5/5/2022

**CREDIT CARD FRAUD DETECTION USING SUPERVISED AND UNSUPERVISED ALGORITHMS - A COMPARATIVE STUDY**

**Module Name:** Postgraduate Major Project

**Module Code:** MOD002726

**StuId:** 2028423

**Student Name:** Anusha Mandalapu

Contents

[1. Introduction 2](#_Toc102734380)

[Detection Methods brief 6](#_Toc102734381)

[Aim and objectives 6](#_Toc102734382)

[2. Literature Review 9](#_Toc102734383)

[3. Methodologies 19](#_Toc102734384)

[Logistic Regression 20](#_Toc102734388)

[Random Forest Classifier 21](#_Toc102734389)

[Support Vector Machine 21](#_Toc102734390)

[Artificial Neural Networks 22](#_Toc102734391)

[3.1. SMOTE Techniques 24](#_Toc102734392)

[Random Undersampling 24](#_Toc102734393)

[Random Oversampling 25](#_Toc102734395)

[ADASYN 26](#_Toc102734396)

[BORDERLINE 27](#_Toc102734397)

[SMOTE 28](#_Toc102734398)

[SMOTE-ENN 29](#_Toc102734399)

[SVM-SMOTE 29](#_Toc102734400)

[3.2. Dataset 30](#_Toc102734401)

[Data Pre processing and Selection 32](#_Toc102734402)

[Metrics Evaluation 32](#_Toc102734403)

[3.3. Implementation-SMOTE Techniques 36](#_Toc102734404)

[4. Results 41](#_Toc102734405)

[Results Analysis and Comparison 76](#_Toc102734406)

[5. Conclusion 77](#_Toc102734407)

[6. References 78](#_Toc102734408)

# Introduction

* **About credit card fraud**

In current digital transaction era, credit cards are one of the financial supports available to people. Nowadays, everyone aspires to have a credit card and uses it to make purchases. This location has become a snare for hackers looking to steal money from consumers using credit cards. When the credit cards are stolen by some fraudsters and misuse them through money transactions or giving it to other criminals, is generally known as **Credit Card Fraud [1].** Credit card fraud can occur mainly in two ways, i) card-present fraud, which occurs rarely in these days and ii) card-not-present fraud which is the most familiar way now a days for fraudulent transactions. Millions of accounts are encountering fraudulent transactions due to the security threats of the database and by attacking the database system to compromise the security. The account holders are unaware of these compromises till they get account statements [2]. As soon as the card is lost or stolen, customers can report to banks through telephone and can block their cards to prevent any transactions. But meanwhile it is possible for the thief to withdraw the money or to make any purchases using the stolen cards.

A number of ways are there for a fraudster to stole the card and to get the access of the credit card. Below are the 4 popular methods of credit card frauds:

1. **Card Skimming**

Credit card details are stolen by fraudsters using a small device without the knowledge of card holder, known as **Card Skimming [3]**. These devices are very small, portable and is generally carried out in busy places like restaurants, bars which is quite easy for fraudsters to copy the card details. Customer care and call centre services are popular ways for this card skimming.

In our daily purchases, it is possible for skimming using a third party device installed into the card machine. This third party device can read the card information like card number, pin when the customer swipe the card to pay the bills. Hence the thief can use this device to steal the customer card information.

Skimming is troublesome for the ordinary cardholder to distinguish, but given a expansive sufficient test, it is decently simple for the card guarantor to identify. It is a process that if a customer complains about unauthorized transactions, then the bank will uses data mining to trace the purchases and merchants at which the customer have used the credit card to pay the bills. It is essential that never ever handle the cards to merchants or shop keepers to pay the bills, there are high chances of frauds by duplicating our card details in that short span of time. Hence, we should always assure the money transactions and payments are doing safely without handling the card to the opposite person.

Timeline

Description automatically generated

**Figure 01: Data-driven approach for Credit Card Fraud Detection [11]**

1. **Lost or stolen credit cards**

It is quiet common way for fraudsters to use the credit cards when it is lost or stolen. By the time the customer realizes about the lost credit card and notifications of unauthorized transactions, it will be very late. So it is important to keep credit cards safe always and to inform banks to block the cards as soon as the customer notifies about the credit card lost. So it will prevent the unauthorized payments by the fraudsters.

1. **Card-not-present fraud**

It is not always necessary to have credit cards physically to make fraudulent transactions. It is possible if the fraudster have your credit card details remotely, then they can do payments actively without the card. It is quite complex to identify these frauds until the customer notices irrelevant transactions or unknown transactions in bank statements. So it is safe to check the transactions and corresponding bank statements regularly to track all the payments in order to avoid any fraud transactions. If any unusual transactions are found, we need to contact the card issuer to block the card to stop any further transactions. CNP is an abbreviation for card-not-present fraud. In USA, 7.9 million clients are casualties of CNP extortion and is slowly raised through e-commerce exchanges in Canada [4].

1. **Phishing**

Sometimes fraudsters send fake emails and make wrong calls to trap the people in order to collect the sensitive information. Phishing comes under social engineering attacks. Generally, this information is used to fill the applications and to create profiles in crucial websites for loan approvals, mortgages and to apply for credit cards as well. The personal details includes full name, date of birth, phone number, address and any nick names to hack the passwords by trail and error methods. They may too inquire for security passwords or your credit card points of interest, which can be utilized to form false instalments. Many people presume that every mail from banks are important and will follow it up unknowingly imaging that it is legitimate.

The key factor of any effective phishing assault could be a well-designed spoofed e-mail or spoofed site, which is why it pays to have a sound level of distrust when it comes to opening emails and going to websites [5].

## Detection Methods brief

To detect fraudulent transactions, various methods have been evolved which are using by many banks, companies. From many years, banks and credit card issuers are struggling to provide a robust security to its customers and to detect any fraudulent transactions as early as possible and to prevent them in early stages. For this, they are using various machine learning algorithms such as decision tree, deep learning models, supervised learning models, hidden Markov models [6]. Various algorithms are used to detect fraudulent transactions, among these decision tree, random forest classifier, K-Nearest Neighbour algorithms are most popular. In addition to these models, many sampling techniques and ensemble methods are used for better performance and accuracy results [7]. As the data is about customers transactions, it is quite confidential and is difficult to collect the datasets to train and test the models. In our research, we have found some publicly available datasets and methods to create synthetic datasets which replicates the original transactions data. More about this dataset and implementation of models, training, testing and picking out the best suited model along with suitable sampling technique is explained later in this paper.

## Aim and objectives

Predicting fraudulent transactions and preventing them is the point of interest for banks and credit card companies. The main aim of this paper is to assess the performance of the SMOTE family techniques across machine learning and deep learning models in identifying the credit card fraudulent transactions. As the dataset used in this paper is imbalanced, predicting the fraudulent transactions is a bit difficult task.

To the best of my knowledge, there is no study so far that compares the given SMOTE family techniques with supervised and deep learning models and conducting comparative research further to assess the performance of the SMOTE techniques and models and to encounter the issue with imbalanced dataset. Therefore, this paper focuses on various sampling techniques such as Random UnderSampling(RUS), Random OverSampling(ROS), Adaptive Synthetic Sampling(ADASYN), Borderline-SMOTE(BORDERLINE), Synthetic Minority Oversampling Technique (SMOTE), SVMSMOTE technique, SMOTE-ENN(combining undersampling technique ENN and oversampling techniques SMOTE).

In this paper, the above SMOTE family approaches are applied across the following machine learning and deep learning models to evaluate the performance of the SMOTE techniques to address the imbalanced data issue. The models discussed in this paper are Logistic Regression(LR), Random Forest Classifier(RF), Support Vector Machine(SVM), Artificial Neural Networks(ANN). While evaluating the performance, we have various performance metrics to compare such as precision, accuracy, f1-score and recall, ROC-AUC score and curve. Among these, few metrics such as recall, precision, ROC-AUC score are best suitable for imbalanced datasets. In this paper, we have used ROC-AUC curve as the important factor to assess the performance. In this study, we have identified that both Logistic Regression and Support Vector Machine algorithms with SMOTE-ENN sampling technique performs well with an AUC-ROC score of 0.9416 and 0.9365 respectively. Logistic Regression is compatible for smaller datasets. Therefore, Support Vector Machine performs well in the given algorithms comparatively.

Figure 02 represents the overall flow of the system of this paper with different model names and sampling techniques.

Diagram

Description automatically generated

**Figure 02: Overview of the Fraud Detection System with SMOTE techniques**

In this paper, the remaining sections are arranged as follows, section 2 discusses about literature review related to this field and the research analysis carried out on different papers and journals about imbalanced data and the techniques they have followed in handling the fraudulent transactions and their conclusions, further future work . The next section 3 discusses about the methodologies used in this paper. This section includes details about the dataset, brief introduction about the algorithms and SMOTE family techniques implemented and used in this paper. It also explains about the data pre-processing, feature extraction of data and implementation of the algorithms with different SMOTE techniques, training and testing the data and using different evaluation metrics. In section 4, the outcome of the assessment of algorithms and SMOTE techniques are explained with graphs and results framed in the tables and discussed about the comparison of the SMOTE techniques. In the final section, the conclusion and future work of the paper are explained. Further the references section containing various resources of the information helped in our research are cited and added in the references section.

# Literature Review

Now a days internet became an essential need in everyone’s life. In addition to telephone and television, online services through internet and social media platforms became more popular in people’s lives. Even online purchasing and online payments are very common and mostly they use credit cards for online payments. As the usage of credit card increases, parallelly the fraudulent transactions also enhanced. This credit card fraud transactions occurs in different ways and their sources are discussed in introduction part of this paper. Sequentially, various detecting methods have been evolved and proposed by the researchers to prevent these fraudulent transactions. Those methods and analysis of the different researchers, their proposed solutions to this problem are discussed below.

In 2017, John O. [John O. Awoyemi](https://ieeexplore.ieee.org/author/37086281308), [Adebayo O. Adetunmbi](https://ieeexplore.ieee.org/author/37085579729), and  [Samuel A. Oluwadare](https://ieeexplore.ieee.org/author/37085472689) [12] conducted the research to evaluate the performance of various algorithms using certain sampling techniques in detection of fraudulent transactions using an imbalanced dataset. In this paper, the authors have assessed the performance of certain algorithms such as Naïve Bayes, K-Nearest Neighbour model, Logistic Regression algorithms using sampling techniques as the dataset is imbalanced. With the use of advanced technology, fraudsters also updated their ways and attacking the customers credit cards in new ways. Hence the fraudulent behaviours also changing dynamically along with the customers.

Sampling techniques are applied for imbalanced dataset to distribute the fraud and not fraudulent classes equally to train the models and to get better accuracy results. A random sampling techniques equally pick and distributes the data into 50:50 ratio. As the dataset is imbalanced and contains more real transactions and very less fake transactions, a hybrid of sampling techniques are used to balance both positive and negative classes of the dataset. Three models selected in this paper are machine learning models which are trained using the sampled dataset and the performance is assessed using various evaluation metrics. The dataset is divided into two distribution datasets using hybrid sampling approaches. Three models are evaluated with performance metrics such as accuracy, sensitivity, precision and balanced classification rate, Matthews Correlation coefficient metrics.

In 1994, Ghosh and Reilly have proposed a deep learning model specifically Neural Networks to detect fraudulent transactions. Large datasets which are labelled were used to train this model and the transactions in this dataset were recorded over two months of duration [13]. Various fraudulent activities such as credit card lost/stolen, counterfeit fraud, phishing, NRI(non-received issue) can be detected using this neural network model by training the datasets in such a way. The results of this study proves that neural networks performs better in detecting fraudulent transactions compared to other rule-based fraud detection methods. The performance of this model is measured in terms of accuracy and duration to detect fraudulent transactions. In real world, this neural networks system is using installed and using at IBM 3090 in Mellon Bank and is current in use to detect the unauthorized transactions in credit card segment of that bank.

In a paper, J. V. V. Sriram Sasank, G. Ram Sahith, K. Abhinav and Meena Belwal [14] discussed about a system which proves that it is important to balance the datasets for both positive and negative cases to get better accuracy results and to detect fraudulent transactions accurately. The sampling methods and SMOTE family approaches performs better because of its synthetic sampling instead of the nearest values. The authors showed that, the model Logistic regression performs best with an accuracy rate of 97.0% and corresponding precision of 99.99% and it performs well among the five techniques used in this system.

The authors S. Makki, Z. Assaghir, Y. Taher, R. Haque, M. Hacid and H. Zeineddine have analyzed and explained that fraudulent transactions leads to heavy financial loss. The existing methods and models to prevent these fraudulent transactions are very expensive, time consuming and requires man power, and many researchers are working on these detection methods to find innovative ways to control the unauthorized transactions. Finally, after rigorous amount of efforts the authors have concluded that the main reason for ineffective results is that the dataset is imbalanced. So they have used a balanced dataset to train and test the models.

In a paper, the authors Ishan Sohony, Rameshwar Pratap, and Ullas Nambiar have conducted their research and proposed that applying ensemble techniques to detect fraudulent transactions is more effective comaparatively. In their study, they have concluded that Random forest performs better and is the best suited model with high accuracy and the neural networks models performs well in detecting the fraudulent activities. In this paper, they have conducted their research using larger datasets and the ensemble technique they have proposed is the hybrid mix of Random forest and neural networks .

Debachudamani Prusti and Santhnu Kumar Rath implemented a system using various machine learning algorithms such as Decision Tree, K-Nearest neighbour model, Extreme Learning Machine, Multilayer Perceptron and Support Vector Machine in order to identify the fraudulent transactions. They have proposed an innovative framework by combining all the above mentioned models. In addition to these machine learning models, they have used two web-based protocols like SOAP and REST to exchange the data efficiently across different platforms. They have concluded that SVM model performs better with an accuracy of 81.63% compared to other algorithms and have used accuracy as the performance metric to evaluate these five models. But the hybrid model implemented by them using five algorithms performs better than SVM model with an accuracy of 82.58% .

M. S. Kumar, V. Soundarya, S. Kavitha, E. S. Keerthika and E. Aswini proposed a framework to detect these fraud transactions of credit cards with the use of Random forest models. In general, Random Forest Classifier is a machine learning supervised model which works based on decision tree approach to identify fraudulent transactions and their classification into real and fake transactions. Their performance is measured using confusion matrix and in their analysis the proposed model performs well with an accuracy of 90% .

In a paper, the authors J. Esmaily and R. Moradinezhad, have implemented a framework which is a combination of Artificial Neural Networks and Decision Tree algorithms. They have implemented the approach as two phase in which the first phase is about collecting the results of decision tree model and by using Multilayer Perceptron a new dataset is generated. This new dataset acts as input to the neural networks to finally categorize the data. This models ensures about reliability by producing a low false detection count .

Siddhartha Bhattacharyya and 4 others in their paper have conducted a detailed research on comparative analysis of three different models such as Support vector Machine, Random Forest and logistic Regression. The outcomes of their analysis shows that among three algorithms, Random Forest technique performs better with high accuracy results and then Logistic regression followed by Support Vector Machine.

Y. Sahin and E. Duman implemented a framework to identify fraudulent transactions using a hybrid model implemented by combining SVM model and Decision Tree model. The results of their analysis proves that Decision tree performs better for smaller datasets whereas SVM model performs better for larger datasets equally with decision trees.

Abhimanyu Roy et al. have implemented a framework using deep learning models to identify the fraudulent transactions. The proposed framework is build based upon artificial neural networks with in the required time and considering parameters such as memory components such as long term and short term memory and also includes several other parameters. By using these segments in identifying fraudulent transactions, almost 80 million online money transfers are classified as real and fake transactions. In this framework they have used a distributed cloud computing environment. The framework proposed by the authors gives an effective information about the sensitivity analysis of the proposed metrics in identifying the fraudulent transactions.

Shiyang Xuan et al. have implemented a random forest model in two ways to identify the fraudulent transactions based on their behaviour. Further, they have conducted a comparative study on this two types of random forests models which are distinguished based on their performance and classifiers in detecting the transactions as legal or not. To train and test the models, the authors have used the data an e-commerce company of China and assessed the performance of these two models. In this study, they have utilized B2C dataset in identifying the transactions as fraud or real. Hence, in the final results of the authors proved that the proposed framework performs well for smaller datasets and is not suitable for imbalanced datasets and becomes ineffective for any such kind of dataset.

Yusuf sahin and Ekrem Duman have conducted a comparative research on the performance of deep learning models such as Artificial Neural Network(ANN) and Logistic Regression model which is a machine learning model. In addition to numerous transactions, there are number of classifications are exist in credit card transactions. In their study, the authors have concluded that Logistic Regression performs well comparatively than Artificial Neural Networks. Further they have carried out research on addressing the imbalanced dataset issues.

In a paper, Leila et.al researched about a strategy of conglomerating profile which misuses the characteristic designs in time arrangement of exchanges, and the extortion location is performed online at the conclusion of a day or at the conclusion of a period individually. In their proposal, they have used and compared the performance of several techniques like Support Vector Machine and Random Forest classifier in detecting the fraudulent transactions of the credit card. Finally, their outcomes shows that random forest performs better in detecting the fake transactions and concluded that among all the techniques Random Forest has best performance metrics in processing with the aggregation.

To detect credit card fraud transactions, a web service-based application is proposed which is a result of machine learning models research analysis. To identify these fraudulent transactions and monitor customer’s behaviour, genetic algorithm is proposed. Based on this algorithm calculation, an innovative fraud detection system is proposed which is flexible to behaviour changes of customers and is an hybrid framework with the mix of classification and clustering approaches, and this scalable algorithm is known as BOAT .

Decision trees and Support Vector Machine (SVM), combined strategy of Choice tree, Neural Systems, Logistic regression, Self-Organizing Map (SOM) combined with Gaussian work, and Fluffy logic combined with Self-organizing map have been presented for money related extortion location strategy. A combined strategy of SVM, Random forests, Logistic regression, Self-Organizing Map Neural Network (SOMNN), Genetic Calculation with behavior based method and Hidden Markov Model (Hmm) has been attempted.

There exists many machine learning techniques to detect fraudulent transactions and many new approaches are coming with the advancement of artificial intelligence, such as proposed a new framework to improve the performance of fraud detection techniques. Totally ten models of machine learning were proposed and trained, tested to check for the accuracy and efficiency in detecting fraudulent transactions. In this study, the authors have used car insurance data claims. After rigorous research and analysis, the authors have concluded that random forest gives best results among all other algorithms in detecting fraud activities.

In a paper, the authors have proposed a framework focusing on imbalanced dataset issue . This study proposes a fraud detection purposes in insurance field to detect fraud activities. Here, the objective is to implement a system known as insurance fraud detection using machine learning models such as Decision Tree, Support Vector Machine, and Artificial Neural Networks with the use of sampling techniques to balanced the dataset. The objective of this approach is to balance the majority and minority classes and to train the models to get the results accurately and efficiently. Finally they have concluded that Decision Tree model performs well compared to other algorithms and it gives better performance with sampling techniques than with original imbalanced data.

Various machine learning algorithms are trained and evaluated as part of fraud detection. In a paper, the authors have proposed few additional algorithms of machine learning were proposed such as support vector machine along with random forest and logistic regression with neural networks. The main objective of this approach is to improve the system of fraud detection and to get the better accuracy results. The fraud detection system which were developed are facing many troubles due to changes in behavioural patterns of the models. Due to these changes it is quite complex to identify which is fake and genuine. The data of credit card transactions of customers is confidential and is extremely imbalanced which is an issue to train and test the models on this imbalanced dataset. Therefore, the sampling approaches came into picture to balance this imbalanced dataset and to train and test the models to measure the accuracy.

There are so many researches and analysis carried out by authors to find effective solutions to credit card fraud detection. These strategies are included with various techniques and are not limited to, neural networks, IDE, Bayesian Network, Optimization algorithms, Meta-Learning agents, Artificial Intelligence, Image processing, Constitution-based systems, Logistic Regression, Support Vector Machine(SVM), Decision Tree, KNN, Adaptive learning and so on. To process the real-time payment processing applications, the structure of neural networks is primarily used with the help of unsupervised techniques. With the help of optimal classification, the issues from every correlated community and self-organizing graph of neural networks solves these kind of problems. With more than fair 95 percent of the whole location framework of ROC request bend false without really causing any other wrong alert gathering cast learning (too regularly broadly known as meta-classifier) upgrades comes about through combining diverse learning calculations optimization calculations to improve measurable comes about.

In a paper, S. Makki, Z. Assaghir, Y. Taher, R. Haque, M. Hacid and H. Zeineddine have explained that the fraudulent activities in credit cards creates heavy financial loss. So many approaches and researches are going on provide best solutions to this credit card fraud issue and to minimize the financial loss to customers and banks as well.But often these methods are very costly and time consuming and requires huge man power. In their research, authors have found that the imbalanced dataset is the major drawback of this system and it leads ineffective results after careful consideration by the researchers in their studies. Using few technical approaches they have used balanced dataset to train and test the models in their experiments and concluded that Logistic Regression, C5.0, Decision Tree and Support Vector Machine, Artificial Neural Networks are the best models with better accuracy, ROC\_AUC score and sensitivity results as well.

It is clear that imbalanced dataset is the major concern and drawback for inaccurate results in credit card fraud detection. So balancing this dataset with sampling approaches is the main objective. Two types of sampling techniques such as random under sampling which balances the data by removing the majority class data [37] and another technique is random over sampling which balances the data by duplicating the data with the same representation of the existing data. There are several advanced techniques are used such as SMOTE techniques, ADASYN, BORDERLINE and so on, which are oversampling techniques to create synthetic data using KNN. It is better to add synthetic data to get better results using SMOTE techniques than just removing or replicating the instances using sampling approaches.

In a paper, the authors have implemented a new approach in sampling techniques which referred as Moving to Adaptive Samples (MASI) in imbalanced dataset and they have concluded that this approach performs better compared to other sampling approaches like random undersampling, random oversampling and synthetic minority oversampling techniques (SMOTE). The sampling approaches SMOTE and undersampling classifier created net instances of data and balances the class distribution before implementation, whereas the MASI approach adds new instances of data based on the density distribution of the imbalanced data and increases the size of the minority class with the change in class labels. The analysts demonstrate this decreases the inclination of the classifier because it moves the tests in minor lesson closer to the choice boundary. Alternatively, the other way to handle this imbalanced data issue is to implement ensemble learning in the framework at algorithmic level. The main functionality of ensemble methods is to reduce the variance in the data using multiple classifiers with the help of bagging and boosting techniques. In bagging approach, various weak classifiers are trained on multiple subsets of the majority and minority classes in the initial phases after which the final classifier is structured with the use of all those weak classifiers. Similarly, using AdaBoost method the same strategy can be applied in different classification problems and it disposes of the require for investigating an ideal course balance ratio whereas reducing the data loss which can be caused by RUS, and overfitting issue caused by ROS and Smote strategies.

In a research, the authors have proposed a new framework which changes the cost function of SVM which yields a cost-sensitive version. This model was trained using 21 datasets from KDD98 which is not included with fraud detection dataset. Moreover, the authors have compared the outcomes with balanced data results which are generated using SMOTE and they have used AUC curve as their metric to compare the results. Finally, they have concluded that this proposed framework produces better results with maximum number of datasetss.

Hensman and Masko have analysed the results of ROS on imbalanced class distribution by creating 10 imbalanced datasets with the help of CIFAR-10 dataset. The ROS technique balances the dataset by replicating the minority class instances randomly until all the classes are in same size and balanced, any way if a class size is less than the size of the largest class is still considered as a minority class. After analysing the results of balanced and imbalanced datasets, the authors have concluded that the balanced dataset with oversampling techniques achieves good results with better performance compared to the results produced with imbalanced dataset which is the original dataset. In another research of Buda et al. explained about the similar results obtained when compared ROS against RUS. Using these ROS and RUS methods, Buda et al. produces balanced datasets with synthetic datasets by oversampling the minority class instances and unsersampling majority class instances in order to balance the data with equal class distribution.

In another research paper, the authors have explained about the essential features of weighted distribution in ADASYN for various minority class instances as per their level of trouble in learning, such as most of the synthetic data generated for minority class instances which are harder to learn are compared against the minority class instances which are easier to learn. Therefore, the ADASYN method enhances its areas of learning regarding the data distribution mainly in two ways: 1) decreasing the partiality nature of class distribution brought by imbalanced class distribution, and 2) Gradually, moving the classification decision boundary regarding the difficult cases. Simulation investigations on a few machine learning information sets appear the adequacy of this strategy over five assessment measurements.

# Methodologies



Many researches and experiments were carried out from so many years on credit card fraud detection. The researchers have identified and proposed various frameworks to detect these fake transactions using machine learning models and deep learning models. As the dataset is imbalanced, the proposed approaches are facing challenges in producing better results and accurate performance in identifying fraudulent transactions. In this paper, the literature review and research indicates that sampling approaches and SMOTE techniques plays key role producing accurate results with better accuracy score for the proposed credit card fraud detection system.

In this paper, mainly four models are trained and tested against the balanced datasets which are created using sampling approaches and SMOTE techniques such as Random Under Sampling(RUS), Random Over Sampling(ROS), Adaptive Synthetic SMOTE, BORDERLINE SMOTE, SMOTE-ENN and SVMSMOTE. Therefore, all the four models are trained and tested with the balanced dataset created using SMOTE family and the results are compared with all possible combinations of algorithms with SMOTE techniques to identify the model that produces better results with corresponding SMOTE technique. After comparing the results of all the models with the balanced data, it was observed that the **Logistic Regression**(LR) model with the sampling technique **SMOTE-ENN** performs better and produces effective results compared to other models. It produces an accuracy rate of 0.98 and ROC\_AUC score is 0.941. As the dataset is imbalanced, the suggested metric to be consider while comparing the results is ROC\_AUC curve. Followed by Logistic Regression, **Support Vector Machine(SVM)** with **SMOTE-ENN** approach performs better with an accuracy rate of 0.98 and ROC\_AUC score is 0.936. More details about the algorithms, SMOTE family techniques used in this study and their implementation are explained in detail here.

### Logistic Regression

Increasingly measurable models were connected at data mining errands incorporate regression analysis, different discriminant investigation, logistic regression, and Probit strategy, etc [16][17]. In situations when we are willing to predict the probability of a feature or outcome of an instinct on the basis of values of a set of predictor variables, then in those situations Logistic Regression performs well. Logistic regression model and Linear regression model are similar to each other but the linear regression model is suitable for models where the dependent variable is dichotomous. The applications of Logistic regression models are like to predict the odds ration for all the dependent variables in the model and it used to wide range of study and research cases than discriminant analysis. Figure 03 represents the generic graph representation of Logistic Regression model functionality:

Chart

Description automatically generated with medium confidence

**Figure 03: Linear Regression vs Logistic Regression Graph [26]**

Two probability methods such as Linear probability and multivariate conditional probability models (Logit and Probit) were proposed into the business failure prediction study and these methods plays a key role in predicting the probability of an organization’s failure [18].

### Random Forest Classifier

These algorithms which are tree based are most common in in machine learning models used to lead the various issues. These measurements are flexible and can illustrate any type of problem easily. The RF algorithms are able to use the highlights or the type for categorical highlights when preparing for forecasts on making tests within the categories they have to place in. Moreover, they gives expectations with **tall** **precision**, **stability**, and **ease of interpretation**. Figure 04 represents the structure of Random Forest model:

Chart, radar chart

Description automatically generated

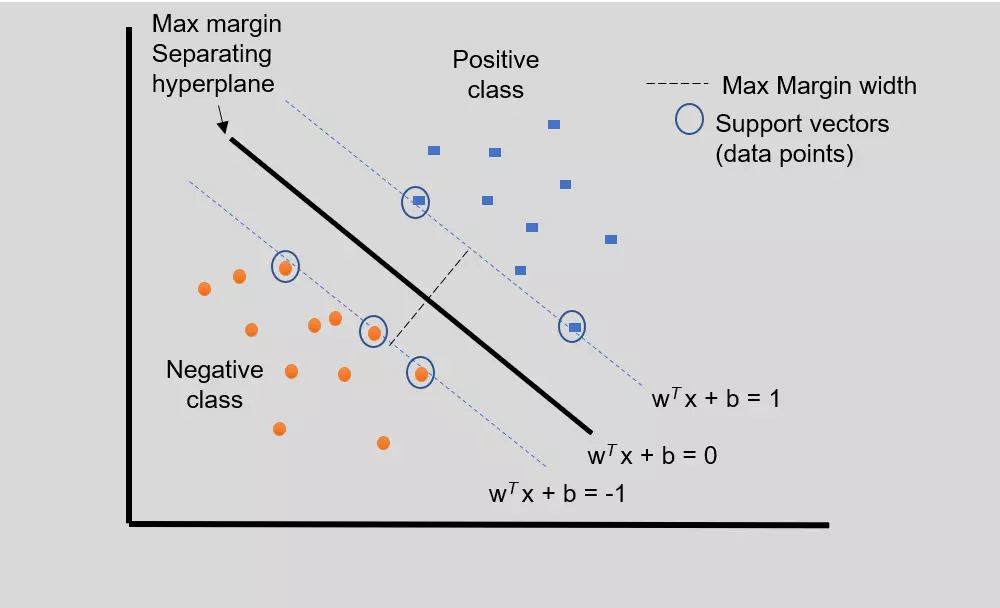
**Figure 04: Representation of Random Forest Classifier**

### Support Vector Machine

Support vector machine may be a strategy utilized in pattern recognition and classification. Support Vector Machine (SVM) could be a supervised machine learning strategy utilized for classification and regression assignments. SVM performs two-class or multi-class information classification by allotting the class names to the perceptions.

The objective of SVM is to outline the input dataset into high-dimensional space and make a decision boundary (isolating hyperplane) by learning to accurately classify the classes. (Isolating) Hyperplane can be characterized as the direct line in a high-dimensional space.

The Hyperplane partitions the input information (preparing information) in a such way that information focuses from one course will be on the same side than data focuses from another class, and maximize the separate between edges. Moreover, the distance between the hyperplane and the closest information point from each course is maximal. Consequently, SVM is additionally known as Maximum Margin Classifiers. Figure 05 shows the graphical representation of the SVM model:



**Figure 05: Representation of SVM**

### Artificial Neural Networks

ANNs are naturally motivated computer programs outlined to reenact the way in which the human brain forms data . ANNs assemble their knowledge by recognizing the designs and relationships in data and learn (or are trained) through encounter, not from programming, and there lies the essential contrast between ANNs and other classical computer programs. Figure 6 and 7 represents the generic model structure of artificial neural networks:

Diagram

Description automatically generated

**Figure 06: Generic ANN model [38]**

Diagram

Description automatically generated

**Figure 07: Conceptual model for ANN [37]**

## SMOTE Techniques

### Random Undersampling

Random Undersampling is the inverse to Random Oversampling. This strategy looks for to randomly select and expel samples from the majority lesson, thus decreasing the number of illustrations within the lion's share course within the changed information. The result of undersampling could be a transformed information set with less cases within the lion's share course — this handle may be rehashed until the number of cases in each class is break even with. Data distribution when RUS is applied can be shown in Figure 08:

A picture containing chart

Description automatically generated

**Figure 08: Illustration of Random UnderSampling Technique [27]**

### Utilizing this approach is compelling in circumstances where the minority class incorporates a adequate sum of cases in spite of the serious imbalance. On the other hand, it is continuously vital to consider the prospects of profitable data being erased as we arbitrarily expel them from our information set since we have no way to identify or store the illustrations that are data wealthy within the larger part of the class.

### Random Oversampling

Oversampling is for the most part utilized more regularly than undersampling, particularly when the point by point information has however to be collected by overview, meet or something else. Undersampling is utilized much less as often as possible. Overabundance of as of now collected information got to be an issue as it were within the "Big Data" time, and the reasons to utilize undersampling are primarily down to earth and related to asset costs. Particularly, whereas one needs a appropriately huge test estimate to draw substantial statistical conclusions, the data must be cleaned some time recently it can be utilized. Cleansing ordinarily includes a critical human component, and is regularly particular to the dataset and the analytical issue, and so takes time and money. Figure 09 represents the outline of the ROS technique:

Diagram

Description automatically generated

**Figure 09: Representation of Random Oversampling [28]**

### ADASYN

Within the handle of model training, the quality of the dataset will too influence the expectation execution of the model. Data preprocessing may be a essential step some time recently training, which includes oversampling the initial imbalanced information tests. ADASYN (Adaptive Synthetic Sampling) was utilized in this study. The calculation gives diverse weights to diverse minority samples and utilizes

Chart, scatter chart

Description automatically generated

**Figure 10: Representation of ADASYN sampling technique [29]**

a few component to consequently decide how numerous composite samples each minority test should deliver in arrange to realize the objective of information balance. ADASYN isn't like Smote, which synthesizes the same number of samples for each minority test. Figure 10 represents the differences between two data plots of original dataset and adasyn dataset:

### BORDERLINE

Hui Han et. at. In 2005 [3] has introduced a variation in SMOTE approach which is known as Borderline-SMOTE. This SMOTE technique mainly aims to create synthetic data by taking the samples which are able to form the border that separates one class from another. Hence, this SMOTE technique helps to identify the samples that are on the border of the class space and applies BORDER-SMOTE technique to these samples. Borderline-SMOTE may be a variety of the Smote. Similar to the title infers, it has something to do with the border. So, not at all like with the Smote, where the manufactured information are made arbitrarily between the two information, Borderline-SMOTE as it were makes synthetic data along the choice boundary between the two classes. Also, there are two sorts of Borderline-SMOTE. Figure 11 represents the Borderline-SMOTE technique functionality through graphs:

Diagram, scatter chart

Description automatically generated with medium confidence

**Figure 11: Borderline SMOTE representation [30]**

### SMOTE

Synthetic Minority Over-sampling Method (Smote) was presented by Nitesh V. Chawla et. to the. in 2002. Smote is an over-sampling method centered on producing engineered tabular information. The common thought of Smote is the era of synthetic information between each test of the minority lesson and its “k” closest neighbors. That's , for each one of the tests of the minority course, its “k” closest neighbors are found (by default k = 5), at that point between the sets of focuses produced by the test and each of its neighbors, a unused synthetic information is created. Figure 12 represents the visual description of SMOTE technique:

Scatter chart

Description automatically generated with medium confidence

**Figure12: Visual description of SMOTE [31]**

### SMOTE-ENN

The method of SMOTE-ENN algorithm: (A) Smote chosen each test from the minority tests progressively as the root test for the blend of the modern test. (B) The taking after result was gotten by utilizing ENN to eliminate noise tests when the method of Smote is caused.

Chart, scatter chart

Description automatically generated

**Figure 13: Process of SMOTE-ENN algorithm [38]**

### SVM-SMOTE

SVM Smote centers on expanding minority focuses along the choice boundary. The argument behind usually that occurrences around this boundary are basic for assessing the optimal decision boundary (which contrasts with the K-Means strategy we saw prior but adjusts with the Borderline variation). Figure 14 represents the comparison scenario of various SMOTE techniques:

Chart, scatter chart

Description automatically generated

**Figure 14: Comparison of different SMOTE algorithms [32]**

## Dataset

A dataset is fundamentally a collection of related information. In this paper, we make utilize of a freely accessible imbalanced dataset. An imbalanced dataset is one in which difference happens within the dependent variables. Imbalanced infers that there's an unequal dispersion of classes. The specific dataset that we utilize is additionally an imbalanced one. This specific dataset contains the record of exchanges made by European cardholders. It has the records of 284,807 exchanges made over a span of two days, out of which 492 were found out to be extortion. The rate of false exchanges is found out to be extremely low. This dataset is profoundly unequal. Since giving exchange details of a client is considered to issue related to privacy, in this manner most of the highlights within the dataset are changed utilizing principal component analysis (PCA). V1, V2, V3,..., V28 are PCA connected features and rest i.e., ‘time’, ‘amount’ and ‘class’ are non-PCA connected highlights, as appeared in Figure 15.

Text

Description automatically generated

**Figure 15: Feature description of the Dataset**

Commonly, fraud location issue is considered as a data mining issue. Data mining is the method of finding significant unused relationships, designs and patterns by filtering through huge sums of information put away in repositories, utilizing pattern recognition innovations as well as factual and numerical strategies. The discovery of credit card fraud is for the most part received a classification show. Figure.16 appears the out line of the show

Graphical user interface, text, application

Description automatically generated

**Figure 16: Outline of the Model**

### Data Pre processing and Selection

Firstly the highlights utilized within the dataset are changing over into numerical data. Feature determination may be a exceptionally imperative organize in extortion discovery. The features within the data productively depict the utilization of behavior of an individual credit card account. In this model, the features which interpret the behavior of the client as it were are chosen for detection. Including irreverent features make the classifier inefficient. Figure 17 represents the class distribution of imbalanced dataset;

Graphical user interface, application, Word

Description automatically generated

**Figure 17: Class distribution with unbalanced data**

### Metrics Evaluation

**Confusion Matrix**

The metrics which is straightforward measurements and normal approach is confusion matrix which can be used to calculate other metrics such as accuracy, precision, recall and specificity. Confusion matrix doesn’t have any metric itself but it contains the information to calculate the metrics with the data inside it as shown in Figure 18.

Table

Description automatically generated with low confidence

**Figure 18: Confusion Matrix**

**True Positives (TP):** True positives are the instances identified as genuine positives both in model and original dataset.

**True Negatives (TN):** True negatives are the instances identified as negatives both in model and original dataset.

**False Positives (FP):** False positives are the instances identified as genuine positives in model but negatives instances in original dataset.

**False Negatives (FN):**False negatives are the instances identified as negatives in model but positive instances in original dataset.

**Accuracy**

In general, accuracy can be defined as the count of genuine positive instances in total number of instances. In terms of confusion matrix, Accuracy can be measured as the sum of true positives and false negatives division with the true positives, false negatives, false positives and true negatives. The representation of accuracy using confusion matrix is shown in Figure 19 and its mathematical formula are as follows:

**Accuracy = (TP+FN) / (TP+FP+FN+TN)**



**Graphical user interface, application

Description automatically generated**

**Figure 19: Accuracy Metrics**

* **Precision**

In general, precision can be defined as the count of genuine positive instances in total number of positive instances. In terms of confusion matrix, precision can be measured as the sum of true positives division with the true positives, false positives. The representation of precision using confusion matrix is shown in Figure 20 and its mathematical formula are as follows:

**Precision = (TP) / (TP+FP)**



**Graphical user interface, application

Description automatically generated**

**Figure 20: Precision Metrics**

* **Recall/Sensitivity**

In general, recall can be defined as the count of genuine positive instances in total number of true positive and false negative instances. In terms of confusion matrix, recall can be measured as the sum of true positives division with the true positives, false negatives. The representation of recall using confusion matrix is shown in Figure 21 and its mathematical formula are as follows:

**Recall = (TP) / (TP+FN)**



**Graphical user interface, application

Description automatically generated**

**Figure 21: Recall Metrics**

* **Specificity**

In general, specificity can be defined as the count of genuine positive instances in total number of false positive and true negative instances. In terms of confusion matrix, specificity can be measured as the sum of true positives division with the false positives, true negatives. The representation of specificity using confusion matrix is shown in Figure 22 and its mathematical formula are as follows:

**Specificity = (TP) / (FP+TN)**



**Table

Description automatically generated with medium confidence**

**Figure 22: Recall Metrics**

* **F1 Score**

It depends on the model and the application, whether to choose recall or precision as our metric. But there are cases where both are necessary. Hence, combining these two metrics into a single metric works out in such cases. The metric which combines both precision and recall is known as F1-score, and is represented in a mathematical formula as follows:

**F1-score= 2\*Precision\*Recall/(Precision+Recall)**

* **AUC-ROC Curve**

The ROC curve stands for Receiver Operating Characteristic curve which indicates the performance of the classifier as a function of its threshold. Mainly, it displays the TPR which means True Positive Rate against FPR which means False Positive Rate for different threshold values. This curve mainly finds out the TPR and FPR values and plots them as curves.

### Implementation-SMOTE Techniques

As soon as the sampling techniques are applied, the imbalanced dataset will turn into balanced dataset and is explained below for each sampling technique. In the graph, ‘0’ represents non fraud and ‘1’ represents fraud. The graphs are plotted between class and frequency as axis.

1. **Random Undersampling**

The following Figure 23 represents the data representation after applying the undersampling technique to the imbalanced dataset. As a result of undersampling, the number of instances of majority classes are reduced and balanced with minority class.

Chart, bar chart

Description automatically generated

**Figure 23: Data distribution of RUS**

1. Random Oversampling

The following Figure 24 represents the data representation after applying the oversampling technique to the imbalanced dataset. As a result of oversampling, the number of instances of minority classes are duplicated and balanced with majority class.

Chart, bar chart

Description automatically generated

**Figure 24: Data distribution of ROS**

1. **ADASYN**

The following Figure 25 represents the data representation after applying the ADASYN technique to the imbalanced dataset. As a result of this technique, it creates the instances of minority class which are ‘harder to learn’ instead of just replicating minority instances and will balance with majority class instances.

Chart, bar chart

Description automatically generated

**Figure 25: Data distribution of ADASYN**

1. BORDERLINE

The following Figure 26 represents the data representation after applying the BORDERLINE technique to the imbalanced dataset. As a result of this technique, the number of instances of which creates the border to separate the classes are created in minority classes and balanced with majority class.

Chart, bar chart

Description automatically generated

**Figure 23: Data distribution of BORDERLINE**

1. **SMOTE**

The following Figure 27 represents the data representation after applying the oversampling technique to the imbalanced dataset. As a result of SMOTE, the number of new instances are created from the existing minority classes and balanced with majority class.

Chart, bar chart

Description automatically generated

**Figure 27: Data distribution of SMOTE**

1. **SVM-SMOTE**

The following Figure 28 represents the data representation after applying the oversampling technique to the imbalanced dataset. As a result of SVM-SMOTE,it uses SVM algorithm to create new samples of data and add to minority classes and balanced with majority class.

Chart, bar chart

Description automatically generated

**Figure 28: Data distribution of SVM-SMOTE**

1. **SMOTE-ENN**

The following Figure 29 represents the data representation after applying the oversampling technique to the imbalanced dataset. As a result of SMOTE-ENN, it measures the distance between any random sample to that of k nearest neighbor samples and multiples this distance to create more samples in minority class and balanced with majority class.

Chart, bar chart

Description automatically generated

**Figure 29: Data distribution of SMOTE-ENN**

# Results

In this paper, various algorithms with SMOTE family approach are analysed to identify the fraudulent transactions in credit card operations and the results are compared to find out the best model with better performance and high accuracy. SMOTE techniques plays a vital role for better performance through balancing the data either by oversampling or undersampling the data since the original data is imbalanced. Various performance metrics are explained in methodologies section to compare the results, among these ROC-AUC curve is the best metric to compare as the data is originally imbalanced. In addition to this evaluation, other metrics such as accuracy, precision and recall are also calculated and compared in this paper. Therefore, the comparison of results are represented in various ways such as numerically with percentage of accuracy and AUC score and by plotting the ROC curves and confusion matrix for better understanding. These results and diagrams are explained below for Logistic Regression(LR), Support Vector Machine(SVM), Random Forest Classifier(RF) and Artificial Neural Networks(ANN) applied with SMOTE family techniques such as Random Undersampling(RUS), Random Oversampling(ROS), Adaptive Synthetic(ADASYN), BORDERLINE, Synthetic Minority Oversampling Technique(SMOTE), SVM-SMOTE, SMOTE-ENN.

**1)Imbalanced Dataset**

The following figure shows the data distribution of imbalanced dataset of credit card transactions. Here Class 0 represents genuine transactions whereas Class 1 represents fake transactions. Both classes are not equally distributed since the real transactions are around 2,84,315 and 492 fake transactions. In the below figure, the area occupied by blue dots is more compared to that of red which indicated fake transactions. Figure 30 represents the data distribution of imbalanced dataset:

**Graphical user interface, application, Word

Description automatically generated**

**Figure 30: Visual representation of distribution of Imabalnced data**

Another way of representing the data distribution of imbalanced data of credit card transactions is through data bars as follow. The following figure represents the imbalanced dataset and this graph is known as Histogram.

Graphical user interface, application, Word

Description automatically generated

**Figure 31: Fraud Class Histogram**

Here the blue bar indicates the real transactions and the fake transactions bar is negligible comparatively which is not clearly visible in this graph. This graph is plotted across class and frequency as axis. X-axis represents the Class which is 0 and 1 and Y-axis represents frequency which is the count of the transactions and the name of this histogram is Fraud Class histogram which is shown in the top of the graph.

**Logistic Regression**

The following figures represents the confusion matrix and ROC curve of the model Logistic Regression with original dataset which is imbalanced. The confusion matrix is plotted with predicted label and actual label as its axis and the colour bar along with confusion matrix describes the shades according to the transactions count and the ROC curve is plotted with True Positive Rate and False Positive Rate as axis and the figure also displays the AUC score of the corresponding model. As explained in earlier section confusion matrix consists of four values which are described as follows for this model:

**True Positive:** There are totally 85,288 transactions which are identified as true positive.

**True Negative:** There are totally 80 transactions which are identified as true negative.

**False Negative:** There are totally 64 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 11 transactions which are identified as fraud transactions but originally these are genuine transactions.

Chart

Description automatically generated Chart, line chart

Description automatically generated

**Figure 32: Confusion Matrix and AUC-ROC curve of LR**

Figure 32 represents ROC curve and AUC score with 0.78. Following are the evaluation metrics for Logistic Regression Model with imbalanced dataset:

**Accuracy :** 0.9991

**Precision :** 0.9992

**Recall :** 0.9998

**AUC Score:** 0.7777

**Random Forest**

The following figures represents the confusion matrix and ROC curve of the model Random Forest Classifier with original dataset which is imbalanced. The confusion matrix is plotted with predicted label and actual label as its axis and the colour bar along with confusion matrix describes the shades according to the transactions count and the ROC curve is plotted with True Positive Rate and False Positive Rate as axis and the figure also displays the AUC score of the corresponding model. As explained in earlier section confusion matrix consists of four values which are described as follows for this model:

**True Postive:** There are totally 85,291 transactions which are identified true positive.

**True Negative:** There are totally 114 transactions which are identified true negative.

**False Negative:** There are totally 30 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 8 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 33: Confusion Matrix and AUC-ROC curve of RF**

Figure 33 represents ROC curve and AUC score with 0.90. Following are the evaluation metrics for Random Forest Classifier with imbalanced dataset:

**Accuracy :** 0.9983

**Precision :** 0.9996

**Recall :** 0.9999

**AUC Score:** 0.8957

**Support Vector Machine**

The following figures represents the confusion matrix and ROC curve of the model Support Vector Machine with original dataset which is imbalanced. The confusion matrix is plotted with predicted label and actual label as its axis and the colour bar along with confusion matrix describes the shades according to the transactions count and the ROC curve is plotted with True Positive Rate and False Positive Rate as axis and the figure also displays the AUC score of the corresponding model. As explained in earlier section confusion matrix consists of four values which are described as follows for this model:

**True Positive:** There are totally 85,289 transactions which are identified as true positive

**True Negative:** There are totally 46 transactions which are identified as true negative.

**False Negative:** There are totally 98 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 10 transactions which are identified as fraud transactions but originally these are genuine transactions.

Chart

Description automatically generated Chart, line chart

Description automatically generated

**Figure 34: Confusion Matrix and AUC-ROC curve of SVM**

Figure 34 represents ROC curve and AUC score with 0.66. Following are the evaluation metrics for Support Vector Machine Model with imbalanced dataset:

**Accuracy :** 0.9983

**Precision :** 0.9988

**Recall :** 0.9998

**AUC Score:** 0.6596

**Artificial Neural Networks**

The following figures represents the confusion matrix and ROC curve of the model Artificial Neural Networks with original dataset which is imbalanced. The confusion matrix is plotted with predicted label and actual label as its axis and the colour bar along with confusion matrix describes the shades according to the transactions count and the ROC curve is plotted with True Positive Rate and False Positive Rate as axis and the figure also displays the AUC score of the corresponding model. As explained in earlier section confusion matrix consists of four values which are described as follows for this model:

**True Positive:** There are totally 82,965 transactions which are identified as true positive.

**True Negative:** There are totally 131 transactions which are identified as true negative.

**False Negative:** There are totally 13 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 2334 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart, treemap chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 35: Confusion Matrix and AUC-ROC curve of ANN**

Figure 35 represents ROC curve and AUC score with 0.57. Following are the evaluation metrics for Artificial Neural Networks model with imbalanced dataset:

**Accuracy :** 0.9994

**Precision :** 0.8613

**Recall :** 0.8194

**AUC Score:** 0.57

**2)Random Under Sampling**

**Logistic Regression**

The following figures represents the confusion matrix and ROC curve of the model Logistic Regression with synthetic dataset generated using Random Undersampling technique. fraud transactions and they are described as fraud in the original dataset as well.

**True Positive:** There are totally 82,965 transactions which are identified as true positive.

**True Negative:** There are totally 131 transactions which are identified as true negative.

**False Negative:** There are totally 13 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 2334 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart, treemap chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 36: Confusion Matrix and AUC-ROC curve of LR**

Figure 36 represents ROC curve and AUC score with 0.94. Following are the evaluation metrics for Logistic Regression Model with balanced dataset:

**Accuracy :** 0.9725

**Precision :** 0.9991

**Recall :** 0.9761

**AUC Score:** 0.9411

**Random Forest**

The following figures represents the confusion matrix and ROC curve of the model Random Forest Classifier with synthetic dataset generated using Random Undersampling technique.

**True Positive:** There are totally 82,785 transactions which are identified as true positive.

**True Negative:** There are totally 127 transactions which are identified as true negative.

**False Negative:** There are totally 17 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 2514 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart, treemap chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 37: Confusion Matrix and AUC-ROC curve of RF**

Figure 37 represents ROC curve and AUC score with 0.93. Following are the evaluation metrics for Random Forest Classifier with balanced dataset:

**Accuracy :** 0.9703

**Precision :** 0.9991

**Recall :** 0.9761

**AUC Score:** 0.9262

**Support Vector Machine**

The following figures represents the confusion matrix and ROC curve of the model Support Vector Machine with synthetic dataset generated using Random Undersampling technique.

**True Positive:** There are totally 80,740 transactions which are identified as true positive.

**True Negative:** There are totally 91 transactions which are identified as true negative.

**False Negative:** There are totally 53 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 4559 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart, treemap chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 38: Confusion Matrix and AUC-ROC curve of SVM**

Figure 38 represents ROC curve and AUC score with 0.79. Following are the evaluation metrics for Support Vector Machine with balanced dataset:

**Accuracy :** 0.9460

**Precision :** 0.9991

**Recall :** 0.9561

**AUC Score:** 0.7892

**Artificial Neural Networks**

The following figures represents the confusion matrix and ROC curve of the model Artificial Neural Networks with synthetic dataset generated using Random Undersampling technique.

**True Positive:** There are totally 85,162 transactions which are identified as true positive.

**True Negative:** There are totally 118 transactions which are identified as true negative.

**False Negative:** There are totally 26 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 137 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 39: Confusion Matrix and AUC-ROC curve of ANN**

Figure 39 represents ROC curve and AUC score with 0.91. Following are the evaluation metrics for Artificial Neural Networks model with balanced dataset:

**Accuracy :** 0.9980

**Precision :** 0.9997

**Recall :** 0.8472

**AUC Score:** 0.9089

**3)Random Over Sampling**

**Logistic Regression**

The following figures represents the confusion matrix and ROC curve of the model Logistic Regression with synthetic dataset generated using Random Oversampling technique.

**True Positive:** There are totally 83,331 transactions which are identified as true positive.

**True Negative:** There are totally 130 transactions which are identified as true negative.

**False Negative:** There are totally 14 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 1968 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart, treemap chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 40: Confusion Matrix and AUC-ROC curve of LR**

Figure 40 represents ROC curve and AUC score with 0.94. Following are the evaluation metrics for Logistic Regression with balanced dataset:

**Accuracy :** 0.9865

**Precision :** 0.9998

**Recall :** 0.9879

**AUC Score:** 0.9398

**Random Forest**

The following figures represents the confusion matrix and ROC curve of the model Random Forest Classifier with synthetic dataset generated using Random Oversampling technique.

**True Positive:** There are totally 85,292 transactions which are identified as true positive.

**True Negative:** There are totally 113 transactions which are identified as true negative.

**False Negative:** There are totally 31 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 7 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 41: Confusion Matrix and AUC-ROC curve of RF**

Figure 41 represents ROC curve and AUC score with 0.90. Following are the evaluation metrics for Random Forest Classifier with balanced dataset:

**Accuracy :** 0.9983

**Precision :** 0.9996

**Recall :** 0.9998

**AUC Score:** 0.9027

**Support Vector Machine**

The following figures represents the confusion matrix and ROC curve of the model Support Vector Machine with synthetic dataset generated using Random Oversampling technique.

**True Positive:** There are totally 83,967 transactions which are identified as true positive.

**True Negative:** There are totally 129 transactions which are identified as true negative.

**False Negative:** There are totally 15 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 1332 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart, treemap chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 42: Confusion Matrix and AUC-ROC curve of SVM**

Figure 42 represents ROC curve and AUC score with 0.94. Following are the evaluation metrics for Support Vector Machine with balanced dataset:

**Accuracy :** 0.9867

**Precision :** 0.9998

**Recall :** 0.9877

**AUC Score:** 0.9401

**Artificial Neural Networks**

The following figures represents the confusion matrix and ROC curve of the model Artificial Neural Networks with synthetic dataset generated using Random Oversampling technique.

**True Positive:** There are totally 84,672 transactions which are identified as true positive.

**True Negative:** There are totally 127 transactions which are identified as true negative.

**False Negative:** There are totally 17 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 627 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 43: Confusion Matrix and AUC-ROC curve of ANN**

Figure 43 represents ROC curve and AUC score with 0.94. Following are the evaluation metrics for Artificial Neural Networks with balanced dataset:

**Accuracy :** 0.9980

**Precision :** 0.9997

**Recall :** 0.8333

**AUC Score:** 0.9156

**4)ADASYN**

**Logistic Regression**

The following figures represents the confusion matrix and ROC curve of the model Logistic Regression with synthetic dataset generated using ADASYN technique.

**True Positive:** There are totally 80,872 transactions which are identified as true positive.

**True Negative:** There are totally 133 transactions which are identified as true negative.

**False Negative:** There are totally 11 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 4427 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart, treemap chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 44: Confusion Matrix and AUC-ROC curve of LR**

Figure 44 represents ROC curve and AUC score with 0.93. Following are the evaluation metrics for Logistic Regression with balanced dataset:

**Accuracy :** 0.9585

**Precision :** 0.9998

**Recall :** 0.9585

**AUC Score:** 0.9358

**Random Forest**

The following figures represents the confusion matrix and ROC curve of the model Random Forest with synthetic dataset generated using ADASYN technique.

**True Positive:** There are totally 85,288 transactions which are identified as true positive.

**True Negative:** There are totally 121 transactions which are identified as true negative.

**False Negative:** There are totally 23 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 11 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 45: Confusion Matrix and AUC-ROC curve of RF**

Figure 45 represents ROC curve and AUC score with 0.92. Following are the evaluation metrics for Random Forest Classifier with balanced dataset:

**Accuracy :** 0.9998

**Precision :** 0.9998

**Recall :** 0.9998

**AUC Score:** 0.9235

**Support Vector Machine**

The following figures represents the confusion matrix and ROC curve of the model Support Vector Machine with synthetic dataset generated using ADASYN technique.

**True Positive:** There are totally 81,473 transactions which are identified as true positive.

**True Negative:** There are totally 130 transactions which are identified as true negative.

**False Negative:** There are totally 14 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 3826 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart, treemap chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 46: Confusion Matrix and AUC-ROC curve of SVM**

Figure 46 represents ROC curve and AUC score with 0.93. Following are the evaluation metrics for Support Vector Machine with balanced dataset:

**Accuracy :** 0.9601

**Precision :** 0.9998

**Recall :** 0.9651

**AUC Score:** 0.9289

**Artificial Neural Networks**

The following figures represents the confusion matrix and ROC curve of the model Artificial Neural Networks with synthetic dataset generated using ADASYN technique.

**True Positive:** There are totally 84,058 transactions which are identified as true positive.

**True Negative:** There are totally 130 transactions which are identified as true negative.

**False Negative:** There are totally 14 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 1241 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart, treemap chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 47: Confusion Matrix and AUC-ROC curve of ANN**

Figure 47 represents ROC curve and AUC score with 0.94. Following are the evaluation metrics for ANN with balanced dataset:

**Accuracy :** 0.9853

**Precision :** 0.9997

**Recall :** 0.8333

**AUC Score:** 0.9441

**5) BORDERLINE**

**Logistic Regression**

The following figures represents the confusion matrix and ROC curve of the model Logistic Regression with synthetic dataset generated using BORDERLINE technique.

**True Positive:** There are totally 84,703 transactions which are identified as true positive.

**True Negative:** There are totally 124 transactions which are identified as true negative.

**False Negative:** There are totally 20 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 596 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 48: Confusion Matrix and AUC-ROC curve of LR**

Figure 48 represents ROC curve and AUC score with 0.93. Following are the evaluation metrics for LR with balanced dataset:

**Accuracy :** 0.9931

**Precision :** 0.9998

**Recall :** 0.9901

**AUC Score:** 0.9276

**Random Forest**

The following figures represents the confusion matrix and ROC curve of the model RF with synthetic dataset generated using BORDERLINE technique.

**True Positive:** There are totally 85,292 transactions which are identified as true positive.

**True Negative:** There are totally 120 transactions which are identified as true negative.

**False Negative:** There are totally 24 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 7 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 49: Confusion Matrix and AUC-ROC curve of RF**

Figure 49 represents ROC curve and AUC score with 0.92. Following are the evaluation metrics for Random Forest Classifier with balanced dataset:

**Accuracy :** 0.9983

**Precision :** 0.9996

**Recall :** 0.9998

**AUC Score:** 0.9131

**Support Vector Machine**

The following figures represents the confusion matrix and ROC curve of the model SVM with synthetic dataset generated using BORDERLINE technique.

**True Positive:** There are totally 84,625 transactions which are identified as true positive.

**True Negative:** There are totally 122 transactions which are identified as true negative.

**False Negative:** There are totally 22 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 624 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 50: Confusion Matrix and AUC-ROC curve of SVM**

Figure 50 represents ROC curve and AUC score with 0.92. Following are the evaluation metrics for SVM with balanced dataset:

**Accuracy :** 0.9947

**Precision :** 0.9998

**Recall :** 0.9941

**AUC Score:** 0.9196

**Artificial Neural Networks**

The following figures represents the confusion matrix and ROC curve of the model Artificial Neural Networks with synthetic dataset generated using BORDERLINE technique.

**True Positive:** There are totally 84,924 transactions which are identified as true positive.

**True Negative:** There are totally 124 transactions which are identified as true negative.

**False Negative:** There are totally 20 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 375 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart, treemap chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 51: Confusion Matrix and AUC-ROC curve of ANN**

Figure 51 represents ROC curve and AUC score with 0.93. Following are the evaluation metrics for ANN with balanced dataset:

**Accuracy :** 0.9853

**Precision :** 0.9997

**Recall :** 0.8333

**AUC Score:** 0.9341

**6)SMOTE**

**Logistic Regression**

The following figures represents the confusion matrix and ROC curve of the model LR with synthetic dataset generated using SMOTE technique.

**True Positive:** There are totally 83,557 transactions which are identified as true positive.

**True Negative:** There are totally 130 transactions which are identified as true negative.

**False Negative:** There are totally 14 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 1742 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart, treemap chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 52: Confusion Matrix and AUC-ROC curve of LR**

Figure 52 represents ROC curve and AUC score with 0.94. Following are the evaluation metrics for LR with balanced dataset:

**Accuracy :** 0.9801

**Precision :** 0.9998

**Recall :** 0.9856

**AUC Score:** 0.9411

**Random Forest**

The following figures represents the confusion matrix and ROC curve of the model RF with synthetic dataset generated using SMOTE technique.

**True Positive:** There are totally 85,287 transactions which are identified as true positive.

**True Negative:** There are totally 120 transactions which are identified as true negative.

**False Negative:** There are totally 24transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 12 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 53: Confusion Matrix and AUC-ROC curve of RF**

Figure 53 represents ROC curve and AUC score with 0.92. Following are the evaluation metrics for RF with balanced dataset:

**Accuracy :** 0.9801

**Precision :** 0.9998

**Recall :** 0.9856

**AUC Score:** 0.9211

**Support Vector Machine**

The following figures represents the confusion matrix and ROC curve of the model SVM with synthetic dataset generated using SMOTE technique.

**True Positive:** There are totally 84,282 transactions which are identified as true positive.

**True Negative:** There are totally 127 transactions which are identified as true negative.

**False Negative:** There are totally 17 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 1017 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart, treemap chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 54: Confusion Matrix and AUC-ROC curve of SVM**

Figure 54 represents ROC curve and AUC score with 0.94. Following are the evaluation metrics for SVM with balanced dataset:

**Accuracy :** 0.9967

**Precision :** 0.9998

**Recall :** 0.9931

**AUC Score:** 0.9202

**Artificial Neural Networks**

The following figures represents the confusion matrix and ROC curve of the model Artificial Neural Networks with synthetic dataset generated using ANN technique.

**True Positive:** There are totally 84,487 transactions which are identified as true positive.

**True Negative:** There are totally 128 transactions which are identified as true negative.

**False Negative:** There are totally 16 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 812 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 55: Confusion Matrix and AUC-ROC curve of ANN**

Figure 55 represents ROC curve and AUC score with 0.94. Following are the evaluation metrics for ANN with balanced dataset:

**Accuracy :** 0.9975

**Precision :** 0.9996

**Recall :** 0.8194

**AUC Score:** 0.9086

**7)SVM-SMOTE**

**Logistic Regression**

The following figures represents the confusion matrix and ROC curve of the model LR with synthetic dataset generated using SVM-SMOTE technique.

**True Positive:** There are totally 84,826 transactions which are identified as true positive.

**True Negative:** There are totally 125 transactions which are identified as true negative.

**False Negative:** There are totally 19 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 473 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 56: Confusion Matrix and AUC-ROC curve of LR**

Figure 56 represents ROC curve and AUC score with 0.93. Following are the evaluation metrics for LR with balanced dataset:

**Accuracy :** 0.9947

**Precision :** 0.9998

**Recall :** 0.9931

**AUC Score:** 0.9312

**Random Forest**

The following figures represents the confusion matrix and ROC curve of the model RF with synthetic dataset generated using SVM-SMOTE technique.

**True Positive:** There are totally 85,287 transactions which are identified as true positive.

**True Negative:** There are totally 120 transactions which are identified as true negative.

**False Negative:** There are totally 24 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 12 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 57: Confusion Matrix and AUC-ROC curve of RF**

Figure 57 represents ROC curve and AUC score with 0.92. Following are the evaluation metrics for Random Forest Classifier with balanced dataset:

**Accuracy :** 0.9983

**Precision :** 0.9996

**Recall :** 0.9998

**AUC Score:** 0.9235

**Support Vector Machine**

The following figures represents the confusion matrix and ROC curve of the model SVM with synthetic dataset generated using SVM-SMOTE technique.

**True Positive:** There are totally 84,722 transactions which are identified as true positive.

**True Negative:** There are totally 122 transactions which are identified as true negative.

**False Negative:** There are totally 22 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 577 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 58: Confusion Matrix and AUC-ROC curve of SVM**

Figure 58 represents ROC curve and AUC score with 0.92. Following are the evaluation metrics for SVM with balanced dataset:

**Accuracy :** 0.9967

**Precision :** 0.9998

**Recall :** 0.9931

**AUC Score:** 0.9202

**Artificial Neural Networks**

The following figures represents the confusion matrix and ROC curve of the model Artificial Neural Networks with synthetic dataset generated using SVM-SMOTE technique.

**True Positive:** There are totally 85,191 transactions which are identified as true positive.

**True Negative:** There are totally 124 transactions which are identified as true negative.

**False Negative:** There are totally 20 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 108 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 59: Confusion Matrix and AUC-ROC curve of ANN**

Figure 59 represents ROC curve and AUC score with 0.93. Following are the evaluation metrics for ANN with balanced dataset:

**Accuracy :** 0.9994

**Precision :** 0.9996

**Recall :** 0.8194

**AUC Score:** 0.9095

**8)SMOTE-ENN**

**Logistic Regression**

The following figures represents the confusion matrix and ROC curve of the model LR with synthetic dataset generated using SMOTE-ENN technique.

**True Positive:** There are totally 83,632 transactions which are identified as true positive.

**True Negative:** There are totally 130 transactions which are identified as true negative.

**False Negative:** There are totally 14 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 1667 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart, treemap chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 60: Confusion Matrix and AUC-ROC curve of LR**

Figure 60 represents ROC curve and AUC score with 0.94. Following are the evaluation metrics for LR with balanced dataset:

**Accuracy :** 0.9833

**Precision :** 0.9998

**Recall :** 0.9824

**AUC Score:** 0.9416

**Random Forest**

The following figures represents the confusion matrix and ROC curve of the model RF with synthetic dataset generated using SMOTE-ENN technique.

**True Positive:** There are totally 85,282 transactions which are identified as true positive.

**True Negative:** There are totally 122 transactions which are identified as true negative.

**False Negative:** There are totally 22 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 17 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 61: Confusion Matrix and AUC-ROC curve of RF**

Figure 61 represents ROC curve and AUC score with 0.92. Following are the evaluation metrics for Random Forest Classifier with balanced dataset:

**Accuracy :** 0.9983

**Precision :** 0.9996

**Recall :** 0.9998

**AUC Score:** 0.9269

**Support Vector Machine**

The following figures represents the confusion matrix and ROC curve of the model SVM with synthetic dataset generated using SMOTE-ENN technique.

**True Positive:** There are totally 83,950 transactions which are identified as true positive.

**True Negative:** There are totally 128 transactions which are identified as true negative.

**False Negative:** There are totally 16 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 1349 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 62: Confusion Matrix and AUC-ROC curve of SVM**

Figure 62 represents ROC curve and AUC score with 0.94. Following are the evaluation metrics for SVM with balanced dataset:

**Accuracy :** 0.9801

**Precision :** 0.9996

**Recall :** 0.9841

**AUC Score:** 0.9365

**Artificial Neural Networks**

The following figures represents the confusion matrix and ROC curve of the model Artificial Neural Networks with synthetic dataset generated using SMOTE-ENN technique.

**True Positive:** There are totally 83,395 transactions which are identified as true positive.

**True Negative:** There are totally 130 transactions which are identified as true negative.

**False Negative:** There are totally 14 transactions which are identified as genuine transactions but originally these are fraud transactions.

**False Positive:** There are totally 1904 transactions which are identified as fraud transactions but originally these are genuine transactions.

**Chart, treemap chart

Description automatically generated Chart, line chart

Description automatically generated**

**Figure 63: Confusion Matrix and AUC-ROC curve of ANN**

Figure 63 represents ROC curve and AUC score with 0.94. Following are the evaluation metrics for ANN with balanced dataset:

**Accuracy :** 0.9982

**Precision :** 0.9996

**Recall :** 0.8194

**AUC Score:** 0.9411

### Results Analysis and Comparison

The imbalanced dataset to create various sets of balanced datasets using sampling and SMOTE techniques. These balanced datasets are used to train and test the models. As soon as the training and testing of the models are completed, the performance of the models are evaluated using different performance metrics. The summary of the results and metrics evaluation of all the models is represented in the following Figure 64.

Table

Description automatically generated

**Figure 64: Results Comparison and Analysis of Models**

# Conclusion

In this study, the results shows that the algorithms are not efficient with imbalanced dataset. The models are performing well with better accuracy when data is balanced and SMOTE techniques are utilized to balance this data. Various SMOTE approaches and sampling techniques are used to generate various balanced datasets from the original imbalanced dataset. After careful evaluation of the performance of the models with these balanced datasets are different to that of imbalanced data. AUC is the best metric to compare the performance of the models as the dataset is imbalanced. In this study, it was observed that hybrid sampling approaches such as SMOTE-ENN performs well than other sampling methods. Among all the machine learning algorithms evaluated in this paper, the Support Vector Machine(SVM) models performs well and gives best accuracy results with SMOTE-ENN approach.

As a result of this study, the recommended sampling approaches are hybrid techniques such as SMOTE-ENN to produce best results from the models for imbalanced dataset.

Furthermore, this study can be extended by researching alternative sampling approaches with different machine learning deep learning models. In addition to the metrics evaluation, the time gap to execute a model can be considered as a evaluation measure while comparing the models. These algorithms can be evaluated with different datasets from various sections which consists of imbalanced data issues and to use hybrid approaches to solves these imbalanced data issues.

# References

1. Equifax.co.uk. 2022. *Credit Card Fraud - How To Spot, Prevent And Report It | Equifax UK*. [online] Available at: <https://www.equifax.co.uk/resources/identity-protection/what-is-credit-card-fraud.html> [Accessed 30 March 2022].
2. En.wikipedia.org. 2022. *Credit Card Fraud - Wikipedia*. [online] Available at: <https://en.wikipedia.org/wiki/Credit\_card\_fraud#cite\_note-5> [Accessed 30 March 2022].
3. Fool UK Personal Finance. 2022. *4 Examples Of Credit Card Fraud | Fool UK*. [online] Available at: <https://www.fool.co.uk/personal-finance/credit-cards/learn/4-examples-of-credit-card-fraud/> [Accessed 30 March 2022].
4. Thales Group. 2022. *Slash Card-Not-Present Fraud With Gemalto Dynamic Code Verification*. [online] Available at: <https://www.thalesgroup.com/en/markets/digital-identity-and-security/banking-payment/digital-banking/dcv> [Accessed 30 March 2022].
5. Phishing.org. 2022. *Phishing | Common Phishing Scams*. [online] Available at: <https://www.phishing.org/common-phishing-scams> [Accessed 30 March 2022].
6. Ieeexplore.ieee.org. 2022. *Review On Fraud Detection Methods In Credit Card Transactions*. [online] Available at: <https://ieeexplore.ieee.org/abstract/document/8321781> [Accessed 30 March 2022].
7. Reader.elsevier.com. 2022. [online] Available at: <https://reader.elsevier.com/reader/sd/pii/S1877050915007103?token=E94996BC04A054042FE1F2787BCA9BA9122659BEAE941817874058400D8A44543C1ABA204E587B7C9AE2FB36A74FAE2A&originRegion=eu-west-1&originCreation=20220406121555> [Accessed 6 April 2022].
8. 2022. [online] Available at: <https://www.researchgate.net/figure/Proposed-data-driven-approaches-for-credit-card-fraud-detection\_fig1\_323858923> [Accessed 6 April 2022].
9. Media.springernature.com. 2022. [online] Available at: <https://media.springernature.com/original/springer-static/image/chp%3A10.1007%2F978-981-15-8752-8\_4/MediaObjects/489176\_1\_En\_4\_Fig1\_HTML.png> [Accessed 6 April 2022].
10. Tokenex.com. 2022. *5 Effective Ways To Prevent Credit Card Fraud In 2021*. [online] Available at: <https://www.tokenex.com/blog/five-ways-to-prevent-credit-card-fraud> [Accessed 6 April 2022].
11. Tokenex.com. 2022. [online] Available at: <https://www.tokenex.com/hs-fs/hubfs/19-09-05-X%20Ways%20to%20prevent%20Credit%20Card%20Fraud%20in%202019-d01-01.jpg?width=663&name=19-09-05-X%20Ways%20to%20prevent%20Credit%20Card%20Fraud%20in%202019-d01-01.jpg> [Accessed 6 April 2022].
12. Ieeexplore.ieee.org. 2022. *Credit Card Fraud Detection Using Machine Learning Techniques: A Comparative Analysis*. [online] Available at: <https://ieeexplore.ieee.org/abstract/document/8123782> [Accessed 14 April 2022].
13. Ghosh and Reilly, "Credit card fraud detection with a neural-network," 1994 Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences, 1994, pp. 621-630, doi: 10.1109/HICSS.1994.323314.
14. J. V. V. Sriram Sasank, G. R. Sahith, K. Abhinav and M. Belwal, "Credit Card Fraud Detection Using Various Classification and Sampling Techniques: A Comparative Study," 2019 International Conference on Communication and Electronics Systems (ICCES), 2019, pp. 1713-1718, doi: 10.1109/ICCES45898.2019.9002289.
15. A. Syaripudin and M. L. Khodra, "A comparison for handling imbalanced datasets," 2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA), 2014, pp. 293-298, doi: 10.1109/ICAICTA.2014.7005957.
16. E. I. Altman, G. Marco and F. Varetto, "Corporate distress diagnosis comparisons using linear discriminant analysis and neural networks", Journal of Banking and Finance, vol. 18, no. 3, pp. 505-529, 1994.
17. A.M. Flitman, "Towards analysing student failures: neural networks compared with regression analysis and multiple discriminant analysis", Computers & Operations Research, vol. 24, no. 4, pp. 367-377, 1997.
18. J.A. Ohlson, "Financial ratios and probabilistic prediction of bankruptcy", Journal of Accounting Research, vol. 18, no. 1, pp. 109-131, 1980.
19. S. Le Cessie and J. C. Van Houwelingen, “Ridge estimators in logistic regression,” J. R. Stat. Soc. Ser. C (Applied Stat., vol. 41, no. 1, pp. 191–201, Mar. 1992. https://doi.org/10.2307/2347628
20. D. H. Pandya, S. H. Upadhyay, and S. P. Harsha, “Fault diagnosis of rolling element bearing by using multinomial logistic regression and wavelet packet transform,” Soft Comput., vol. 18, no. 2, pp. 255–266, Feb. 2014. https://doi.org/10.1007/s00500-013-1055-1
21. Hussein, A.S., Khairy, R.S., Najeeb, S.M.M. and ALRikabi, H.T., 2021. Credit Card Fraud Detection Using Fuzzy Rough Nearest Neighbor and Sequential Minimal Optimization with Logistic Regression. International Journal of Interactive Mobile Technologies, 15(5).
22. Dheepa, V. and Dhanapal, R., 2012. Behavior based credit card fraud detection using support vector machines. ICTACT Journal on Soft computing, 2(4), pp.391-397.
23. M. Syeda, Y. Zhang and Y. Pan, "Parallel granular neural networks for fast credit card fraud detection", Proceedings of the 2002 IEEE International Conference on Fuzzy Systems, 2002.
24. D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic and A. Anderla, "Credit Card Fraud Detection - Machine Learning methods," 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH), 2019, pp. 1-5, doi: 10.1109/INFOTEH.2019.8717766.
25. D. S. Sisodia, N. K. Reddy and S. Bhandari, "Performance evaluation of class balancing techniques for credit card fraud detection," 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), 2017, pp. 2747-2752, doi: 10.1109/ICPCSI.2017.8392219.
26. miro.medium.com. 2022. [online] Available at: <https://miro.medium.com/max/1400/1\*dm6ZaX5fuSmuVvM4Ds-vcg.jpeg> [Accessed 6 May 2022].
27. Editor.analyticsvidhya.com. 2022. [online] Available at: <https://editor.analyticsvidhya.com/uploads/30469resampling.png> [Accessed 6 May 2022].
28. Dataaspirant.com. 2022. [online] Available at: <https://dataaspirant.com/wp-content/uploads/2020/08/10-oversampling.png> [Accessed 6 May 2022].
29. Miro.medium.com. 2022. [online] Available at: <https://miro.medium.com/max/700/1\*iXHQaRrdIJLRsxxjEnr67Q.jpeg> [Accessed 6 May 2022].
30. Miro.medium.com. 2022. [online] Available at: <https://miro.medium.com/max/700/1\*p2Ugh8xOQjjwunXmRtN9Cg.jpeg> [Accessed 6 May 2022].
31. Miro.medium.com. 2022. [online] Available at: <https://miro.medium.com/max/700/1\*CG712DHQL\_ZMS3gAEGd\_5w.jpeg> [Accessed 6 May 2022].
32. Miro.medium.com. 2022. [online] Available at: <https://miro.medium.com/max/700/1\*0jwntVGaj7qQkr-MeueQcQ.jpeg> [Accessed 6 May 2022].
33. Medium. 2022. *SMOTE: Synthetic Data Augmentation For Tabular Data*. [online] Available at: <https://towardsdatascience.com/smote-synthetic-data-augmentation-for-tabular-data-1ce28090debc> [Accessed 6 May 2022].
34. Imbalanced-learn.org. 2022. *Sensitivity\_Specificity\_Support — Version 0.10.0.Dev0*. [online] Available at: <https://imbalanced-learn.org/dev/references/generated/imblearn.metrics.sensitivity\_specificity\_support.html> [Accessed 6 May 2022].
35. freeCodeCamp.org. 2022. *Random Forest Classifier Tutorial: How To Use Tree-Based Algorithms For Machine Learning*. [online] Available at: <https://www.freecodecamp.org/news/how-to-use-the-tree-based-algorithm-for-machine-learning/> [Accessed 6 May 2022].
36. Salah Alaloul, W. and Hannan Qureshi, A., 2022. *Data Processing Using Artificial Neural Networks*.
37. 2022. [online] Available at: <https://www.intechopen.com/media/chapter/71673/media/F6.png> [Accessed 6 May 2022].
38. Researchgate.net. 2022. [online] Available at: <https://www.researchgate.net/profile/Xiu-Jin/publication/345482304/figure/fig1/AS:955461788524544@1604811250763/The-process-of-SMOTE-ENN-algorithm-A-SMOTE-selected-each-sample-from-the-minority.png> [Accessed 6 May 2022].
39. 2022. [online] Available at: <https://www.intechopen.com/media/chapter/71673/media/F7.png> [Accessed 6 May 2022].
40. Researchgate.net. 2022. [online] Available at: <https://www.researchgate.net/profile/Elizabeth-Thomas-36/publication/343326638/figure/fig3/AS:919284234592257@1596185849753/llustration-of-random-undersampling-technique.jpg> [Accessed 6 May 2022].
41. Brownlee, J., 2022. *Random Oversampling And Undersampling For Imbalanced Classification*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/> [Accessed 6 May 2022].
42. Miro.medium.com. 2022. [online] Available at: <https://miro.medium.com/max/1202/1\*dOHFA1ffNyKbQ55Vw0PNzA.png> [Accessed 6 May 2022].
43. Arxiv.org. 2022. [online] Available at: <https://arxiv.org/pdf/2105.04301> [Accessed 6 May 2022].
44. Medium. 2022. *5 SMOTE Techniques For Oversampling Your Imbalance Data*. [online] Available at: <https://towardsdatascience.com/5-smote-techniques-for-oversampling-your-imbalance-data-b8155bdbe2b5> [Accessed 6 May 2022].
45. Medium. 2022. *SMOTE: Synthetic Data Augmentation for Tabular Data*. [online] Available at: <https://towardsdatascience.com/smote-synthetic-data-augmentation-for-tabular-data-1ce28090debc> [Accessed 6 May 2022].
46. Imbalanced-learn.org. 2022. [online] Available at: <https://imbalanced-learn.org/stable/\_images/sphx\_glr\_plot\_comparison\_over\_sampling\_thumb.png> [Accessed 6 May 2022].
47. Ieeexplore.ieee.org. 2022. *Credit Card Fraud Detection - Machine Learning methods*. [online] Available at: <https://ieeexplore.ieee.org/document/8717766> [Accessed 6 May 2022].