FMDB *Transactions on Sustainable Computer Letters*

Enhancing Security in Digital Payments: A Comparative Evaluation of Machine Learning Models for Credit Card Fraud Detection

Rejwan Bin Sulaiman[1, *], Mohammad Aljaidi[2], Amjad A. Alsuwaylimi[3], Usman Butt[4], Md. Simul HasanTalukder[5], Syeda Sadia Alam[6], Maruf Farhan[7], Sanjoy Ranjon Das[8], Md Wahidul Alam[9]

[1,7,] Department of Computer Science, Northumbria University, Tyne, England, United Kingdom.
[2] Department of Computer Science, Zarqa University, Zarqa, Jordan.
[3]Department of Computer Science, Northern Border University, Arar, Saudi Arabia.
[4]Department of Computer Science, British University, Dubai, United Arab Emirates.
[5]Department of Computer Science, Bangladesh Atomic Energy Regulatory Authority (BAERA), Dhaka, Bangladesh.
[6]Department of Computer Science, Metropolitan University, Sylhet, Bangladesh.
[8]Department of Computer Science, Shipley College, Shipley, United Kingdom.
[9]Department of Computer Science, University of York, Heslington, York, England, United Kingdom.
rejwan.sulaiman@northumbria.ac.uk[1], mjaidi@zu.edu.jo[2], Amjad.alsuwaylimi@nbu.edu.sa[3], usman.butt@buid.ac.ae[4], simulhasantalukder@gmail.com[5], Syeda.alam.juti@gmail.com[6], Marufrigan9@gmail.com[7], Sonjoyict@gmail.com[8], alamw69@gmail.com[9]

Abstract: Nowadays, credit card transactions are exponentially increasing as a form of online payment. That creates technical pressure on the financial institution to incur losses due to credit card fraud. It is no wonder that credit card fraud is making people feel insecure and unsafe about using the services provided by banks. Data mining reasons on offline and online transactions can cause fraud detection. Number one is the changes in the behaviour of the frauds that are changing, and number two is the fraudulent dataset, which is asymmetric. In addition, the variables and techniques used by the researcher can impact fraud detection activities for credit cards. Therefore, the paper intends to investigate the suitability of the k-nearest-neighbour, decision tree, support vector machine, logistic regression, and Catboost. The data set has a total of 284807 transactions coming from the European credit card holder. Evaluation of the performance can be measured through the specificity, accuracy, sensitivity, precision, and lastly, the recall rate. After completing the comprehensive cross-validation, it was discovered that the catboost's accuracy was extraordinary, which amounted to 93.39%, leaving the other classification matrix far behind.

Keywords: Enhancing Security; Digital Payments; Machine Learning Models; Credit Card Fraud Detection; K-nearest-neighbour (KNN); Decision Tree (DT); Support Vector Machine (SVM); Logistic Regression (LR).

Received on: 05/12/2023, **Revised on:** 09/02/2024, **Accepted on:** 01/04/2024, **Published on:** 09/06/2024

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSCL

DOI: https://doi.org/10.69888/FTSCL.2024.000182

Cite as: R. B. Sulaiman, M. Aljaidi, A. A. Alsuwaylimi, U. Butt, M. S. H. Talukder, S. S. Alam, M. Farhan, S. R. Das, and M. W. Alam, "Enhancing Security in Digital Payments: A Comparative Evaluation of Machine Learning Models for Credit Card Fraud Detection," *FMDB Transactions on Sustainable Computer Letters*., vol. 2, no. 2, pp. 63–84, 2024.

1. Introduction

[*]Corresponding author.

Credit card fraud is a serious problem, and banks and other financial institutions have worked tirelessly to find solutions. Finding solutions to unusual cases of credit card theft is a top priority for academic institutions and academics. The globe loses billions of pounds annually due to fraudulent transactions; in 2020, that number is projected to reach 31.67 billion [1]. The development of e-commerce has substantially raised the possibility of credit card fraud occurring during online transactions. Customers' credit card information is more susceptible to theft if shared with an outside entity. Detecting and preventing fraud are the primary methods to circumvent this [16]. To train the system to make predictions based on the data, machine learning (ML) is an effective method. This approach is used to detect fraud. But there's a catch: sharing the credit card transaction dataset is extremely delicate. Second, there is a significant imbalance between fraudulent and non-fraudulent transactions in the credit card dataset due to skewness [17]. In light of the issues raised by these datasets, we are developing a model that can draw its training data from real-time datasets, protect the confidentiality of consumers' information, and improve the accuracy of its fraud detection capabilities [18]. Finding ways to prevent credit card fraud has recently captured the researchers' interest. The use of machine learning algorithms across industries is all the rage [19]. A large body of research across disciplines has sought to answer this question to determine whether machine learning is the optimal implementation strategy. Examples include the following: the use of deep neural networks for information efficiency prediction [2], the enforcement of EV and instruction regulations [4], and the detection of credit card, potato, and jüte pests by Sulaiman and Schetinin [23]. Researchers are striving to acquire or train the power of ML [20] despite all the constraints of its usage.

2. Literature Review

2.1. What is Credit card fraud detection

To begin with, the first and foremost task is to comprehend the complex nature of the detection system in the modern era. Anti-fraud systems are unique detection systems that are familiar among people due to the immense spread of fraudulence with credit cards. From one point of view, anti-fraud can have a problem that is fairly regarded as dichotomous [29]. However, after several investigations, scholars and scientists found that anti-fraud should be regarded as a multi-classification problem since all the fraudulent acts are different and have a minimum resemblance. Therefore, it can be said that fraudulent acts change over time [30]. Banks and insurance companies are the two entities engulfed by the danger of fraud. Hence, it is immensely appreciated if companies take necessary actions, such as changing the security model now and then, to make it difficult for fraudsters to commit fraud and create chaos in society [31].

2.2. Credit Card Fraud Detection (CCFD) Anatomy

Credit card fraud detection can be categorized into two folds:

- The dataset needs to be labelled. Then, it will require supervised and structured learning. However, one of the significant fallbacks is the steadiness of taking updates [32].
- One of the most important challenges of having supervised learning comprised of the label is the low chance of detecting fraud. This happens because it can only detect fraud from historical data and cannot detect new modes of fraudulent acts. Therefore, the researchers must improve the detection system for credit card fraud. To do this, data mining may add value to the research [33].

2.3. Credit Card Fraud Detection

Fraud detection for credit cards is crafted to ensure that no unauthorized transactions occur since businesses and customers suffer a lot from accidents. The fraudster keeps changing his/her method of operation to prevail. The anti-fraud committee must push the financial mechanism to a rigorous level that the fraudster cannot reach and match. Below, the popular actions of the anti-credit card fraud systems are given [34].

Validation transaction using merchant trade: To identify the user, the merchants need to possess a list that contains the credit card information along with the card number that has been referenced instead of the present card number. Doing so generates important information such as a PIN or security code [35].

Validating Geo-IP address: Geo-location, a modern technology, is used based on the IP addresses of multiple computers. It can pinpoint the fraudster on a real-time basis. This may permit the merchants to use authentication to transact the applications to the pragmatic examples with which the fraudsters can be kept aloof from the fraudulence [36].

Flagging high-risk using IP address: The detection system ensures the identity of the billing address and IP address country [37]. The detection system can detect IP addresses to embrace fraud prevention technology. For example, if a customer orders

something using an Indian IP address, but the shipping and billing address are from the UK. It must undergo a sophisticated review since the anti-fraud precautions will come into play [38].

Mailboxes and proxies: People tend to use email addresses that require no money and are convenient. Also, fraudsters use this kind of email address to remain hidden and anonymous. One key way to fight against fraud prevention is to find the newly registered domain addresses [39]. The anonymous proxy servers may remain hidden under the authenticated IP address. The principal motto is to remain anonymous so credit card fraud can be handled [40].

Machine learning detection: Another prominent detection system is based on the machine learning classification algorithm, which detects fraudulent credit card fraud. It is very popular and effective in terms of detecting fraud. Scientific studies revealed that machine learning classification algorithms are better suited to find and identify the fraud of credit cards [41]. The scholars concluded that this is one of the most effective tools to fight credit card fraud against fraudsters. Regarding credit card fraud, the techniques depend highly on oversampling algorithms and deep learning [42]. LSTM, Long Short-Term Memory Networks, is one of the detection models for fraud against credit cards. Synthetic Minority Over-sampling Technique (SMOTE) along with the k-nearest Neighbor (kNN) are combined to devise "kNN-Smote-LSTM," which works for the fraud detection program. This program enhances the performance of the fraud detection system [43].

2.4. Credit Card Fraud Identification

Challenges persist in detecting credit card fraud since people who use credit cards are not yet familiar with the fraudulent behaviour of the fraudsters. It is because the fraudsters target the companies to score big rather than targeting individuals. Therefore, companies must update their security to protect the stakeholders' interests [44]. However, individuals often become victims of fraudsters since the fraudsters might have the credit card number and the expiry date of the credit card holder. Afterwards, the fraudsters do not need permission from the actual credit card holder [45].

2.5. Consequence of Credit Card Fraud

The user and the company are the direct sufferers once the fraudsters get a hold of the individuals' credentials. The following bullets are the outcomes generated due to the incident of credit card fraud.

- Monetary losses of the businesses and the users
- Information and privacy breaches
- Lack of trust in enterprise for information security

There are numerous methods that companies undertake to prevent credit card fraud. However, the fraudsters are becoming much more intelligent and desperate to steal. This keeps the companies on their toes and strengthens their security measures. From mid-July 2005 to January 2007, 45.6 million credit cards have been disclosed since the system of TJX had been compromised [46]. Albert Gonzalez was accused of theft and was the leader of the incident. Also, in August 2009, Albert Gonzalez was indicted for the massive 130 million credit card theft. 2016, a coordinated cyber-attack took place in Japan. A team of a hundred people organized the attack on 1600 credit cards and stole $12.7 million. In Tokyo, it targeted 1400 convenience stores, and it took about three hours to steal and escape [47].

2.6. Countermeasure

For the customers, companies tend to develop the countermeasures for the transactions. For example, countermeasures are added with different signs to different areas of designation. The company's cardholders should be secured and protected from the unauthorized usage of their financial assets. The secured areas for making the transaction should be done through the financial institutions' security system. In addition to this measure, the cardholders' cards have CHIP identification, so card theft is likely to decrease.

User Training and Education: Customers must have sound knowledge of credit cards and how to file a complaint if fraud-related losses occur. Therefore, customers must check the bills charged to identify their transactions and detect unauthorized transactions immediately. Afterwards, they must let know the financial institutions about the mishaps. Customers need software to record their account numbers, addresses, and telephone numbers. Hence, users and customers need to be careful when making online purchases. It is recommended that harmful websites are to be avoided. In no circumstances should credit card information be shared with anyone or any sites.

Government Legislation: The government is responsible for shaping the legislation against fraud. For instance, a law for the enactment is germane to credit card transaction fraud. This law can enable the customers to ensure safety in the market of

transactions. Also, according to the GDPR principles of the EU, companies that issue credit cards are bound to share the standard procedure of guidelines so that the customers can protect themselves from fraudulent activities [8].

Machine Learning Classifiers: Logistic regression, Decision Tree (DT), KNN, Support Vector Machine (SVM), Category, and Boosting (CatBoost) are the five methods of classification that have been used in this project. These types of classification models address problems like differential training datasets. In addition to this, these can be utilized to learn the classification. Therefore, the comparison between them has been addressed in different studies.

2.7. Logistic Regression Algorithm

It is one of the prominent ways to determine and detect credit card fraud using logistic regression. It is chosen due to the simplicity of the model. Also, it is a sophisticated method to understand and predict the results. A multinomial logistic regression is used to construct the model. One of the unique attributes of this model is that it can process faster and is an appropriate tool for the bi-categorical problem. In addition, it is easy to understand for anyone with some knowledge of information technology. Regarding the bottlenecks, the limit for the data is one, and the other is the adaptability capacity of the scene. Researchers also tried to combine LR with a Neural network and decision tree, which was fascinating. It is important to evaluate the models and to do so, validation and testing (2006) and transaction for training (2005) were used. Hence, the performance comparison among these are 5.88%, 5.84%, and 3.89%, respectively. Ileberi et al. [10] studied the comparison between the logistic regression (LR) and ANN. This piece of the study had been invested in resources by the researchers to evaluate the performance of the credit card fraudulence referring to the test dataset. The main methods of logistic regression method:

Objective: The objective is to find the risk factor for which the specific transactions can be suspected of becoming fraud.

Prediction: The prediction of the fraud being committed can be made based on numerous models of algorithms.

Judgment: Judgment is most likely identical to the prediction. The judgment is based on specific models that show the likelihood of a transaction being the risk factor in each situation.

H-function: It is important to construct the logistic function h(x), also known as the sigmoid or predictive function. Afterwards, the predictive process is expected to be constructed since suitable parameters can be set for the training data to be sent to the vector. The following Figure 1 depicts the function:

$$g(z) = \frac{1}{1+e^{-z}}$$

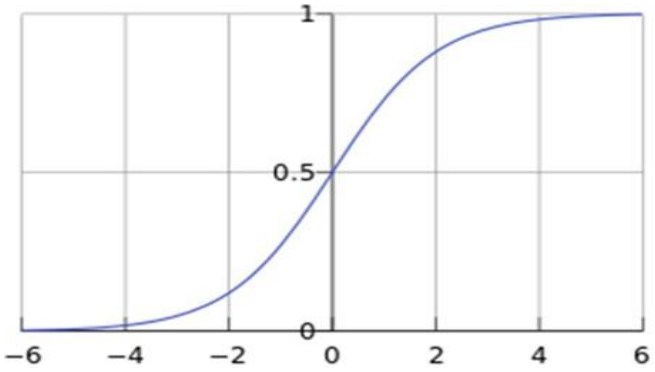

Figure 1: Logical function expressions [24]

Loss Function: The next phase consists of generating the loss function-j. Generally, there are several samples with n types of attributes. The derivation of J functions and the cost depend on the optimum estimation.

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & if\ y = 1 \\ -\log(1 - h_\theta(x)) & if\ y = 0 \end{cases}$$

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost(h_\theta(x_i), y_i) = -\frac{1}{m}\left[\sum_{i=1}^{m}(y_i log h_\theta(x_i) + (1 - y_i)\log(1 - h_\theta(x_i)))\right]$$

The following step involves solving the lowest value of θ using the gradient descent. The operation of generating θ can be shown below:

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i) x_i^j$$

Decision Tree: The decision tree is utilized for different scenarios with known probability. The reason behind the formation of the decision tree is to identify the possibility of whether the present value is equal to zero or greater than that, which will be used to determine the risk of the process. In addition to that, it utilizes the possibility and capability of the decision analysis methodology. Afterwards, based on that, the decision branch is portrayed and looks like a tree. Therefore, the terminology is fixed on a 'decision tree'. Decision trees can be folded into three categories: selection of features, decision tree generation, and decision tree pruning. In machine learning, the decision tree is a predictive model that portrays a map between the object values and properties. The decision tree is also utilized for the classification method. The decision tree is a technique that can be utilized to analyze data and for prediction purposes. Therefore, it is chosen so that the training can be done on fraud detection. The following is an example of a decision tree classification model where the red boxes are called features (Figure 2).

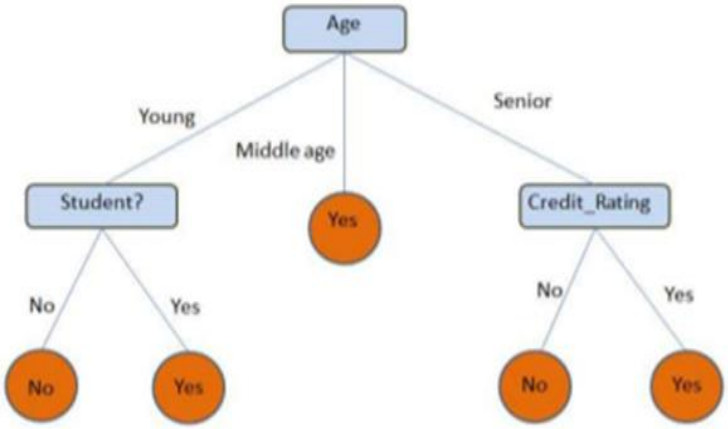

Figure 2: A simple decision tree [25]

The decision trees are involved in human horizontal thinking. Therefore, understanding is not difficult to comprehend and can make interpretations good. The decision tree can make people see the rationality behind the interpretation, whereas other algorithms like SVM NN cannot make it. In the tree, each node stands for the element or the attribute, each branch or link stands for a decision being made, and each leaf stands for a value.

K-nearest Neighbour (KNN): In terms of simplicity in the classification techniques of data mining technology, KNN is one of the simplest and most mature tools in the available resources. The term 'k-nearest neighbours' stands for the values that can denote each of the samples. It is a simple method of classification to record in the data set. Awoyemi et al. [3] used a different strategy to encounter the problem. A dataset containing 284807 trades was used. The dataset was drawn from the European trading market and utilized as a hybrid technique later executed in Python. The accuracy of the logistic regression was 54.86%, whereas the accuracy of the KNN was 97.69%. The result proves the effectiveness of KNN, among other techniques. KNN can be implemented if it knows the category of all the available unknown and known samples. Afterwards, the distance is calculated between the models, and therefore, the known examples of K are chosen based on the majority-majority voting system.

$$d((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

The 'scikit learn' is determined and adjusted by the parameters of n_neighbors, where five is known as the default value. According to Figure 3, K= 3 stands for identifying the green process that will belong to the class of red triangles since the red triangles' proportion is calculated to 2/3. On the other hand, if K=5, the blue square will have belonged to the green circle because the proportion is calculated to be 3/5.

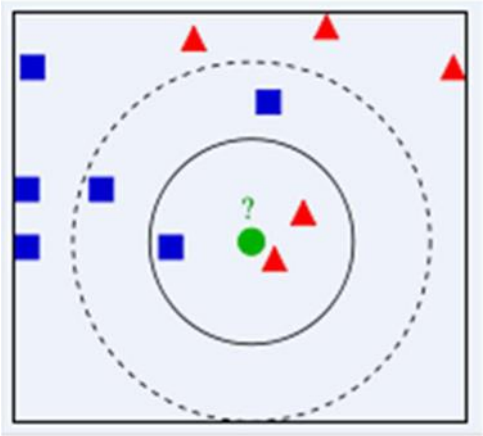

Figure 3: The k-nearest neighbour sample [26]

Support Vector Machine (SVM): SVM is a supervised method and a popular tool for regression analysis and statistical classification. This method is the centre point of this project as well. It is recognized as the 'maximum edge region classifier. The job of SVM is to map the vectors to form a high-dimension space. It will allow the hyperplane to establish itself. Parallel hyperplanes are constructed to divide the data [21]. It is the understanding that the smaller the error will be if the space is higher in between the hyperplanes set parallel.

$$r = y\frac{\mathbf{w}^T\mathbf{x} + b}{\|\mathbf{w}\|}$$

Distance from the example to the separators. The example located near the hyperplane is known as the support vector. ρ is the width between different support vectors inside the classes [22] (Figure 4).

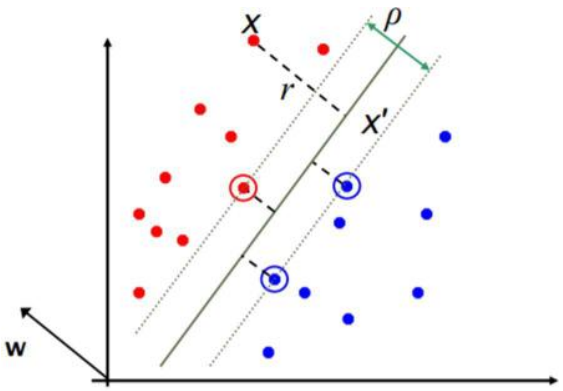

Figure 4: Geometric Margin [27]

Category & Boosting (Catboost): This framework is popular and easy to use. It is a framework of both categorical features with gradient boosting based upon the decision trees of gradient boosting [9]. It is capable of handling different types of data. It is important to compare the different performances of all the libraries using different classification algorithms [23]. The following are the advantages of boost;

- Catboost is unique and robust, with features of numerous categories [6]. The catboost handles categorical characteristics differently. It first computes the frequency of a type, like the fraudulent transaction class in this case, and then adds hyper-parameters to build additional numerical characteristics [6]. In addition, Catboost is powerful: it lowers the need to tune numerous hyperparameters and the risk of overfitting, which makes the model more adaptable. catboost is useful.
- Catboost is pragmatic in terms of use. It can augment the dimension of the features. It can handle numerical and categorical characteristics and employs combined category features to benefit from element linkages, greatly increasing the feature dimension.

- Catboost is an optimized version that provides enhanced performance. The fundamental model of catboost employs symmetric trees, and the leaf value is calculated differently from the usual booster process. However, catboosts use different techniques to prevent overfitting. That's why the catboost algorithm can outperform any modern machine learning system.
- Catboost has an interface with Python and scikit to facilitate instant calls; thus, it is easier to use. Catboost is simple to use and features Python, R, and command-line interfaces that can be integrated with scikit. His extensibility is seen in his custom loss function.

2.8. Synthetic Minority Over-sampling Technique (SMOTE)

The Synthetic Minority Oversampling Technique, known as SMOTE, is a development of oversampling random algorithm. The principle of the SMOTE algorithm is to analyze the sample data by adding different samples to the dataset. On the other hand, a problem in this project called the 'class imbalance' refers to the odd distribution of numerous classes in the given training set. There are three approaches; these are-

Adjusting the Value of Θ: It is important to properly adjust θ based on the positive and negative samples in the training set. A prediction or an educated guess is formulated on the training set.

Oversampling: To mitigate the imbalance of the classes, new models are used for the synthesis. In addition, the classes with fewer samples tend to be oversampled at times in the training set.

Undersampling: To mitigate the class imbalance, under-sampling of different classes is required in the training set [5]. For this project, it is important to conduct both under-sampling and oversampling operations to make a comparison. The comparisons will let the researcher know what the suited operation will be to be conducted for this research and how that operation should be performed in this research project. In addition, comparisons of the results and the disadvantages and advantages of each available technique can be made [1]. The principal job of the synthetic minority oversampling technique, SMOTE, is to generate extra models from the minority class samples. Suppose the value of K can be specified beforehand. In that case, the k-nearest minority sample to xi can be found from a minority sample xi using the k-nearest neighbour method. The distance can be identified as the Euclidean distance that persists among the different models. In the end, random selection is made in the k-nearest neighbours so that a new sample can be generated and used in the formula given.

$$\mathbf{x}_{new} = \mathbf{x}_i + (\hat{\mathbf{x}}_i - \mathbf{x}_i) \times \delta$$

The equation above denotes the x^ as the elected neighbour point for the k-nearest. Also, δ ∈[0,1] is considered as a random number. Figure 5 shows a sample that is SMOTE generated and has used 3-nearest neighbours; also, the model generated by the SMOTE is kept online by the x^i and xi.

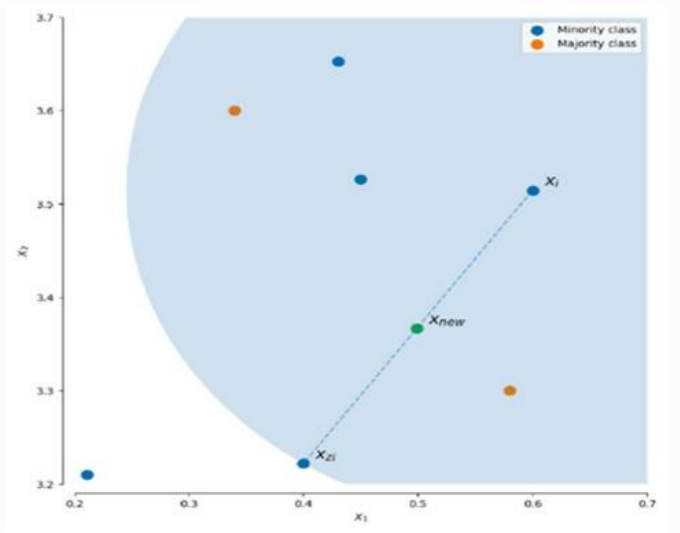

Figure 5: The Example of SMOTE Formation Sample [28]

2.9. Comparative Literature Analysis

The research study has limitations, such as having a single data set and being confined to using some staple classifier algorithms. However, the features used to detect credit card fraud might be identical even though the methods of collection and selection are not the same. Therefore, different research will be prone to generate different outputs. According to Dighe et al. [7], KNN is considered the most accurate algorithm, whereas the logistic regression occurs at the bottom of their interest list. Surely, the choice of algorithms depends on the process and the asking of the research.

2.10. Comparative Analysis Without Cross-Validation

The comparison has been made based on the results of the experiment with the results of the experiments that took place before and done by scholars or researchers. The intent behind doing this is to search for credit card fraud and how the techniques have been used. To evaluate clearly, it is important to consider the size of the dataset, classifier, and evaluation method, which are salient to determine the model's accuracy level.

Table 1: Comparison of different studies without CV

Author	Classifier	Sample Size	Accuracy
Kareem et al., [14]	ANN & Logistic Regression	Training: 2723 Testing: 1163	94.51%
Kumar et al., [15]	Logistic Regression, KNN, Support Vector Machine (SVM), Decision Tree(DT), Category Boosting(CatBoost)	Training: 698 Testing: 299	93.1%
Awoyemi et al., [3]	Naïve Bayes, K -nearest neighbour(KNN), Logistic Regression(LG)	Training:159,238 Testing: 68,236	97.37%

Table 1 formulated above comprises the parameters of two different types of studies. It has been seen that 97.37% accuracy has been achieved, and the achievement was achieved by collecting the highest number of samples in the given sets. The selection of the research was the Naïve Bayes, K-nearest neighbour, and the logistic regression [3]. According to Awoyemi et al. [3], neural networks can be a huge number of datasets that have an optimum number of datasets that have 95.84% accuracy. Therefore, it can be said that, as the number of samples grows, Bayesian and KNN are not optimal even after the cross-validation of the tuning.

2.11. Comparative Analysis With Cross-validation

Upon the cross-validation activity, the output, accuracy and dataset will experience changes in different ways.

Table 2: Comparison of different research using Cv

Author	Classifier	Sample Size	Accuracy
Itoo et al., [11]	ANN & Logistic Regression	Training: 2723 Testing: 1163	94.51%
Manderna et al., [12]	Logistic Regression, KNN, Support Vector Machine (SVM), Decision Tree(DT), Category Boosting(CatBoost)	Training: 688 Testing: 295	94.0%
Awoyemi et al., [3]	Naïve Bayes, K -nearest neighbour(KNN), Logistic Regression(LG)	Training:159,238 Testing:68,236	97.37%
Kabiraj et al., [13]	Decision Tree, Neural Networks and Logistic Regression	Training:159,238 Testing:68,236	95.38%

Cross-validation may not be regarded as one of the most crucial factors in the case of classifier technology. However, based on the table's results, it does affect specific classifiers. Awoyemi et al. [3] achieved a 97.69% accuracy rate, the highest in the Table 2. Afterwards, it was found that there are other points in which the number of datasets can be a crucial factor.

For example, when the training dataset has increased, there is always an increase in the accuracy rate. Therefore, the conclusion can be drawn from this aspect by stating that the training set size has a crucial impact on the accuracy of the performance. Since the training set size increased, the performance improved. Nevertheless, the performance of the classification can change abruptly when a specific size is reached and stability is earned in the process. On the contrary, it is important to remember that in the case of the actual experiment, if the size of the training set increases and the performance summits, the training time will increase as a result. Consequently, the growth of the potential and the time of classification will also tend to improve. In future work, it will be important for the researcher to consider both the requirements of time and the classification performance together.

3. Methodology

The section occupies the study's methodological operation to identify fraudulent transactions from the pool of non-fraudulent transactions. The following Figure 6 clarifies the phases that are kept in this research study. Before moving forward with the phases of the methodological operation for this research study, the dataset discussion is required to poise, which is given in Figure 6.

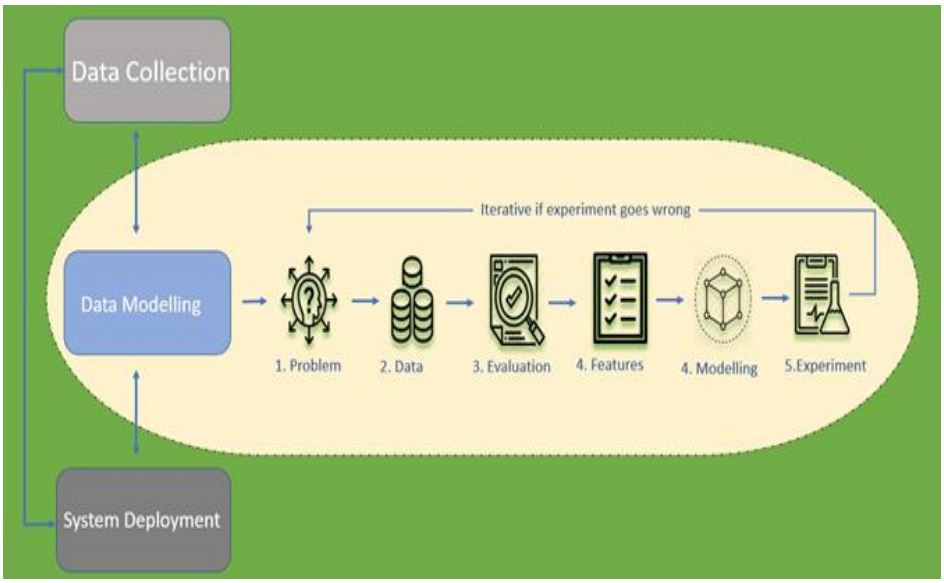

Figure 6: Classification Methodology

3.1. Tools and Software Used

Python has been used for execution, and it is one of the most popular languages worldwide. The data analysis and mining can be projected based on the basic pandas, the NumPy, and different open-source libraries. NumPy, in addition to different Python libraries, are built together. This consists of regression, reduction of the dimensionality, drinking pre-processing, regression, classification, model selection, and different types of functions. Therefore, it can save resources. Hence, it can be said that the catboost library should be improved. It is important to download and install the full package of the catboost since it is an algorithm library for all purposes. Thus, the project can use the catboost and sklearn open-source library properly.

3.2. Data Description

In terms of describing the dataset, it has 31 unique features. The features are put on with labels, and the labels are given for V1 to V28. These are considered anonymous. The rest of the three features are known as the total amount of the transaction, time, and the justification for whether the transaction can be considered fraudulent. PCA, Principal Component Analysis, is the form to which the anonymized 28 variables are being modified. Afterwards, it is put on to upload to the Kaggle. It is mentionable that before the initiation of the analysis, it is salient for the researcher to go through the dataset and check rigorously so that no non-null values or missing values can reside in the dataset (Figure 7).

```
In [9]: data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 284807 entries, 0 to 284806
Data columns (total 31 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   Time    284807 non-null  float64
 1   V1      284807 non-null  float64
 2   V2      284807 non-null  float64
 3   V3      284807 non-null  float64
 4   V4      284807 non-null  float64
 5   V5      284807 non-null  float64
 6   V6      284807 non-null  float64
 7   V7      284807 non-null  float64
 8   V8      284807 non-null  float64
 9   V9      284807 non-null  float64
 10  V10     284807 non-null  float64
 11  V11     284807 non-null  float64
 12  V12     284807 non-null  float64
 13  V13     284807 non-null  float64
 14  V14     284807 non-null  float64
 15  V15     284807 non-null  float64
 16  V16     284807 non-null  float64
 17  V17     284807 non-null  float64
 18  V18     284807 non-null  float64
 19  V19     284807 non-null  float64
 20  V20     284807 non-null  float64
 21  V21     284807 non-null  float64
 22  V22     284807 non-null  float64
 23  V23     284807 non-null  float64
 24  V24     284807 non-null  float64
 25  V25     284807 non-null  float64
 26  V26     284807 non-null  float64
 27  V27     284807 non-null  float64
 28  V28     284807 non-null  float64
 29  Amount  284807 non-null  float64
 30  Class   284807 non-null  int64
dtypes: float64(30), int64(1)
memory usage: 67.4 MB
```

Figure 7: Dataset Information

Upon checking, it has been discovered that the dataset possesses no missing values, and no values are non-null. Therefore, it can be said that this dataset is set to be explored in the exploratory data analysis.

3.3. Class Distribution

How many transactions can potentially become fraudulent is a matter of question. Well, it is understood and believed that most of the transactions are considered non-fraudulent. To the dataset for this project, it is discovered that only 0.17% of transactions are fraudulent, whereas 99.83% of transactions fall under the pie of non-fraudulent. Figure 8, portrayed in the following, can significantly depict the contract between non-fraudulent and fraudulent.

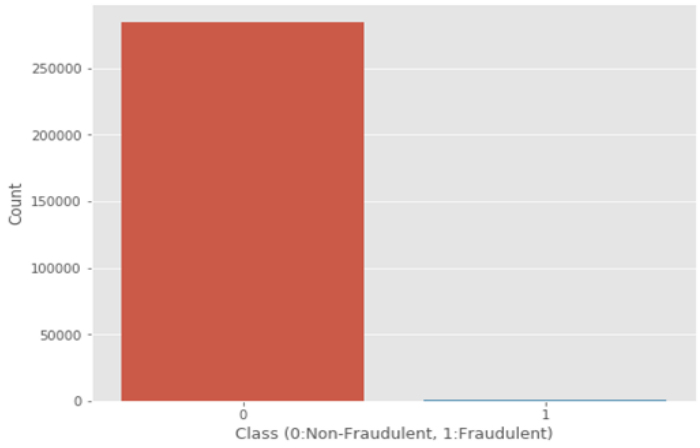

Figure 8: Unbalanced data points (Fraud vs genuine)

The dataset for this project contains many transactions that can avidly provide the best view for the researcher in this study. To be exact, there are 284,807 transactions overall in this dataset. The highest value of this dataset can be mounted up to €25,691.16. In this dataset, the mean value was found to be €88.35. Therefore, it is evident that the data are quite dispersed and spread out from the mean value. Therefore, it is perfectly understood and comprehended that the reason for the monetary value distribution is right-skewed. In addition, most of the transactions of the dataset are smaller due to the small mean value, whereas only some of the transactions are closer to the highest transactional value. Hence, the right-skewed distribution of monetary value occurs by the transactional value (Figure 9).

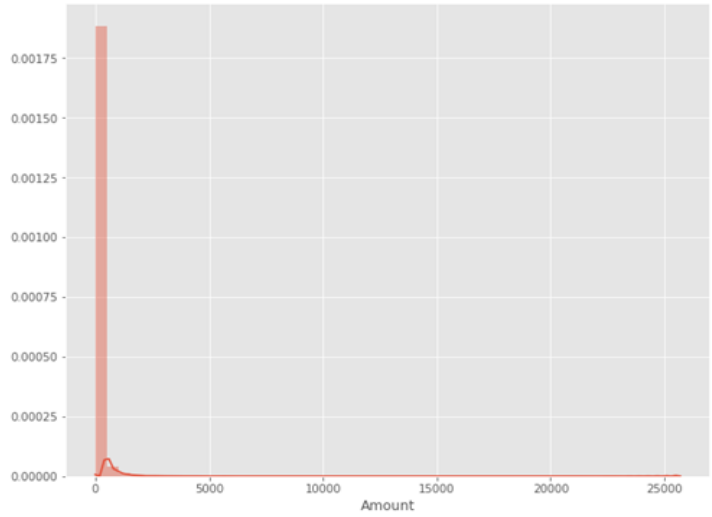

Figure 9: Distribution of the data points

Time has been recorded in seconds so that the transactions' beginning persists in the dataset. Hence, it is evident that all the transactions have been recorded in this dataset. Data that took place for two days have been included in the given dataset. The monetary value of all the given transactions has a bimodal distribution. This signifies the drop in the overall transaction history after 28 hours of transactions. Though the time has not been found for the first transaction, whether it occurred in which part of the day, it seems reasonable and justifiable that the overall transaction's drop occurred most of the night (Figure 10).

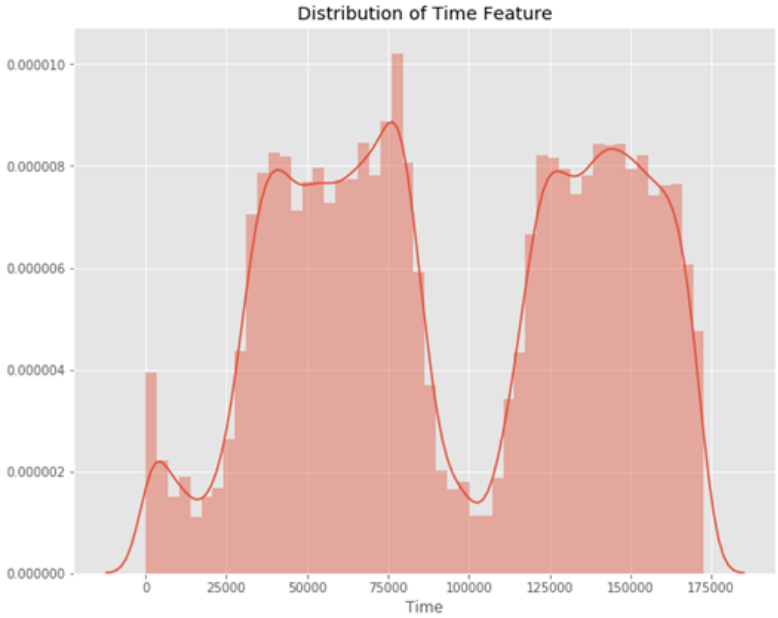

Figure 10: Time Frame Density Plot and Hist Plot

It will be interesting to know whether there is a correlation between the predictors and the variables found in the classes. A heat map can be used to identify the different ways to use visual representations.

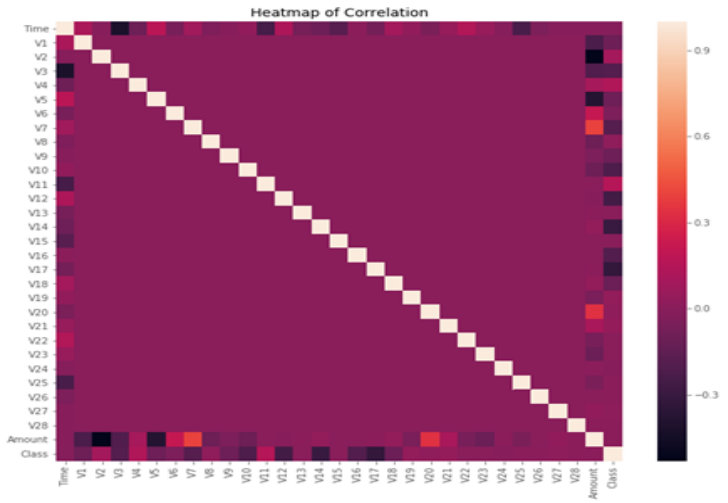

Figure 11: Correlation Heatmap of the dataset

As can be seen in the diagram, the predictors have some correlation with the class variable (Figure 11). However, the significance of the correlation seems to be comparatively little for a huge number of acting variables in this case. Therefore, it can be attributed to two factors: Since the data have been prepared with the assistance of PCA, the predictors can be considered the principal components. The imbalance that persists in the class may be able to disorient the need for specific correlations that can be related to the class variable.

3.4. Balancing The Imbalanced Set of Data

After doing all these hurdles, it does not seem to end; rather, this is the part where the challenge grows to the optimum level. Now, it is required to devise a training data set that will permit the algorithms to comprehend and specify the characteristics to identify fraudulent transactions. Utilizing the original data set will not be wise because almost 99% of transactions are standard and not fraudulent. Therefore, transactions that are not fraudulent will incur an accuracy of more than 99%. However, the thing is, it is neither expected nor needed because it is not needed to possess the accuracy of 99% of transactions that are not fraudulent since the researcher is in the business of identifying the fraudulent transactions in this research study.

Processing of Sample: The imbalance can be seen in a massive sample where the target column class has been presented. It is possible that the sample imbalance can impact the learning of the model. Therefore, the Synthetic Minority Oversampling Technique (SMOTE) has been considered for managing the sample imbalance.

Under-Sampling Process: The under-sampling process is not very complex. It allows the researcher to draw a specific sample from many samples. This is where the new dataset is created; the training starts for the machine to be trained. Afterwards, the under-sampling process yielded 50% fraudulent and standard transactions for each one, whereas the sample size was reduced to 984 (Figure 12).

Percentage of normal transactions: 0.5
Percentage of fraud transactions: 0.5
Total number of transactions in resampled data: 984

	V1	V2	V3	V4	V5	V7	V9	V10
154670	-2.296987	4.064043	-5.957706	4.680008	-2.080938	-4.490847	-1.593249	-8.993811
282830	-2.019495	1.418367	-0.726150	-1.466264	1.779066	-2.125326	0.114217	-1.041870
12696	1.264678	-0.409435	0.311049	-1.085468	-0.050964	-0.862553	2.946587	-1.330825
11198	1.271861	-0.291513	-0.890908	-1.008252	1.777980	-0.831441	1.539442	-0.540662
157734	-0.513183	0.817151	2.394285	-0.113539	0.140662	0.563751	1.479868	-1.158718

Figure 12: Under-Sampling Process (After)

It is time to classify the sample categories using the sns.countplot function (Figure 13).

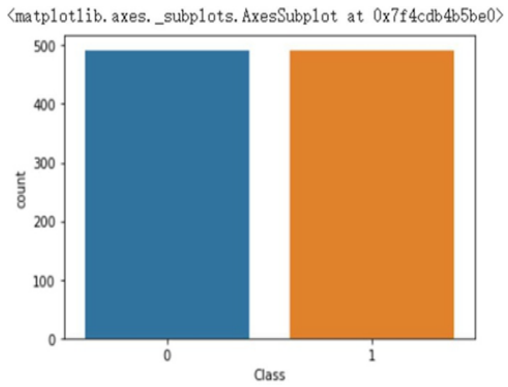

Figure 13: Classification of the Sample

Oversampling Process**:** The oversampling process increases the total number of negative and positive models by increasing the number of positive samples. Afterwards, these tend to learn the overall process of constructing the oversampled data. After being done with the dataset, the total samples of '1' are counted as 227454. That certifies that the overall samples in terms of '0' have also augmented to 227454. Therefore, the 50% of the total size of the sample is 454908. It is wise to utilize the SMOTE algorithm to do the up-sampling technique, which does more than copying the samples from the original data set. Furthermore, it can select an interval vested for each feature, which has fluctuations even for a small margin. Afterwards, it is possible to conduct feature generation and combine all the generated features into a class of new samples. The models developed using this process can be regarded as common sense. Because of the downsample dataset that was created upon the primary logistic regression, the parameters that were selected before may not turn out to be appropriate for the unsampled data. Therefore, it is important to have discovered the optimum parameter for this purpose of the study. To do that, it is necessary to utilize the 'RandomizedSearchCV' for the tuning process. Unlike GridSearchCV, this 'RandomizedSearchCV' will be dedicated to consuming fewer resources such as time and memory.

3.5. Pre-processing of the Data

The study needs to undergo pre-processing since it is one of the important operations for executing the algorithm for machine learning. Data training can be impacted by the consideration of different models that have been generated to produce predictions. Data pre-processing aims to formulate existing data into a specific form, including precise prejudice for variation and missing values. The original data includes both categorical and numerical data. Therefore, the researcher must encode the categorical data before executing the modelling. In addition, data outliers were detected and removed. Furthermore, the independent variables with the expected range were discovered by conducting the feature scaling. The box-cox transformation has been initiated to reduce the skewness of the feature. The process of under-sampling and oversampling has been conducted to reduce the bias and overcome the imbalance that persists in the primary dataset. Therefore, adopting the machine learning library sci-kit and Python data manipulation library pandas tend to increase the responsibilities vested in pre-processing. The stages are portrayed in the following Figure 14:

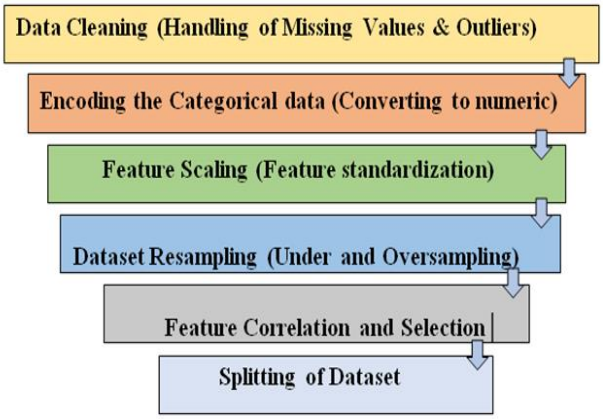

Figure 14: The Steps for Data Pre-processing

3.6. Engineering of the Features

In this section, a diagram will be formulated to construct the feature engineering so that a holistic view of the data and the comprehensive distribution can be conducted. It is salient to have extracted almost all the features from the huge pool of raw data with the assistance of the model and the algorithm. In addition to this, these will allow the researchers to choose, accumulate, and scale all the other elements to ensure excellent performance. Therefore, the researcher in this project intends to facilitate the precision and accuracy of the model by training the model based on the performance of the feature engineering. Figure 15 compares the time dimensions regarding the standard and fraud classes. It depicts that the regularity can be varied due to the regular transactions of the time distribution. Meanwhile, it has been perceived that there should be no exact pattern of time to commit fraudulent transactions.

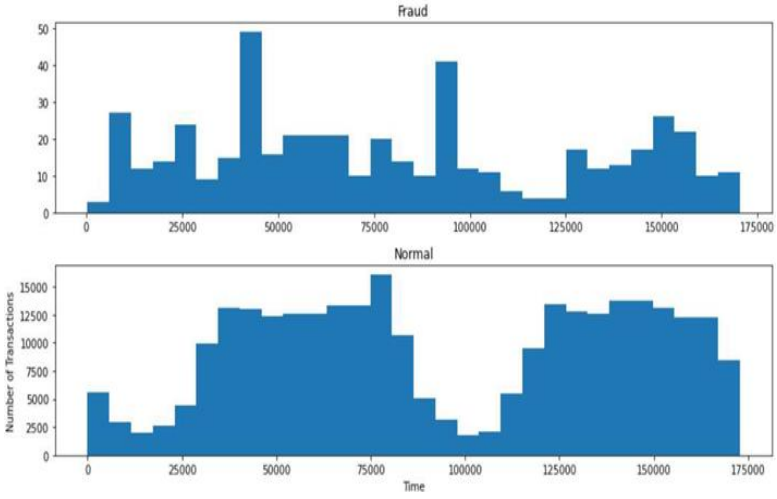

Figure 15: Fraud and standard classes time dimension comparison

However, another type of comparison persists between the amounts of the order. Here, both transaction types are present as a distribution, which is a long tail. At the same time, the fraudulent orders seem to be smaller and do not exceed $1000. Generally, the standard transactions can range up to $15000. Figure 16 depicts the plot that shows the amount against the time. In addition, this exhibits the standard transactions are poised evenly during different times. In contrast, the outliers for the transactional amount have been proven to have lesser frequency than before. Fraudulent transactions are generally scattered over a significant amount of time, and outliers are bound to occur frequently.

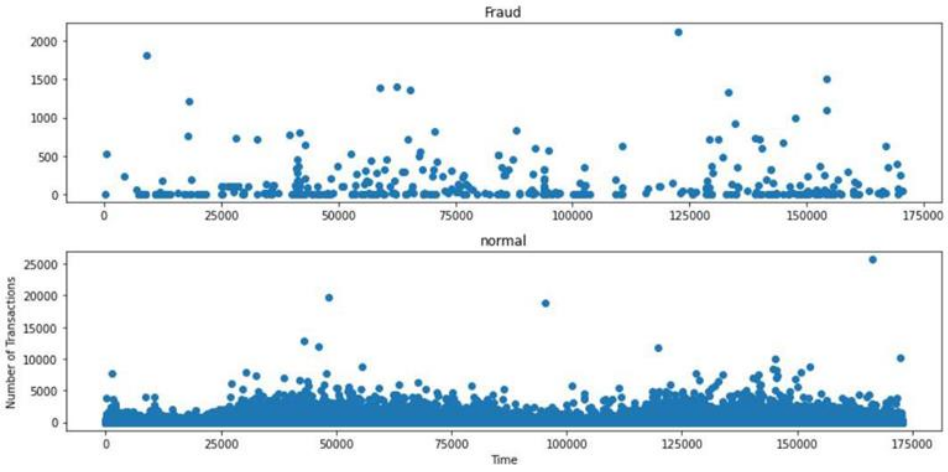

Figure 16: Time against amount scatter diagram

Afterwards, it is time to export all the distribution of PCA-processed features. Here, the observation will occur on distributing all the fraud and standard class elements. The observation found that the distribution on V6, V8, V13, V20, V22, V23, V24, V25, and V26 is identical in both categories. Moreover, the identical shape regarding the distribution stands for the feature having minimal impact on predicting the final output (Figure 17).

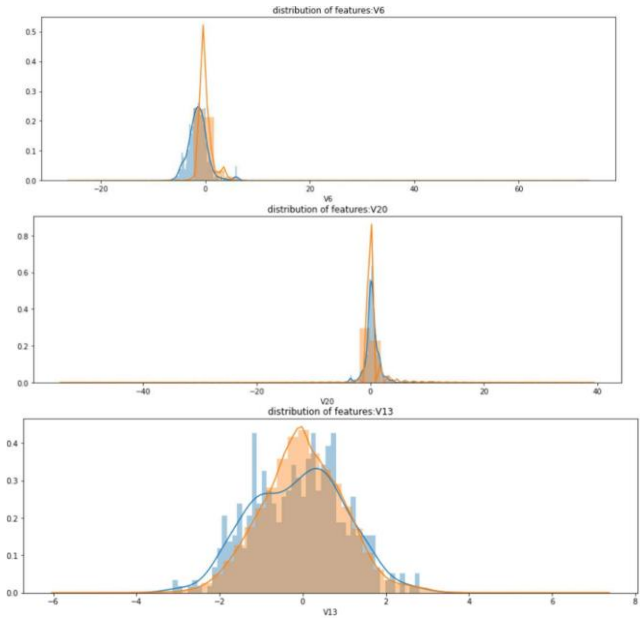

Figure 17: Identical shape of the distribution

In this process of conducting machine learning, in the last step, it has been mentioned that the autumn column contains data that float more than it should. It is important to do this so that the difference in the eigenvalue is not extremely big. Therefore, the need for pre-processing and standardizing the data came to light. Normalization was done by embracing the mean-standard deviation method of the two dynamic spheres related to hour and amount (Figure 18).

	V1	V2	V3	...	V28	Class	normAmount
0	-1.359807	-0.072781	2.536347	...	-0.021053	0	0.244964
1	1.191857	0.266151	0.166480	...	0.014724	0	-0.342475
2	-1.358354	-1.340163	1.773209	...	-0.059752	0	1.160686
3	-0.966272	-0.185226	1.792993	...	0.061458	0	0.140534
4	-1.158233	0.877737	1.548718	...	0.215153	0	-0.073403

[5 rows x 30 columns]

Figure 18: Data collected right after standardization

3.7. Summary of Factors Influencing Choice of Project

This section interprets the visual analysis regarding the review of the sample, the review of the data, and the data reduction. Afterwards, the following attributes were found, and it was inspected whether the specific attribute was salient or not to the need of the study.

Samples with no missing values found for the study: It has been found that the sample itself has no missing values, which certifies that the sample is complete. It can be regarded as a benefit for the researcher since the sample will not need any data manipulation and can be utilized directly for the study.

The class of the sample is not balanced: The study needs to perform an operation called SMOTE sampling. Therefore, it is important to have the sampling class proportion of 0 and 1, which is not balanced in this case. If it is not taken care of, it can impact the accuracy level of the training model.

Irregularity in the transaction time concerning the class of fraudulence: This is a unique dataset attribute. Even though this attribute does not have a specific impact on the training of the sample, it is possible and suggested that it can be done using intuitive comparison for the prediction done at the last result.

Small value orders detected in the fraud class: The transaction orders for the fraud class are hardly more than $1000. Also, it is found that this attribute is indicative. Hence, comparing different predictions found in the metadata results can help predict the transactions that can match the attributes found in the fraud class and the original data.

Scattered time distribution of different fraudulent transactions: To compare different results, the discovered anomalies of numerous anomalies can be considered, which are extensively large in number. The similarity was found in two areas, including the distribution of features belonging to the PCA process: The deletion process is expected and will be needed because the category in this project has identical attributes.

4. Result Analysis

The study aims to devise an appropriate model for the fraud detection system, and the researcher needs to conduct experiments to ensure the model's output. The dire need for pre-processing, as well as the output generated due to various hyperparameters, have been depicted in the section. Lastly, a comparative analysis was formulated to compare numerous algorithms and select the best one.

4.1. Preserving the original dataset

Long before the training begins, it is of utmost importance that the original dataset be split and reserved. To do that, dynamite is introduced, which serves the purpose of sampling the dataset later and will be responsible for altering the original dataset. The method adopted for this operation is known as the StratifiedShuffleSplit, and it is important to reserve a copy of the data at the beginning. StratifiedShuffleSplit is the combination of both ShuffleSplit and StratifiedKFold. These ensure the folds carry on identical samples belonging to all the categories. In the meantime, ShuffleSplit and StratifiedKFold randomize the test pairs based on the parameters and divide them. This is how the maintenance of the original imbalance is ensured. Afterwards, the findings will be utilized to validate the anticipated output.

```
sss = StratifiedShuffleSplit(n_splits=5, test_size=2, random_state=42)
```

4.2. Evaluation of different classification models

The dataset needs to be diced and sliced right before the creation of the model. To do so, train_test_split has been adopted to slice and turn the dataset into x_train,x_test,y_train, and y_test.

```
x_train, x_test, y_train, y_test = train_test_split(x_new, y_new, test_size=.2, random_state=42)
```

Four types of pre-chosen model have been identified, and an observation has been made on their expected performance. Logistic Regression, Decision Tree, SVM, KNN, and Catboost are the five types of models considered for the study. Afterwards, the final selection will be made based on the model's performance.

Table 3: Accuracy after initial training

Classifier Name	Accuracy
Logistic Regression (LR)	94.0%
K-Nearest (KNN)	93.0%
Support Vector Classifier (SVM)	85.0%
Decision Tree Classifier (DT)	90.0%
CatBoostClassifier (CBT)	93.0%

Upon the observation made on the models stated above, by the performance concern, catboost and logistic regression provides better performance (Table 3). Though logistic regression is quite complex, the performance improves once this model is adopted. On the other hand, the performance of catboost is also good since this model comprises different algorithms. However, in this research study, the models have been formed using default parameters, and it is important to adjust the parameters so that the model's accuracy can be verified. It is important to mention that the model will be built on an accuracy rate and not based on the recall rate. The model's accuracy will become more convincing once the data from the post-sampling category balancing is utilized to assess the model better.

4.3. Classifier evaluation with cross-validation

One of the bottlenecks of the model training can be the usage of the identical model training for the training conducted on the identical dataset. This might lead to a scenario where the model itself will be overfitting. Therefore, the sample will be divided, and then the dataset's cross-validation will occur. Then, the model will be learned from the training set, and the parameters will tune the validation set. Afterwards, the test dataset will evaluate the model's overall performance. Grid search will be used to select optimal parameters. Figure 19, attached below, depicts the parameters and the tuning models.

```
#log_reg_params  =  {"penalty":['20','12'],'C':[0.001,0.01,0.1,1,10,100,1000]}
log_reg_params  =  {"penalty":  ['11',  '12'],  'C':  [0.001,  0.01,  0.1,  1,  10,  100,  1000]}

grid_log_reg  =  GridSearchCV(LogisticRegression(),  log_reg_params)
grid_log_reg.fit(x_train,  y_train)

#automatically  get  the  logistic  regression  with  the  be4st  parameters
log_reg  =  grid_log_reg.best_estimator_

knears_params  =  {"n_neighbors":list(range(2,5,1)),'algorithm':['auto','ball_tree','kd_tree','brute']}

grid_knears  =  GridSearchCV(KNeighborsClassifier(),knears_params)
grid_knears.fit(x_train,y_train)

knears_neighbors  =  grid_knears.best_estimator_

svc_params  =  {'C':[0.5,0.7,0.9,1],  'kernel':['rbf','poly','sigmod','linear']}
grid_svc  =  GridSearchCV(SVC(),  svc_params)
grid_svc.fit(x_train,y_train)

svc  =  grid_svc.best_estimator_

tree_params  =  {'criterion':['gini','entropy'],'max_depth':list(range(2,4,1)),'min_samples_leaf':list(range(5,7,1))}
grid_tree  =  GridSearchCV(DecisionTreeClassifier(),tree_params)
grid_tree.fit(x_train,y_train)

tree_clf  =  grid_tree.best_estimator_

#scoring="roc_auc"
catboots_params  =  {'depth':  [6]}
grid_cboots  =  GridSearchCV(CatBoostClassifier(),catboots_params,scoring="neg_mean_squared_error",iid=False,n_jobs=-1,cv=5)
grid_cboots.fit(x_train,y_train)

cboots_clf  =  grid_cboots.best_estimator_
```

Figure 19: Setting the parameters for different models

A grid search was used to set up the parameters. Afterwards, the grid search will invest in finding different combinations of the parameters so that the best location can be found according to the evaluation mechanism. With respect to the grid search, two types of parameters are adjusted:s kerne and C. C is regarded as the penalty parameter. If the default value is calculated as 1.0, then the higher degree of C can be equivalent to the variable of penalty relaxation. It is expected to find the relaxation variable to be proximate to zero. In most likely cases, the training sets are automatically divided into pairs, which is accurate regarding dataset testing. However, it does not have a strong ability for generalization. Different kernel functions can be changed. The kernel arguments are represented using a short form like 'Linear', 'poly', 'rbf', 'sigmoid', 'precomputed', etc. 'rbf' is used by default. We have used 5-fold cross-validation.

Table 4: Results of cross-validation after changing the parameter

Classifier Name	Accuracy After Cross-Validation	Accuracy Before Cross-Validation	Change
Logistic Regression (LR)	94.78%	94.0%	0.78%
K-Nearest (KNN)	93.52%	93.0%	0.52%
Support Vector Classifier (SVM)	93.14%	93.0%	0.14%
Decision Tree Classifier (DT)	92.25%	90.0%	2.25%
CatBoostClassifier (CBT)	93.39%	93.0%	0.39%

Table 4 shows that the accuracy for each of the respective models has been enhanced to a level right after the parameter has been adjusted. In addition, the results also show that logistic regression can be regarded as one of the most appropriate models for the research study, followed by vector machine, KNN, and Catboost. However, a decision tree should be out of the question when the implementation is concerned due to poor performance. Logistic regression has an accuracy of 94.78%, whereas logistic regression is used as a default parameter in tan. The tuning model assisted the logistic regression with an improved accuracy of 0.78%. For the k-nearest (KNN), 93.52% accuracy was found once cross-validation was conducted.

However, it was 93.0% before the tuning, which shows an improvement in accuracy of 0.52%. To the SVM, it was found that the accuracy of the SVM can be calculated as 93.14% after the adjustment is done. Before the adjustment, the accuracy was 93%, 0.14% lower than the one with the adjustment. The rate of accuracy for this can be positioned in the third-last position. Regarding the decision tree, it did not perform as expected. However, after the adjustment, an improvement was seen since the accuracy rose from 90% to 92.25%, with a total jump of 2.25%. Despite this jump, the position of the decision tree is at the bottom of the table in terms of accuracy. Lastly, the Catboost rose from 93% to 93.39%, with a total improvement of 0.39% accuracy.

Nevertheless, it was the third most accurate algorithm for this research study. The difference in accuracy persists between the logistic regression and the vector machine. The difference is minimal since it is observable that the learning curve shapes the degree of fit. In this case, the test and training set accuracy for the logistic regression are close. This means the model is not trying to overcompensate, overfit, or even underfit. On the other hand, the accuracy of the training set for the vector machine is comparatively higher, with a hint of overfitting. Therefore, selecting the logistic regression as the prediction model for the job will be wise.

4.4. Data modelling using under-sampling

An under-sampling model will be created to predict the real data that will utilize the original data reserved at the beginning when the parameters were set with the StratifiedShuffleSplit. Possessing an unbalanced test set is important to make sense of the predictions. In addition, specific methods are to be used for the 'learn', including NearMiss, a method for downsampling.

The make-pipeline is somewhat like the mechanism that persisted in the sklearn. However, the setup is organized separately for sampling. Afterwards, the optimal model is selected, and the training begins after tuning the respective parameters. After that, individual scores are printed from the model, and the graph shows the recall as good. However, since the accuracy is low, training was done on the wrong positive sample to clear the negative samples. Then, the oversampling will be done for the smooth operation (Table 5).

Table 5: Undersampling benchmark

Type (undersampling)	Score
Accuracy	0.97
Recall	0.87
Precision	0.05
F1	0.12
Roc_auc	0.93

4.5. Test set/confusion matrix

Now, the oversampling method will be considered, and the process for training will remain identical. The intent is to use the logistic regression model to train the oversampling (Table 6).

Table 6: Oversampling Benchmark

Type (oversampling)	Score
Accuracy	0.97
Recall	0.87
Precision	0.0068
F1	0.11
Roc_auc	0.92

As can be seen from different metrics in the given classifiers, accuracy is pretty good regarding the oversampling method, unlike the one with the undersampling method. However, a reduction in the recall rate has been seen. However, the final prediction needs to be made based on the test set, and afterwards, a confusion matrix needs to be drawn for comparison purposes (Figure 20).

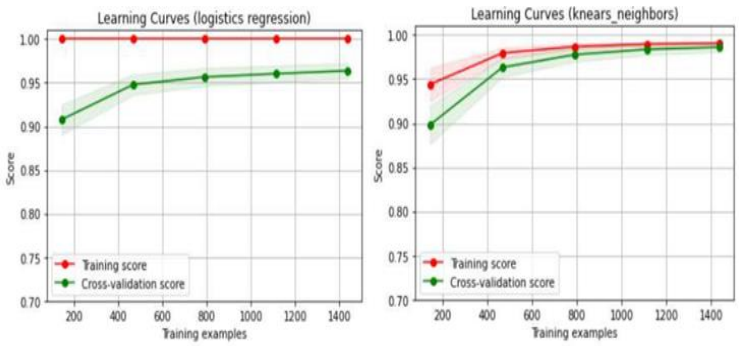

Figure 20: Learning Curve of KNN and LR

An identical operation, such as the one, has been done using the over-sampling method, yielding the following output. The confusion matrix provides a comparison between the two methods known for sampling.

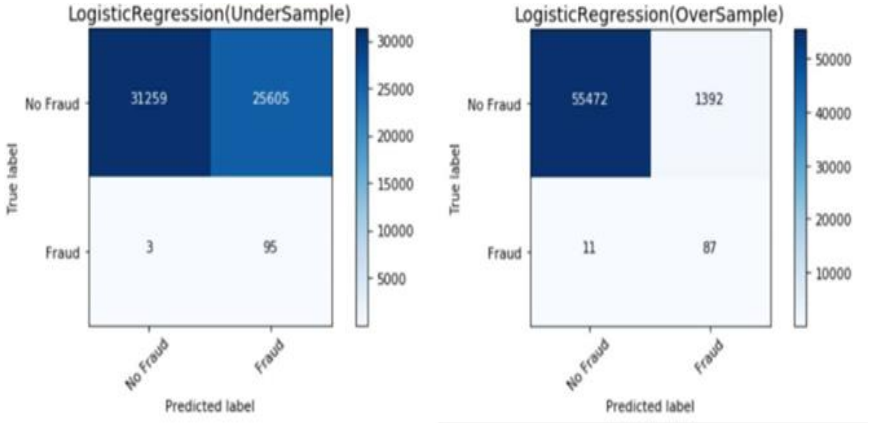

Figure 21: Confusion matrix of LR with oversampling and undersampling

Oversampling and undersampling operations were performed using the calculation of the confusion matrix in the logistic regression model. Figure 21, the first confusion matrix belongs to the undersampling. It is seen that the model is quite accurate enough to detect the fishy samples. Half of the overall transactions have been anticipated as fraudulent, which is what the samples make, not regular. However, this model can predict fraudulent samples, and it is commercially possible to see the impact and effect. 'Sklearn', 'Imblearn', and 'Keras' are the packages the researcher needs to use for the anti-fraud project. There are techniques such as grid search, random search, oversampling, downsampling, cross-validation, confusion matrix, learning curve, etc., which have been used in building the learning models such as logistic regression, neural network, decision tree, catboost, knn, and vector machine. Lastly, the chosen model was logistic regression since it has 97% accuracy with an 87% recall rate.

4.6. Conclusion of the Result

From kaggle.com, the credit card dataset has been considered. For the engineering of the feature and pre-processing, SMOTE has been used to deal with datasets that are not balanced. Afterwards, the model for detecting the fraudulent activities of credit card fraud has been devised using five algorithms: logistic regression, KNN, support vector machine, decision tree, and category and boosting (catboost). Afterwards, a confusion matrix was used so that a comparison could be made between the two sampling methods. Logistic regression is one of the best solutions for serving the expectations of this research study since it has an overall accuracy of 97.0%. Though decision trees and logistic regression are vastly used for credit card spoofing detection, this paper uses SVM and catboost to compare and generate new ideas for better performance. At the end of the analysis, we came up with two important conclusions. Firstly, although KNN and catboost algorithms performed extremely well, we believe they might do better in notation if we train the dataset in the training phase after the integration. Secondly, as the credit card data increases exponentially over time, it is not suggested that SVM be used on the credit card data. Moreover, the SVM algorithm takes a longer time to train. It is better to avoid using SVM to minimize the institution's costs.

4.7. Reflection Learning

This research study has embraced different features of machine learning. This project is also a requirement the school sets to pass the degree. Machine Learning Applications are one of the rigorous strategies we undertook to gain a clear understanding of the subject matter and to find relevance and resemblance to how data are used in the pragmatic world. Since the beginning of the degree, projects related to machine learning have been carried out, such as natural language learning, text classification, and so on. Compared to those, credit card fraud detection is something we are not familiar with. However, it gave me an immense opportunity to investigate a real-life challenge to make the world a bit better than before. On another note, this project has been challenging for me since five classification algorithms have been embraced to train the model with a view to cross-validating and tuning the optimization. One of the key things that we learned by doing it is breaking traditional barriers. In this project, unlike many others, we introduced a comparatively new algorithm, Catboost, to credit card fraud detection. No researcher has used Catboost in the credit card fraud detection literature. We firmly believe my attempt at the project is full of enthusiasm and integrity to learn. Lastly, a direction for future research has been poised for the researchers interested in credit card fraud detection research.

5. Conclusion

The research study invests all the resources in investigating the fraud detection models using different algorithms. Doing so is to train the dataset and test the system. Different processes have been taken to discover the best possible fraud detection model, and comparisons have been made using different models and theories on credit card and fraudulent transactions. The journey of choosing the appropriate model began with selecting five classifiers with ten different types of combinations of algorithms as well as different sampling methods so that a proper evaluation could be done on the prediction of the performance. Lastly, cross-validation was applied to achieve maximum accuracy. The study's discoveries can be summed up here: Using oversampling to minimize the damage of the unbalanced credit card dataset yielded identical results when introduced with the confusion matrix. Logistic regression has been utilized mostly due to the benefit of targeting data processing, which the SVM and Catboost have followed. The comparison was done using prior literature to test and train the dataset.

5.1. Future Work

The future work of this research study will comprise different suggestions and recommendations that other researchers can undertake to delimit the limitations of this research study. Though this research study has successfully detected credit card fraud, the improvement persists and should be looked at closely to make future research flawless. After completing the training model, it would be wise to amalgamate more than two classifiers so that the detection performance might have increased, and it would open the research to more possible outcomes. Using the deep learning, which is quite equivalent to the neural network. Deep learning is unlike machine learning since it is not supervised and utilizes unlabeled or unstructured data. It would be used to train the models for the fraud detection model in an easier way. Though catboost is an effective algorithm, the time limit is a concern. The researcher needs to adjust all the parameters to optimize the overall performance. The given data source is not authentic; it is someone's dataset. The amount of data will increase if it is needed to extract data from the specific network. Therefore, the performance of the model will be predicted and will be improved as a matter of fact, which will increase the accuracy of credit card fraud detection. Different types of attacks are tested in the machine learning of the classifier, and therefore, the analysis is performed under the attack. This might be another way to develop the security measure and to enhance the accuracy. To test the model, in the beginning, the existing dataset will be used to detect the performance against fraud detection. Afterwards, the evaluation will be done based on the bank's present credit card system so that real fraud detection can be seen in the pragmatic world through experiments and testing.

Acknowledgement: N/A

Data Availability Statement: The data for this study can be made available upon request to the corresponding author.

Funding Statement: This manuscript and research paper were prepared without any financial support or funding

Conflicts of Interest Statement: The authors have no conflicts of interest to declare. This work represents a new contribution by the authors, and all citations and references are appropriately included based on the information utilized.

Ethics and Consent Statement: This research adheres to ethical guidelines, obtaining informed consent from all participants.

References

1. M. Alghamdi, M. Al-Mallah, S. Keteyian, C. Brawner, J. Ehrman, and S. Sakr, "Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project," PLoS One, vol. 12, no. 7, p. e0179805, 2017.
2. M. Aljaidi, N. Aslam, X. Chen, O. Kaiwartya, Y. A. Al-Gumaei, and M. Khalid, "A reinforcement learning-based assignment scheme for EVs to charging stations," in 2022 IEEE 95th Vehicular Technology Conference: (VTC2022-Spring), London, United Kingdom, 2022.
3. J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," in 2017 International Conference on Computing Networking and Informatics (ICCNI), Lagos, Nigeria, 2017.
4. R. Bin Sulaiman, V. Schetinin, and P. Sant, "Review of machine learning approach on credit card fraud detection," Hum-Cent Intell Syst, vol. 2, no. 1–2, pp. 55–68, 2022.
5. A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, "Credit card fraud detection: A realistic modeling and a novel learning strategy," IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 8, pp. 3784–3797, 2018.
6. F. Yao, J. Sun, and J. Dong, "Estimating daily dew point temperature based on local and cross-station meteorological data using CatBoost algorithm," Comput. Model. Eng. Sci., vol. 130, no. 2, pp. 671–700, 2022.
7. D. Dighe, S. Patil, and S. Kokate, "Detection of credit card fraud transactions using machine learning algorithms and neural networks: A comparative study," in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018.
8. M. Foulsham, Living with the New General Data Protection Regulation (GDPR). Financial Compliance. Cham. Springer International Publishing, pp.113-136, 2019.
9. J. T. Hancock and T. M. Khoshgoftaar, "Gradient boosted decision tree algorithms for medicare fraud detection," SN Comput. Sci., vol. 2, no. 4, p.268, 2021.
10. E. Ileberi, Y. Sun, and Z. Wang, "Performance evaluation of machine learning methods for credit card fraud detection using SMOTE and AdaBoost," IEEE Access, vol. 9, no.12, pp. 165286–165294, 2021.
11. F. Itoo, Meenakshi, and S. Singh, "Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection," Int. J. Inf. Technol., vol. 13, no. 4, pp. 1503–1511, 2021.
12. A. Manderna, S. Kumar, U. Dohare, M. Aljaidi, O. Kaiwartya, and J. Lloret, 2023. "Vehicular network intrusion detection using a cascaded deep learning approach with multi-variant metaheuristic". Sensors, vol. 23, no.21, p.8772, 2023.
13. S. Kabiraj, L. Akter, M. Raihan, N. J. Diba, E. Podder, and M. M. Hassan, "Prediction of recurrence and non-recurrence events of breast cancer using bagging algorithm," in 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020.
14. A. Kareem, H. Liu, and P. Sant, "Review on pneumonia image detection: A machine learning approach," Hum-Cent Intell Syst, vol. 2, no. 1–2, pp. 31–43, 2022.
15. G. Kumar, S. Kumar, and A. A. Prakash, "Credit card fraud detection using machine learning," Int. J. Eng. Adv. Technol., vol. 10, no. 4, pp. 124–126, 2021.
16. S. R. Lenka, R. K. Barik, S. S. Patra, and V. P. Singh, Modified Decision Tree Learning for Cost-Sensitive Credit Card Fraud Detection Model. Advances in Communication and Computational Technology. Singapore; Singapore: Springer, vol.668, no.1, pp. 1479-1493, 2020.
17. E. F. Malik, K. W. Khaw, B. Belaton, W. P. Wong, and X. Chew, "Credit card fraud detection using a new hybrid machine learning architecture," Mathematics, vol. 10, no. 9, p. 1480, 2022.
18. A. Nandi, K. K. Randhawa, H. S. Chua, M. Seera, and C. P. Lim, "Credit card fraud detection using a hierarchical behavior-knowledge space model," PLoS ONE, vol. 17, no. 1, p.e0260579, 2022.
19. S. Poojitha and K. Malathi, "An Innovative Method to Enhance the Accuracy of Credit Card Fraud Detection Using Logistic Regression Algorithm by Comparing Random Forest Algorithm," vol. 107, no.1, pp. 14205–14218, 2022.
20. X. Zhang, Y. Han, W. Xu, and Q. Wang, HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture. Vol.557, no.5, pp. 302–316, 2021.
21. A. Singh and A. Jain, "A novel framework for credit card fraud prevention and detection (CCFPD) based on three layer verification strategy," in Proceedings of ICETIT 2019, Cham: Springer International Publishing, vol.605, no.1, pp. 935–948, 2019.
22. C. Sudha and D. Akila, "Credit Card Fraud Detection System based on Operational & Transaction features using SVM and Random Forest Classifiers," in 2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM), Dubai, United Arab Emirates, 2021.
23. R. B. Sulaiman and V. Schetinin, "Deep neural-network prediction for study of informational efficiency," in Lecture Notes in Networks and Systems, Cham: Springer International Publishing, 2022, vol.295, no.8, pp. 460–467.
24. R. Yedida and S. Saha, "Beginning with machine learning: a comprehensive primer," Eur. Phys. J. Spec. Top., vol. 230, no. 10, pp. 2363–2444, 2021.

25. A. Kumar, “Decision Tree Concepts, Examples, interview questions,” Analytics Yogi, 17-Feb-2023. [Online]. Available: https://vitalflux.com/decision-tree-algorithm-concepts-interview-questions-set-1/. [Accessed: 02-Oct-2023].
26. D. N. Dimid, “Supervised Learning algorithms cheat sheet,” Towards Data Science, 21-Sep-2021. [Online]. Available: https://towardsdatascience.com/supervised-learning-algorithms-cheat-sheet-40009e7f29f5. [Accessed: 02-Oct-2023].
27. Z. Hadjadj, “A cooperative framework for automated segmentation of tumors in brain MRI images,” Multimed. Tools Appl., vol. 82, no. 26, pp. 41381–41404, 2023.
28. F. Aguilar, “SMOTE-NC in ML categorization models for imbalanced datasets,” Analytics Vidhya, 09-Oct-2019. [Online]. Available: https://medium.com/analytics-vidhya/smote-nc-in-ml-categorization-models-fo-imbalanced-datasets-8adbdcf08c25. [Accessed: 02-Oct-2023].
29. M. M. Abbassy, "Opinion mining for Arabic customer feedback using machine learning," Journal of Advanced Research in Dynamical and Control Systems, vol. 12, no. SP3, pp. 209–217, 2020.
30. M. M. Abbassy and A. Abo-Alnadr, "Rule-based emotion AI in Arabic customer review," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 9, no.1, p.12, 2019.
31. B. Senapati and B. S. Rawal, "Adopting a deep learning split-protocol based predictive maintenance management system for industrial manufacturing operations," in Big Data Intelligence and Computing. DataCom 2022, C. Hsu, M. Xu, H. Cao, H. Baghban, and A. B. M. Shawkat Ali, Eds., Lecture Notes in Computer Science, vol. 13864. Singapore: Springer, pp. 25–38, 2023.
32. B. Senapati and B. S. Rawal, "Quantum communication with RLP quantum resistant cryptography in industrial manufacturing," Cyber Security and Applications, vol. 1, no.10, p. 100019, 2023.
33. B. Senapati et al., "Wrist crack classification using deep learning and X-ray imaging," in Proceedings of the Second International Conference on Advances in Computing Research (ACR’24), K. Daimi and A. Al Sadoon, Eds., Lecture Notes in Networks and Systems, vol. 956. Cham: Springer, pp. 72–85, 2024.
34. D. A. A. Al-Maaitah, T. A. M. Al-Maaitah, and O. H. M. Alkharabsheh, “The impact of job satisfaction on the employees turnover intention at public universities (Northern Border University),” International Journal of Advanced and Applied Sciences, vol. 8, no. 5, pp. 53–58, 2021.
35. E. Vashishtha and H. Kapoor, "Enhancing patient experience by automating and transforming free text into actionable consumer insights: a natural language processing (NLP) approach," International Journal of Health Sciences and Research, vol. 13, no. 10, pp. 275-288, 2023.
36. W. M. Ead and M. M. Abbassy, "A general cyber hygiene approach for financial analytical environment," in Financial Data Analytics, pp. 369–384. Cham: Springer International Publishing, Switerzland, 2022.
37. F. M. Masad, T. A. Al-Maaitah, D. A. Al-Maaitah, E. F. Qawasmeh, and N. A. Qatawneh, “Harnessing artificial intelligence for human resources management: Tools, advantages, and risks in the energy sector,” in E3S Web of Conferences, vol. 541, EDP Sciences, 2024.
38. K. Shukla, E. Vashishtha, M. Sandhu, and R. Choubey, "Natural Language Processing: Unlocking the Power of Text and Speech Data," Xoffencer International Book Publication House, USA, p. 251, 2023.
39. D. Köseoğlu, S. Ead, and W. M. Abbassy, "Basics of Financial Data Analytics," in Financial Data Analytics, pp. 23–57. Cham: Springer International Publishing, Switerzland, 2022.
40. M. M. Al-Ajlouni, D. A. Al-Maaitah, and T. A. Al-Maaitah, “Managing Supply Chains Using Business Intelligence,” Kurdish Studies, vol. 12, no. 2, pp. 5328–5337, 2024.
41. N. Alrawashdeh, A. A. Alsmadi, M. Alsaaideh, D. A. Maaitah, M. Al-Okaily, and A. Al-Okaily, “Embracing cryptocurrency in the financial landscape: An empirical study,” in Studies in Systems, Decision and Control, Cham: Springer Nature Switzerland, pp. 721–733, 2024.
42. S. Temara, “Harnessing the power of artificial intelligence to enhance next-generation cybersecurity,” World Journal of Advanced Research and Reviews, vol. 23, no. 2, pp. 797–811,2024.
43. S. Temara, “Maximizing Penetration Testing Success with Effective Reconnaissance Techniques Using ChatGPT”, Asian Journal of Research in Computer Science, vol. 17, no. 5, pp. 19–29, 2024.
44. S. Temara, “The Ransomware Epidemic: Recent Cybersecurity Incidents Demystified”, Asian Journal of Advanced Research and Reports, vol. 18, no. 3, pp. 1–16, Feb. 2024.
45. T. A. Al-Maaitah et al., “Strategies for success: A theoretical model for implementing business intelligence systems to enhance organizational performance,” Int. J. Adv. Appl. Sci., vol. 11, no. 5, pp. 55–61, 2024.
46. T. Maaitah, “The role of business intelligence tools in the decision making process and performance,” Journal of intelligence studies in business, vol. 13, no. 1, pp. 43–52, 2023.
47. T. Matin, “The Impact of Social Media Influencers on Brand Awareness, Image and Trust in their Sponsored Content: An Empirical Study from Georgian Social Media Users,” International Journal of Marketing, Communication and New Media, vol. 10, no. 18, p.12, 2022.