuracy, multiple algorithms might be joined as modules and linked together.	Maniraj et al., (2019). [63]
3	Using an Optimised Light Gradient Boosting Device for CCFD. Focus is on utilizing an enhanced light gradient boosting machine for detecting fraud in credit card transactions (OLightGBM**)**	Compared to other machine learning approaches, the suggested method achieved the highest levels of accuracy, AUC, precision, and F1-score.	Taha et al., (2020). [64]

4	Using Weighted Extreme Learning Machines for imbalanced classification in CCFDs. Focus is to Optimize a Weighted Extreme Learning Machine using various clever optimization techniques.	Better classification performance can be attained using the three improved WELMs.	Zhu et al., (2020). [65]
5	Awareness and prevention of credit card fraud The focus is on being conscious of the occurrence of credit card fraud	More measures must be implemented to stop and identify credit card theft.	Barker et al., (2008). [66]
6	Detection of credit card fraud using data analytic methods. Focus is on anticipate the identification of fraud	The KNN method outperformed logistic regression in terms of accuracy value.	Vengatesan et al. (2020). [67]
7	Utilizing Feature Selection Method-Based Adaptive Credit Card Fraud Detection Techniques. Focus is on feature selection strategies for CCFDs at the application level.	The J48 classifier and PART classifier prediction accuracy has improved by 4% and 2%, respectively. Additionally, the J48 classifier, AdaBoost classifier, and random forest all gotten better in terms of precision.	Singh et al., (2019). [68]
8	Utilizing specific ML methods to identify CCF. The focus is on how well ML algorithms perform?	There is barely any performance difference between logistic regression and random forest. LR displays superior outcomes in an incremental setup	Puh et al., (2019). [69]
9	CCFD using real-time data-driven methods Focus is on anomaly detection using real-time data	The T2 control chart and one-class support vector machine's results demonstrate that the strategy had a low rate of false alarms and high detection accuracy.	Tran et al. (2018). [70]
10	CCFD Using Gradient Boosting Techniques. Focus is on a variety of supervised and unsupervised machine learning methods for identifying fraudulent behaviour.	In terms of accuracy, the CatBoost Model is efficient, When used with huge datasets, Light Gradient Boosting Machine is effective.	Sarkar et al., (2022). [71]

Makki, A.Z., et al. [62] compared the efficacy of the numerous imbalanced categorization algorithms in the presence of extreme imbalance. According to the performance measures taken into consideration, it was discovered that the LR 70, C5.0 decision tree algorithm, SVM 71, and ANN are the best approaches.

Maniraj, S.A., et al.,[63] focused on data set analysis and preprocessing, as well as the application of several anomaly detection techniques to PCA-converted Credit Card Transaction data, as well as the Isolation Forest methodology and the Local Outlier Factor.

Taha, M.S.,[64] illustrates the efficiency of an optimised light gradient boosting machine for identifying fraud in credit card transactions, using tests on two actual, publicly available credit card transaction data sets that included both valid and fraudulent transactions.

Zhu H ,L.G.,[65] explains during situations involving imbalanced classification, three enhanced dandelion algorithms with probability-based mutation are proposed, along with three optimised WELMs. It is also discussed Application of proposed algorithm to the detection of credit card fraud

Barker, K. J.,[66] explains Fraud relies on already available technology and the simplicity of obtaining equipment to steal people's identities and account information and create fake credit cards. The way technology evolves has an impact on fraud prevention as well.

Vengatesan, K.,[67] presents a model which is based on a classification algorithm, and also based on a training and test set of data, performance is assessed using logistic regression and the KNN72 [72] algorithm.

Singh, A. [68] proposes CCF at the application level, the study compares and examines the performance of five ML techniques. For the performance evaluation, the key metrics are considered.

Puh, M., [69] examines how well three machine learning algorithms—Random Forest, Support Vector Machine [73], and Logistic Regression [74]—perform at spotting fraud using real-world data that contains credit card information.

Tran, P.,[70], proposes using one class SVM, the optimal kernel parameter selection, and T2 control charts for data-driven approaches for CCFD. Algorithms' efficacy is measured against a dataset comprising actual e-commerce transactions conducted online.

5. CURRENT STATUS & NEW RELATED ISSUES :

The effectiveness of a categorization system can be summarised in the form of a table that is known as a confusion matrix. A confusion matrix is used to display the results of a classification system and provide a summary of those results.

Table 5: Confusion Matrix

		True Class	
		Positive	Negative
Predicted class	Positive	TP	FP
	Negative	FN	TN

The total number of accurate outcomes or hypothesis where the actual class was positive is known as the true positive rate (TP). The frequency of inaccurate findings or forecasts made while the actual class was positive is known as the false positive rate (FP). The total number of accurate outcomes or forecasts, when the actual class was negative, is known as the true-negative rate (TN). The total number of inaccurate outcomes or forecasts made when the actual class was negative is known as the false negative rate (FN). FP and FN should be minimized

A false positive occurs when something is found to be true when it is actually false. Type I error is another name for FP. A false negative is when something is labeled as true when it is actually false. This is called as Type II error.

The listing of a legitimate user as a fraudster by algorithms that is FP may cause numerous issues. Denial of service to a trustworthy customer could damage the financial organization's reputation. As a result, the business risks losing a real customer. Non-identification of the FN fraudster may result in financial loss for the organisation. Following Table 6 shows Metrics used:

Table 6: Performance Metrics used ML

S. No.	Metric	Formula
1	Accuracy	TP+TN/(TP+FP+FN+TN)
2	Precision	TP/(TP+FP)
3	Recall (sensitivity)	TP/(TP+FN)
4	Specificity	TN/(TN+FP)
5	F1 Score	2 * Precision * Recall / (Precision + Recall)

The accuracy ratio is the number of accurately predicted observations to all observations. When we have symmetric datasets with almost equal values for false positives and negatives, accuracy is a good metric.

The term precision refers the proportion of positive observations that were accurately predicted, compared to the total number of positive observations that were anticipated. The precision can be determined by dividing the number of genuine positive results by the sum of the true and false positive results.
Recall is the ratio of accurately predicted positive observations to all of the actual class's observations. The recall estimates how well the model can identify positive samples. More positive samples are found when the recall is higher.
The specificity of a classifier is measured by the proportion of correctly classified negative data to the total number of actual negative data. The capacity of a test to appropriately exclude is known as specificity.
When added together, Precision and Recall make up what is known as the F1 Score. In calculating this score, both false positives and false negatives are taken into account. It is not as simple to understand as accuracy. The F1-score is a fair metric that accurately measures how accurate models are across different domains.

5.1 Major types of Credit Card Fraud Types:

(i) ***Card Not Present Fraud:*** When someone uses a debit or credit card to make an online purchase without the owner's consent, this is referred to as card not present fraud (CNP fraud). The most prevalent CNP fraud is E-commerce fraud.
Telephone and online transactions can begin without a physical payment card being present. Instead, all that is required from the customer to complete the transaction is the card number and other easily accessible information.
However, the phrase applies to any credit card transaction in which the card is not physically provided to the merchant to complete the transaction, including phone fraud and mail order fraud. Most credit card frauds involve card not present fraud. When a fraudster has access to genuine information, CNP fraud protection is difficult [75].

(ii) ***Skimming:*** The use of a skimmer to obtain credit card information is a form of theft known as skimming. Skimmer is card reader hardware that is installed and gathers card numbers. Fraudster may modify ATM equipment to collect magnetic stripe information from used cards and the PINs associated with cards. A fraudster can wirelessly acquire stolen information by simply being there and having access to a PIN.
Skimming is challenging to catch. Often, skimming is discovered unexpectedly. Businesses may become suspicious of fraud if their revenue is lower than anticipated [76].

(iii) ***Phishing:*** Phishing is when someone impersonates a trustworthy website, email, or message to gain sensitive financial information such as credit card numbers. A lot of the time, it might be difficult to determine whether these links and websites can be trusted.
Attacks involve sending spoof emails to users to get them to respond with their login information and password. Even while such assaults may be successful right now, from the attackers' perspective, the success rate is reduced because many users have been educated not to transmit important information via email.
Certain phishing scams are more technologically sophisticated and utilise well-known flaws in widely used web browsers in Internet, to gather sensitive data about the victim [77].

(iv) ***Lost or Stolen card***: A card could be taken by a thief or could be accessed from an unsafe location. Until the cardholder learns of the loss and blocks the card, the card may be mistreated and used without permission. As cards are kept in wallets that include all of the card's information, a thief who steals a card may have immediate access to the information. [78].

(v) ***Application fraud:*** Application fraud occurs when a dishonest person submits an application for a loan or line of credit using fraudulent or stolen identification with the intention of not paying back the money borrowed from the lender.
Once the application has been processed, the thief uses the new credit card to make large withdrawals of money, leaving the person whose identity was stolen accountable for making payments on the loan.

Even though the victim is irresponsible for the phony debt, it is possible that they may not be discovered as a victim of this form of fraud until after they have already paid off the money. This presented a significant risk to the victim [79].

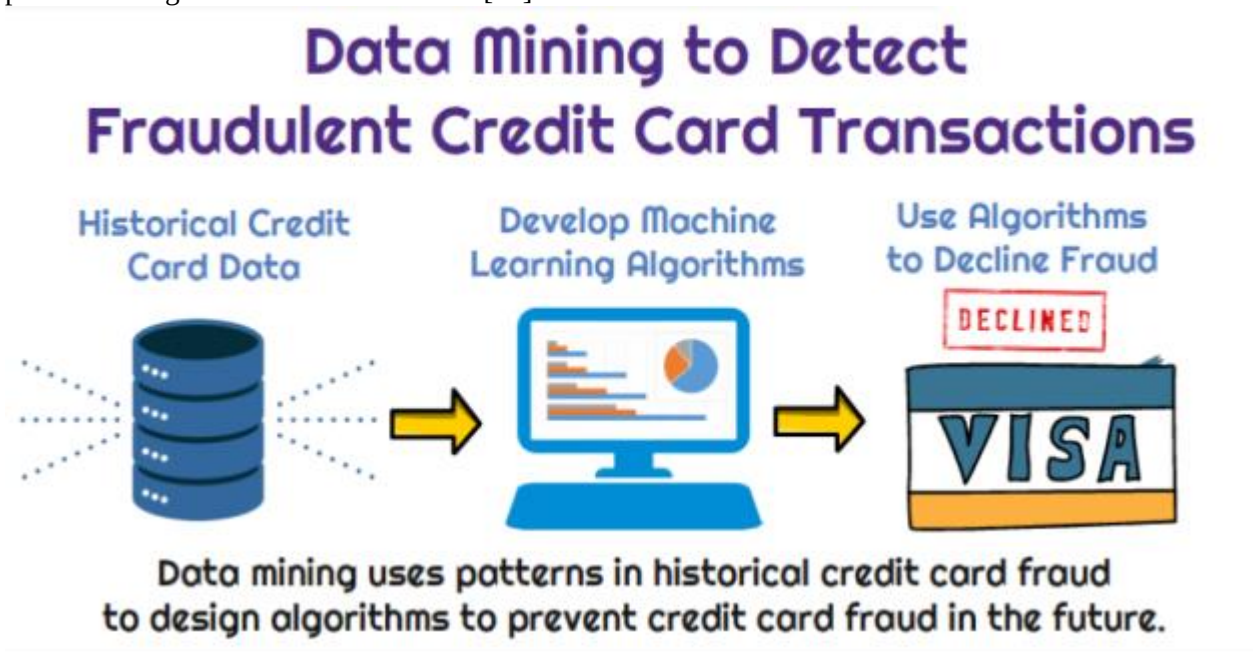

Fig. 1: image source: www.universalcpareview.com [80]

6. IDEAL SOLUTION, DESIRED STATUS & IMPROVEMENTS REQUIRED :

KNN, Logistic Regression, SWM, Random Forest, and Neural Networks are a few algorithms from which conclusions are drawn after evaluating their metrics. To make a choice, we must evaluate all the algorithms using the dataset.

Online transactions are a major source of difficulty for fraudster detection, and metrics alone cannot solve this problem. We must devise a technique, which could be a hybrid algorithm or a single algorithm, that provides the best possible outcome. Algorithms for online machine learning must be developed.

When working with unbalanced data set in CCFD, accuracy is not a good indicator of performance. A good measure is the F1 score. When calculating the F1 measure, precision and recall metrics are both considered. Our main problem with confusion matrices is False Positive and False Negative results. Predicted as Fraduster, but the customer is Legitimate user (FP) and Predicted as Legitimate, the user is Fraduster (FN) needs to be much less. Precision considers both FP and TP. To compute recall, FN and TP are required. If the algorithm defeats FP and FN, then F1 should be equal to 1 or close to 1.

7. RESEARCH GAP :

After performing a literature review to date, the following research gaps were identified:

(i) Credit card fraud can be detected through various ML and DM technologies. Without considering all the other algorithms, a comparison of a small number of algorithms shows that random forest outperforms the others.

(ii) Metrics used to evaluate algorithms are offline, they will be used to assess algorithm performance after the fraud has occurred. By the time metrics identify the fraudster, the fraudster may have engaged in several further instances of deception. Anti-fraud technology that works in real time is an absolute must.

(iii) When PCA is applied to a dataset, information is lost and improperly understood.

8. RESEARCH AGENDAS BASED ON THE RESEARCH GAP :

(1) Study of different ML and DM algorithms used in CCFD and their performance.
(2) Comparison of confusion matrix results with respect to different algorithms in ML and DM.
(3) To study datasets available and find features important in data sets in a large data set.

(4) Study different tools available to compare algorithm performance.

9. ANALYSIS OF RESEARCH AGENDAS :

Machine learning may be broken down into four main categories, which are supervised [81], semi-unsupervised [82], semi supervised [83], and reinforcement learning [84]. Each of these subcategories has its own unique set of advantages and disadvantages. Supervised ML algorithms are used to train algorithms to correctly classify data or to anticipate outcomes using named data sets. Using training data, unsupervised machine learning technique applies models. Models instead use the provided data to uncover hidden patterns and insights. Both labeled and unlabeled data are utilized in semi-supervised learning. Labeled data has important tags so the computer can comprehend it, whereas unlabeled data does not. The machine learning algorithm is provided with a list of operations, parameters, and output values. After generating the rules, the ML algorithm seeks to explore multiple options and possibilities, evaluating each result to determine which the best is.

9.1 Supervised learning algorithm:

Naive Bayes: According to the Naive Bayes classifier, which is based on the Bayes theorem, each value is classified as independent of every other value. It enables us to make predictions about groups or categories with probability while utilising a certain collection of features.

Support Vector Machine: They effectively categorise the data by providing a collection of training examples, each of which is labelled to indicate that it belongs to one of the two categories. The method then creates a model by assigning new values to one or more categories.

Linear Regression: The use of simple linear regression makes it feasible to understand the relationships that exist between two continuous variables. It is possible to make an accurate prediction about the value of one variable by employing linear regression analysis and basing that guess on the value of another variable.

Decision Trees: It is a tree structure resembling a flowchart that employs branching to show every decision's potential results. Every node in the tree represents a test on a particular variable, and every branch reflects the result of that test Logistic

Regression: It is concerned with calculating the likelihood of an event happening based on the preceding information given. It is utilised to deal with dependent variables that are binary, that is, have just two possible values, 0 and 1.

Random Forests: The ensemble learning technique uses multiple algorithms to provide superior classification, regression, and other outcomes. Each classifier works poorly on its alone but excels when paired with others.

Nearest Neighbours: The K-Nearest-Neighbor method determines the likelihood that a data point falls into a particular category. In order to determine which cluster a given data point belongs to, it looks at neighbouring data points.

9.2 Unsupervised learning algorithms:

K Means Clustering: The method finds groups within the data, with the variable representing the number of groups. Using the given attributes, it then works iteratively to categorise each data point into one of K groups. To separate groups with similar qualities and group them into clusters is the goal. The number of groups found in the data using the k-means method is denoted by the variable "k."

Independent Component Analysis: To separate distinct sources from a mixed signal, ICA is utilised. In contrast to PCA, which emphasises maximising data point variance, independent component analysis emphasises independence, or independent components. According to the ICA, the independent components must have a non-gaussian distribution and are presumed to be statistically independent of one another.

Apriori algorithm: The phrase "apriori algorithm" refers to the algorithm that establishes the principles of item association. The development of an association rule between a number of different items is the fundamental purpose of the apriori algorithm. The association rule provides an explanation of the relationships that exist between two or more objects. The Apriori algorithm is sometimes referred to as frequent pattern mining in some circles.

Singular Value Decomposition: The factorization of a single matrix into three distinct matrices. Important geometrical and theoretical ideas regarding linear transformations are communicated by it and possesses a number of noteworthy algebraic properties. Additionally, it has several significant applications in the field of data science.

Hierarchical Cluster Analysis: Hierarchical cluster analysis is the technique used to cluster unlabeled data sets. The dendrogram is a tree-like structure that depicts the existing cluster hierarchy. K-means clustering and hierarchical clustering can provide results that resemble one another on occasion, despite the fact that they operate differently. As with the K-Means technique, there is no prerequisite that states that the number of clusters must be determined in advance.

9.3 Semi supervised learning algorithms:

It uses both supervised and unsupervised learning methods in a hybrid fashion. labelled and unlabelled data are present in the training data. But there is a relatively tiny amount of labelled data, and there is a vast amount of unlabeled data. An unsupervised learning method is first used to cluster similar data, and it also helps label the unlabeled data into labelled data by turning it into clusters of similar data. This explains why labelled data is more expensive to acquire compared to unlabeled data.

9.4 Reinforcement learning algorithms:

Artificial Neural Networks: The way that biological systems like the brain process information is an inspiration for artificial neural networks (ANNs). ANNs are simply a vast collection of interconnected processing components that cooperate to address particular issues. It is a computational model made up of various processing components that, in accordance with their established activation functions, take inputs and produce results.

State–action–reward–state–action with eligibility traces: (SARSA – Lambda): SARSA It is a learning algorithm for Markov decision-making procedures. SARSA employs the ON-policy learning method, in which the agent picks up new information from the agents' most recent set of actions. The SARSA algorithm is autonomous from the prior learning or greedy learning policy because there is no maximum operation that is performed in it.

Q-learning algorithm: The Q-learning algorithm uses model-free reinforcement learning to determine the worth of a given action in a given situation. Without the need for modifications, it is able to solve problems associated with stochastic transitions and rewards and does not require an environment model. It essentially amounts to a dynamic programming approach that is gradual and places modest processing demands.

Deep Deterministic Policy Gradient: It is a method that depends on optimising parametrized rules in light of the anticipated return. Deep Q Learning (DQN) and DPG are combined in the model-free off-policy actor-critic approach known as DDPG. Given that it produces the action immediately, the policy is deterministic. The policy-determined action is supplemented with some Gaussian noise to encourage exploration.

Asynchronous Advantage Actor-Critic Algorithm: A3C is made up of numerous autonomous agents, each with its own weights, interacting concurrently with a different copy of the environment. As a result, they can quickly and efficiently explore a larger portion of the state-action space. Agents engage in environment-specific interactions asynchronously, gaining knowledge from every conversation. The learning agent modifies the optimal policy function using the value of the Value function. Agent will be made aware of the acts that were rewarded and those that were penalised.

Most Popular Machine Learning Software Tools are Scikit-learn, PyTorch, TensorFlow, Weka, KNIME, Colab, Apache Mahout, Accord.Net, Shogun, Keras.io, Rapid Miner, etc.

10. FINAL RESEARCH PROPOSAL IN CHOSEN TOPIC :

A comparative study of credit card fraud detection using various ML and DM techniques and modification to existing approaches.

11. SWOT ANALYSIS :

A SWOT [85] analysis is useful for evaluating both the internal as well as external factors that may have an impact on a business. A capability or resource that an organisation can successfully utilise in order to meet its goals is referred to as strength of that organisation. A company's weaknesses [86] are

the result of internal variables. The process of locating these can assist in the localization of potential development areas. It enables organisations to design solutions to correct and control their problematic areas, which in turn helps them develop. Opportunities [87] can be defined as external variables that are open and available for the organisation to use to its advantage in order to benefit from them. Threats [88] enable organisations to gain visibility into their shortcomings as well as potential areas for development.

11.1 Strengths:

Learning by machine occurs without human intervention: A computer is responsible for the entirety of the data interpretation and analysis processes, both of which are carried out on the data. Prediction or interpretation of the data does not require the involvement of humans. Machines begin to learn and predict which algorithm or program will produce the greatest results as part of the overall machine learning process.

Improved fraud detection accuracy: Machine learning techniques are more precise and produce more relevant results compared to rule-based solutions since they take many additional elements into account. This is because ML algorithms can consider a vast number of extra data points, such as minute features of behaviour patterns associated with a particular account.

Multiple data formats are supported: In a fluid and uncertain setting, It is capable of managing a diverse range of data types. It has a wide range of capabilities and operates at multiple levels.

Applied to different areas: It finds use in many aspects of modern life, including healthcare, engineering, teaching, and more. Data-analysis and prediction tools can range from the simplest of programs to the most complicated of structured machines.

11.2 Weakness:

Require a massive dataset: Large data sets are suitable for machine learning techniques. The results might not be perfect with little data. Large data sets are necessary for accurate machine learning of models. This data volume is a problem for smaller firms, but it is not for big organisations.

Noise in data: There could be a huge amount of noise in the data we get. The algorithm needs to be able to tell the difference between authentic user input and skewed or made-up user input.

Unbalanced Data: The data used to detect credit card fraud are unbalanced. Few fraudulent credit card transactions occur overall. Due to this, fraud transaction detection is exceedingly challenging and inaccurate.

Difficult to gather reliable data: For a machine learning model to be effective over the course of its lifetime, It is essential to get a significant amount of data. The information that is entered will be of high quality, which is something that rarely occurs in real life.

Unstandardized metrics: There is no one set of evaluation criteria that can be applied across the board when comparing and contrasting the results of different fraud detection systems.

The financial burden of fraud detection: The system ought to consider, at the same time, both the cost of the detected fraud as well as the cost of the prevention of further fraud.

11.3 Opportunities:

To prevent or identify fraudulent conduct, model creation processes can be used in industries related to commerce, such as the insurance and telecommunications sectors.

Gathering and organising raw data, which is subsequently utilised to train the model to forecast the likelihood of fraud.

Modern data-driven statistics and machine-learning (ML) techniques can produce prediction models with statistical resemblances that output the likelihood that a transaction is fraudulent and can solve the aforementioned problem.

Machine learning also improves with experience and grows more accurate and effective at work, just like humans do. This leads to decisions that will be better. As more data is gathered, a machine can learn more, and it can also understand patterns and trends which can be used in business.

11.4 Threats: (Challenge)

False positives are instances in which legitimate transactions are incorrectly suspected of being fraudulent when, in fact, they are legitimate (and vice versa). Transactions that are not genuine might

also give the appearance of being legitimate (false negative). As a result, obtaining low rates of false positives and false negatives is one of the key challenges that faces fraud detection systems.
Illegitimate transactions are less often compared with legitimate transactions. This type of data is called skewed data. Working with skewed data is difficult.
Imbalanced data are the result of skewed data. To balance dataset synthetic minority oversampling technique is used, which may result in overlapping and more noise in the data.
Working with a large dataset is a tedious task and data is confidential that result in Principal Component Analysis. PCA have its own assumptions like correlation between features, missing values etc.
Different models are used to test fraud detection. Each model uses a different algorithm. The detection of new frauds will be difficult.
Real life data contains privacy and sensitive and is not easily gettable because of the risk involved in it. Institutions may be held responsible for data leakage.
Data mining techniques will need more time to deal with a huge amount of data.
Building a proper, error-free machine learning model is essential. To do that, businesses require educated professionals with experience of creating such systems and in-depth knowledge of the peculiar area of payment fraud.
Inflexibility: Recognizing new kinds of valid or fraudulent patterns is a regular difficulty for classification algorithms. This is because of the lack of flexibility in their design. Both supervised and unsupervised fraud detection algorithms find it challenging to recognise new patterns of both normal and fraudulent behaviour.

12. SUGGESTIONS TO IMPLEMENT RESEARCH ACTIVITIES ACCORDING TO PROPOSAL :

(1) Different forms of attack are effectively tackled by different types of algorithms. A single algorithm may unable to handle all types of attacks.
(2) The F1 score needs to be considered compared to other metrics for performance evaluation.
(3) Customers sharing the same characteristics will act in the same way.
(4) Innovative algorithms work better than conventional data mining techniques.

13. CONCLUSION :

Today's need is for digital money transactions. There has been an increase in card usage and digital money transactions. A rise in cybercrime is seen along with an increase in digital money transactions. The key focus in preventing cybercrime is the identification of unauthorised users. Emphasis is placed on the machine learning techniques utilised in the identification of CCF. The many ML algorithms employed in CCFD are the main topic of this study. The same algorithms are utilised in data mining to identify fraud. The paper also focuses on measures for CCFD algorithm evaluation. CCFD is a real-time necessity, and the model should be accurate to catch fraud early. The dataset used, training data, and testing data are also emphasised. A confusion matrix is used to assess metrics and, ultimately, algorithm performance**.**

REFERENCES :

[1] Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision support systems*, *50*(3), 602-613. Google Scholar↗

[2] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260. Google Scholar↗

[3] Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS quarterly*, *45*(3), 1433-1450. Google Scholar↗

[4] Ashok Kumar, D., & Venugopalan, S. R. (2018). A novel algorithm for network anomaly detection using adaptive machine learning. *Progress in Advanced Computing and Intelligent Engineering*, *564*(1), 59-69. Google Scholar↗

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

[6] Popat, R. R., & Chaudhary, J. (2018). A survey on credit card fraud detection using machine learning. *2nd international conference on trends in electronics and informatics (ICOEI)*, *1*(1), 1120-1125. Google Scholar↗

[7] Dornadula, V. N., & Geetha, S. (2019). Credit card fraud detection using machine learning algorithms. *Procedia computer science*, *165*(1), 631-641. Google Scholar↗

[8] Thennakoon, A., Bhagyani, C., Premadasa, S., Mihiranga, S., & Kuruwitaarachchi, N. (2019). Real-time credit card fraud detection using machine learning. *9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, *1*(1), 488-493. Google Scholar↗

[9] Mishra, A., & Ghorpade, C. (2018), Credit card fraud detection on the skewed data using various classification and ensemble techniques. *IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, *1*(1), 1-5. Google Scholar↗

[10] Jiang, Z., Pan, T., Zhang, C., & Yang, J. (2021). A new oversampling method based on the classification contribution degree. *Symmetry*, *13*(2), 1-13. Google Scholar↗

[11] Almseidin, M., Alzubi, M., Kovacs, S., & Alkasassbeh, M. Evaluation of machine learning algorithms for intrusion detection system. *IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY)*, *1*(1), 277-282). Google Scholar↗

[12] Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017). Credit card fraud detection using machine learning techniques: A comparative analysis. *International conference on computing networking and informatics (ICCNI)*, *1*(1), 1-9. Google Scholar↗

[13] Maes, S., Tuyls, K., Vanschoenwinkel, B., & Manderick, B. (2002). Credit card fraud detection using Bayesian and neural networks. *Proceedings of the 1st international naiso congress on neuro fuzzy technologies*, *261*(1), 270-277. Google Scholar↗

[14] Choudhary, R., & Gianey, H. K. (2017). Comprehensive review on supervised machine learning algorithms. *International Conference on Machine Learning and Data Science (MLDS)*, *1*(1), 37-43. Google Scholar↗

[15] Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, *187*(1), 27-48. Google Scholar↗

[16] Wang, S. C. (2003). Artificial neural network. In *Interdisciplinary computing in java programming*,*73(*1*)*, 81-100. Google Scholar↗

[17] Hofmeyr, S. A., & Forrest, S. (2000). Architecture for an artificial immune system. *Evolutionary computation*, *8*(4), 443-473. Google Scholar↗

[18] Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *18*(6), 275-285. Google Scholar↗

[19] Sarker, I. H., & Kayes, A. S. M. (2020). ABC-RuleMiner: User behavioral rule-based machine learning method for context-aware intelligent services. *Journal of Network and Computer Applications*, *168*(1), 49-58. Google Scholar↗

[20] Carney, M., Webster, B., Alvarado, I., Phillips, K., Howell, N., Griffith, J., Jongejan, J., Pitaru, A., Chen, A. (2020). Teachable machine: Approachable Web-based tool for exploring machine learning classification. *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, *1*(1), 1-8. Google Scholar↗

[21] Sadineni, P. K. (2020). Detection of fraudulent transactions in credit card using machine learning algorithms. *Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*,*1*(1), 659-660. Google Scholar↗

[22] Leite, R. A., Gschwandtner, T., Miksch, S., Kriglstein, S., Pohl, M., Gstrein, E., & Kuntner, J. (2017). Eva: Visual analytics to identify fraudulent events. *IEEE transactions on visualization and computer graphics*, *24*(1), 330-339. Google Scholar↗

[23] Shah, N., Lamba, H., Beutel, A., & Faloutsos, C. (2017). The many faces of link fraud. *2017 IEEE International Conference on Data Mining (ICDM),1*(1), 1069-1074. Google Scholar↗

[24] Gordon, M. L., Zhou, K., Patel, K., Hashimoto, T., & Bernstein, M. S. (2021). The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, *1*(1), 1-14. Google Scholar↗

[25] Pan, F., Wang, W., Tung, A. K., & Yang, J. (2005). Finding representative set from massive data. *Fifth IEEE International Conference on Data Mining (ICDM'05),1*(1), 1-8. Google Scholar↗

[26] Liu, X. Y., Wu, J., & Zhou, Z. H. (2008). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *39*(2), 539-550. Google Scholar↗

[27] Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, *16*(5), 412-424. Google Scholar↗

[28] Kalapanidas, E., Avouris, N., Craciun, M., & Neagu, D. (2003). Machine learning algorithms: a study on noise sensitivity. *Proc. 1st Balcan Conference in Informatics,1*(1), 356-365. Google Scholar↗

[29] Kapp, E. A., Schütz, F., Connolly, L. M., Chakel, J. A., Meza, J. E., Miller, C. A., ... & Simpson, R. J. (2005). An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. *Proteomics*, *5*(13), 3475-3490. Google Scholar↗

[30] Zhu, Q. (2020). On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset. *Pattern Recognition Letters*, *136*(1), 71-80. Google Scholar↗

[31] Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168-192. Google Scholar↗

[32] Nikou, M., Mansourfar, G., & Bagherzadeh, J. (2019). Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms. *Intelligent Systems in Accounting, Finance and Management*, *26*(4), 164-174. Google Scholar↗

[33] Mishina, Y., Murata, R., Yamauchi, Y., Yamashita, T., & Fujiyoshi, H. (2015). Boosted random forest. *IEICE TRANSACTIONS on Information and Systems*, *98*(9), 1630-1636. Google Scholar↗

[34] Junker, M., Hoch, R., & Dengel, A. (1999). On the evaluation of document analysis components by recall, precision, and accuracy. *Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR'99*, *1*(1), 713-716. Google Scholar↗

[35] Yacouby, R., & Axman, D. (2020, November). Probabilistic extension of precision, recall, and F1 score for more thorough evaluation of classification models. In *Proceedings of the first workshop on evaluation and comparison of NLP systems,1*(1), 79-91. Google Scholar↗

[36] Susmaga, R. (2004). Confusion matrix visualization. In *Intelligent information processing and web mining*, *25*(1), 107-116. Google Scholar↗

[37] Bhavsar, H., & Ganatra, A. (2012). A comparative study of training algorithms for supervised machine learning. *International Journal of Soft Computing and Engineering (IJSCE)*, *2*(4), 2231-2307. Google Scholar↗

[38] Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons. b*, *4*(1), 51-62. Google Scholar↗

[39] Bhanusri, A., Valli, K. R. S., Jyothi, P., Sai, G. V., & Rohith, R. (2020). Credit card fraud detection using Machine learning algorithms. *Journal of Research in Humanities and Social Science*, *8*(2), 04-11. Google Scholar↗

[40] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*(1), 321-357. Google Scholar↗

[41] Baum, E. B. (1988). On the capabilities of multilayer perceptrons. *Journal of complexity*, *4*(3), 193-215. Google Scholar↗

[42] Lee, C. H., Lin, C. R., & Chen, M. S. (2001). Sliding-window filtering: an efficient algorithm for incremental mining. *Proceedings of the tenth international conference on Information and knowledge management*, *1*(1), 263-270. Google Scholar↗

[43] Nadiammai, G. V., & Hemalatha, M. J. E. I. J. (2014). Effective approach toward Intrusion Detection System using data mining techniques. *Egyptian Informatics Journal*, *15*(1), 37-50. Google Scholar↗

[44] Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook*, 875-886. Springer, Boston, MA. Google Scholar↗

[45] Yee, O. S., Sagadevan, S., & Malim, N. H. A. H. (2018). Credit card fraud detection using machine learning as data mining technique. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, *10*(1-4), 23-27. Google Scholar↗

[46] Sahu, A., Harshvardhan, G. M., & Gourisaria, M. K. (2020). A dual approach for credit card fraud detection using neural network and data mining techniques. *IEEE 17th India council international conference (INDICON)*, *1*(1), 1-7. Google Scholar↗

[47] Akhilomen, J. (2013, July). Data mining application for cyber credit-card fraud detection system. In *Industrial Conference on Data Mining* 218-228. Google Scholar↗

[48] Maes, S., Tuyls, K., Vanschoenwinkel, B., & Manderick, B. (2002). Credit card fraud detection using Bayesian and neural networks. *Proceedings of the 1st international naiso congress on neuro fuzzy technologies*, *261*(1), 270-277. Google Scholar↗

[49] Patil, Vipul, and Umesh Kumar Lilhore.(2018). A survey on different data mining & machine learning methods for credit card fraud detection. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *3*(5), 320-325. Google Scholar↗

[50] Ata, O., & Hazim, L. (2020). Comparative analysis of different distributions dataset by using data mining techniques on credit card fraud detection. *Tehnički vjesnik*, *27*(2), 618-626. Google Scholar↗

[51] John, S. N., Anele, C., Kennedy, O. O., Olajide, F., & Kennedy, C. G. (2016). Realtime fraud detection in the banking sector using data mining techniques/algorithm. *International conference on computational science and computational intelligence (CSCI)*, *1(1)*, 1186-1191. Google Scholar↗

[52] Lerner, B., & Malka, R. (2011). Investigation of the K2 algorithm in learning Bayesian network classifiers. *Applied Artificial Intelligence*, *25*(1), 74-96. Google Scholar↗

[53] Ma, S. C., & Shi, H. B. (2004). Tree-augmented naive Bayes ensembles. *Proceedings of 2004 International Conference on Machine Learning and Cybernetics*, 3(1), 1497-1502. Google Scholar↗

[54] Chauhan, H., Kumar, V., Pundir, S., & Pilli, E. S. (2013). A comparative study of classification techniques for intrusion detection. *International Symposium on Computational and Business Intelligence*, *1*(1), 40-43. Google Scholar↗

[55] Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), 433-459. Google Scholar↗

[56] Khoshgoftaar, T. M., Seiffert, C., Van Hulse, J., Napolitano, A., & Folleco, A. (2007). Learning with limited minority class data. *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, *1*(1), 348-353. Google Scholar↗

[57] Borges, T. A., & Neves, R. F. (2020). Ensemble of machine learning algorithms for cryptocurrency investment with different data resampling methods. *Applied Soft Computing*, *90*(1), 5-42. Google Scholar↗

[58] Hashemi, M., & Karimi, H. (2018). Weighted machine learning. *Statistics, Optimization and Information Computing*, *6*(4), 497-525. Google Scholar↗

[59] Dozono, H., Niina, G., & Araki, S. (2016, December). Convolutional self organizing map. *International conference on computational science and computational intelligence (CSCI)*, *1*(1), 767-771. Google Scholar↗

[60] Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36. Google Scholar↗

[61] Eng, J. (2005). Receiver operating characteristic analysis: a primer1. *Academic radiology*, *12*(7), 909-916. Google Scholar↗

[62] Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M. S., & Zeineddine, H. (2019). An experimental study with imbalanced classification approaches for credit card fraud detection. *IEEE Access*, *7*(1), 93010-93022. Google Scholar↗

[63] 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. Google Scholar↗

[64] 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*(1), 25579-25587. Google Scholar↗

[65] Zhu, H., Liu, G., Zhou, M., Xie, Y., Abusorrah, A., & Kang, Q. (2020). Optimizing weighted extreme learning machines for imbalanced classification and application to credit card fraud detection. *Neurocomputing*, *407*(1), 50-62. Google Scholar↗

[66] Barker, K. J., D'amato, J., & Sheridon, P. (2008). Credit card fraud: awareness and prevention. *Journal of financial crime,15(4), 398-410*. Google Scholar↗

[67] Vengatesan, K., Kumar, A., Yuvraj, S., Kumar, V., & Sabnis, S. (2020). Credit card fraud detection using data analytic techniques. *Advances in Mathematics: Scientific Journal*, *9*(3), 1185-1196. Google Scholar↗

[68] Singh, A., & Jain, A. (2019). Adaptive credit card fraud detection techniques based on feature selection method. *Advances in computer communication and computational sciences*, *1*(1), 167-178. Google Scholar↗

[69] Puh, M., & Brkić, L. (2019). Detecting credit card fraud using selected machine learning algorithms. *42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, *1*(1), 1250-1255. Google Scholar↗

[70] Tran, P. H., Tran, K. P., Huong, T. T., Heuchenne, C., HienTran, P., & Le, T. M. H. (2018). Real time data-driven approaches for credit card fraud detection. *Proceedings of the 2018 international conference on e-business and applications*, *1*(1), 6-9. Google Scholar↗

[71] Sarkar, T., & Shah, D. (2022). Modelly: An open source all in one python package for developing machine learning models. *Software Impacts, 14(1)*,1-4. Google Scholar↗

[72] Zhang, S., Cheng, D., Deng, Z., Zong, M., & Deng, X. (2018). A novel kNN algorithm with data-driven k parameter computation. *Pattern Recognition Letters*, *109*(1), 44-54. Google Scholar↗

[73] Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. *Machine learning* , *1*(1), 101-121. Academic Press. Google Scholar↗

[74] Domínguez-Almendros, S., Benítez-Parejo, N., & Gonzalez-Ramirez, A. R. (2011). Logistic regression models. *Allergologia et immunopathologia*, *39*(5), 295-305. Google Scholar↗

[75] 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), 1-20. Google Scholar↗.

[76] Guo, H., & Jin, B. (2010). Forensic analysis of skimming devices for credit fraud detection. *International Conference on Information and Financial Engineering* , *1*(1), 542-546. Google Scholar↗

[77] Tasmin, S., Sarmin, A. K., Shalehin, M., & Haque, A. B. (2022). Combating the Phishing Attacks: Recent Trends and Future Challenges. *Advanced Practical Approaches to Web Mining Techniques and Application*, *1*(1), 106-137. Google Scholar↗

[78] Panthakkan, A., Valappil, N., Appathil, M., Verma, S., Mansoor, W., & Al-Ahmad, H. (2022). Performance Comparison of Credit Card Fraud Detection System using Machine Learning. *International Conference on Signal Processing and Information Security*, *1*(1), 17-21. Google Scholar↗

[79] Singh, A., & Jain, A. (2019). Adaptive credit card fraud detection techniques based on feature selection method. *Advances in Computer Communication and Computational Sciences: Proceedings of IC4S 2018*, *1*(1), 167-178. Google Scholar↗

[80] universal cpa review, (2022). Data mining to detect fraudulent credit card transactions. https://www.universalcpareview.com/wp-content/uploads/2021/07/data-mining-credit-card-fraud.png. Retrieved on 13/01/2023.

[81] Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, *48*(3), 128-138. Google Scholar↗

[82] Amruthnath, N., & Gupta, T. (2018). A research study on unsupervised machine learning algorithms for early fault detection in predictive maintenance. *5th international conference on industrial engineering and applications (ICIEA)*, *1*(1), 355-361. Google Scholar↗

[83] Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine learning*, *109*(2), 373-440. Google Scholar↗

[84] Aithal, P. S., & Kumar, P. M. (2015). Applying SWOC analysis to an institution of higher education. *International Journal of Management, IT and Engineering*, *5*(7), 231-247. Google Scholar↗

[85] Kesavan, V., & Srinivasan, K. S. (2022). A Case Study on the Digital Payment Systems in India. *Compendium of Management Case Studies*, *1*(1), 17-26. Google Scholar↗

[86] Helms, M. M., & Nixon, J. (2010). Exploring SWOT analysis–where are we now? A review of academic research from the last decade. *Journal of strategy and management*, *3*(3), 215-251. Google Scholar↗

[87] Berry, T. (2018). What is a SWOT analysis?. *B Plans*, *1*(1), 1-10. Google Scholar↗

[88] Gretzky, W. (2010). Strategic planning and SWOT analysis. *Essentials of strategic planning in healthcare*, *1*(12), 91-108. Google Scholar↗

****