National College of Ireland

Credit Card Fraud Detection Using Conditional Tabular Generative Adversarial Networks (CT-GAN) and Supervised Machine Learning Techniques.

MSc Research Project
Data Analytics

Tushar Patil
Student ID: X19199988

School of Computing
National College of Ireland

Supervisor: Dr. Bharathi Chakravarthi

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Tushar Patil
Student ID:	X19199988
Programme:	Data Analytics
Year:	2020-2021
Module:	MSc Research Project
Supervisor:	Dr. Bharathi Chakravarthi
Submission Due Date:	16/08/2021
Project Title:	Credit Card Fraud Detection Using Conditional Tabular Generative Adversarial Networks (CT-GAN) and Supervised Machine Learning Techniques.
Word Count:	7550
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Tushar Patil
Date:	28th September 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Credit Card Fraud Detection Using Conditional Tabular Generative Adversarial Networks (CT-GAN) and Supervised Machine Learning Techniques.

Tushar Patil
X19199988

Abstract

Credit card fraud has been a major concern for financial institutes and business stakeholders for a long time and due to its ever-increasing nature, it has been a global topic of interest for researchers. Machine learning has proved to be one of the promising approaches for the detection and prediction of frauds. Despite having various advantages, there is no ideal model to handle this task due to the various factors involved. On the other hand, the class imbalance is one of the major and frequently occurring challenges while dealing with fraud detection tasks which hamper the model performance. There are several previously explored studies combining machine learning algorithms with various data-pre-processing techniques to handle class imbalance challenges. To take this research further we have used a novel approach of combining supervised machine learning algorithms like Logistic regression, Random Forest, XGBoost with Conditional Tabular Generative Adversarial Networks (CT-GAN) for balancing skewed data by data augmentation. We have used the SelectKBest feature selection method for selecting the most significant feature for our analysis. After testing the proposed technique on our machine learning algorithms which are trained on both unbalanced and balanced data, we have observed a significant increase in model performances in terms of F1-score, recall, AUC score and Gmean. The results show that the Random Forest model outperforms other models in all terms with 100% recall value followed by XGBoost having recall of 91% after applying our proposed technique whereas Logistic Regression has shown the most significant increase in performance from 78% recall to 90% after trained on balanced data.

1 Introduction

The astonishing growth of the e-commerce sector in the past few years have made cashless payments using credit cards an important aspect of the financial industry. With the increased usage of credit cards, they have become a hot target for cyberattacks and financial frauds throughout the industry. Credit card fraud happens when an unauthorised person uses a credit card for purchase without the authorization of the credit card owner. Credit card frauds account for billion-dollar losses throughout the world. Total financial loss due to such frauds was as high as 22.8 billion USD in 2017 and was expected to reach a count of 31 billion USD by 2020 (Oblé and Bontempi; 2019; John and Naaz; 2019). As detection of small fraudulent transactions would lead to avoiding large financial losses,

efficient credit card fraud detection systems are important for researchers and financial organizations. Though fraud detection methods were there for quite a long time, a continuous evolution of distribution over time because of new attacks and seasonality and a small percentage of fraud transactions makes them a challenging and ever-mutating problem for researchers.

Class imbalance (a rare percentage of fraud transactions in data) is a major challenge faced in the fraud detection task, as it dominates the performance of machine learning techniques. Various techniques are proposed to tackle this class imbalance challenge in fraud detection tasks. Roughly, we can categorize the techniques as following: Data-driven approaches and algorithm-driven approaches. In this research, we are making use of the data-driven methodology for handling the class imbalance challenge. It includes resampling methods like Condensed Nearest Neighbour Rule (CNN), Wilson Edited Nearest Neighbour Rule (ENN), random oversampling, focused oversampling, and synthetic oversampling Etc. All these methods incorporate some shortcomings like uncontrolled loss of information, increased training time, increased probability of overfitting of the models Etc.(Thabtah et al.; 2020a)

In this research, we are proposing a novel approach to tackle the class imbalance challenge by making combined use of Conditional generative adversarial networks(CT-GAN) and machine learning techniques to credit card fraud detection tasks. In this proposed method we are employing well-evolved machine learning algorithms like logistic regression, random forest, SVM, etc. for fraud classification tasks as they perform efficiently in fraud detection tasks (Bhattacharyya et al.; 2011). On the other hand, these machine learning models are highly prone to overfitting due to class imbalance problems (Bhattacharyya et al.; 2011). The proposed approach is majorly concerned with handling the class imbalance challenge with data augmentation using an advanced version of GAN which is CT-GAN. A generative adversarial network(GAN) consists of two neural networks working together for the generation of synthetic data samples from simple noise. The CT-GAN is a modified version of GAN which is specially designed for tabular data augmentation. We would be utilizing the same for the generation of minority class data samples. Unlike samples generated by the random oversampling method, samples augmented by CT-GAN are more random and genuine because of the ability of neural networks to work like the human mind and uncover the hidden patterns present in the data.(Fiore et al.; 2019; Asha and KR; 2021)

Hence, the research question for this study would be- "Can we improve the credit card fraud detection accuracy by using the novel combination of CT-GAN and machine learning techniques as compared to other state-of-art techniques available to handle class imbalance challenges?". The major contribution of this research is a novel approach of using an open-source python library(CTGAN) for tabular data augmentation, which removes the complexity of developing and training the GAN. Due to this, we are proposing a more simple, lightweight, and scalable credit card fraud detection architecture. The performance of the proposed intelligent method is tested on publicly available data and evaluated by using performance evaluation metrics.

2 Related Work

Credit card frauds are the major type of fraud, causing some serious financial loss to financial institutes every year. Due to the severity of this issue, there has been some

research and practices for the detection and prediction of fraudulent transactions from legitimate ones for a long time. The evolving nature of frauds and cyberattacks gives rise to the need for more accurate and advanced methodologies for the detection and prevention of credit card fraud. Machine learning has been one of the most successful techniques for fraud detection. Also, the class imbalance is a crucial challenge while dealing with frauds detection problems using machine learning and needs to be addressed. Several studies and research have been done in the same field and provided important contributions to the field. We would review these various researches in this section.

2.1 Existing Techniques for credit card Fraud Detection

(Khatri et al.; 2020) has carried out an experimental case study using publicly available data for credit card fraud classification task. Here they have utilized supervised machine learning models to decision tree, K- nearest neighbour[KNN], logistic regression, random forest, naïve Bayes and evaluated and compared their performances based on sensitivity, precision, and time for selection of best performing model for a given set of data. They have identified the decision tree method outperforming others in terms of overall performance and time for processing the data when having the threshold changed to 0.4. (Khatri et al.; 2020) have obtained a slightly lower sensitivity of 79.21 for the decision tree model as compared to KNN having a sensitivity of 81.19. On the other hand, decision tree methods take as little time as 0sec when compared to KNN which took almost 462 sec. As the time taken for detection of fraudulent transactions is a crucial factor, they have suggested a decision tree method more suitable for real-time applications. Here, authors have highlighted the class imbalance as a major challenge in the field of fraud detection. A similar critical review has been carried out by (Al Smadi and Min; 2020) exploring potential machine learning techniques for fraud detection. In this study, authors have highlighted the critical nature of credit card fraud and identified that it contributes highest among various types of fraud, which is 29%. Here authors have analysed the performance of various fraud detection approaches using deep learning, machine learning algorithms, HMM, fuzzy logic-based systems, etc. and provided critical comparative analysis on the same. (Al Smadi and Min; 2020) have analysed the trade-off between cost, computational time, efficiency and the data handling capability of various fraud detection techniques. Neural networks provide high accuracy and can work with large data but lack simplicity and requires high computational resources which make a much more costly approach. Fuzzy logic-based systems and HMM comes with the same disadvantages, as they are more complex and costly to implement. In contrast to the above-mentioned techniques' machine learning algorithms like logistic regression, decision tree, random forest, etc. are easy to implement, efficient and cost-efficient which makes them more scalable and robust. (Al Smadi and Min; 2020) have also identified class imbalance as the major issue hampering the performance of machine learning techniques while dealing with fraud detection problems.

In the context of the above-discussed research work, (Sadineni; 2020) carried out a similar experiment by considering SVM as another potential algorithm in addition to previously used methods for the fraud detection task. Here the authors have done a comparative analysis using accuracy, precision, and false alarm rate as evaluation metrics. (Sadineni; 2020) have discussed the unsuitability of SVM due to large time consumption, despite having advantages like reduced risk of overfitting and suitability to handle high dimensionality data. It also performs poorly in terms of accuracy, having an accuracy of

95.16% and gets surpassed by ANN which having an accuracy of 99.92%. It also lacks in terms of false rate alarm(4.9%) which is much higher with respect to ANN(0.1%). In this research, authors have concluded the superiority of ANN over other algorithms, but also highlighted the high computational requirements for the fraud detection task. In contradiction of evaluations, metrics used by (Sadineni; 2020), (Bhattacharyya et al.; 2011) have highlighted the unsuitability of accuracy for evaluating the model performances due to the presence of high data imbalance in the real-world transaction data. Due to this class imbalance, the majority class which is legitimate transactions in the real world would dominate the performance of the machine learning model in terms of accuracy. Here authors have used F1-score-(2*precision*recall/[precision+recall]- Harmonic mean of precision and recall), Geometric mean (2*precision*recall/[precision +recall]- Harmonic mean of precision and recall) in addition to precision and recall for evaluating the performance of machine learning algorithms. According to the experiments carried out by (Bhattacharyya et al.; 2011), the Random Forest classifier outperforms logistic regression and SVM in terms of overall performance using above-mentioned evaluation metrics.

(Lucas et al.; 2020) has proposed another approach for credit card fraud detection using a feature engineering strategy. Apart from relying just on raw transactional data for fraud detection, authors have identified historical transactional patter as a crucial feature for the detection of frauds. With help of the Hidden Markov Model(HMM), based features (Lucas et al.; 2020) have obtained a hidden sequential feature from historical transactions and used the same for the application of machine learning models like Random Forest. In this research, the authors have highlighted the advantages of the proposed approach like robustness for different classifiers and hyperparameter change. They have recorded a significant increase in precision-recall AUC using this technique when used with the Random Forest. According to the authors, the major limitation of this approach is not being able to yield good results when applied to short transaction history and handling missing values. (Asha and KR; 2021) has carried out an experimental study using a deep learning approach. Here, the authors have focused their analysis on the supervised approach using ANN. By using an ANN having 15 hidden layers with RELU as an activation function, they have achieved some outstanding results for the fraud detection task. They have evaluated the performance of various models using evaluation metrics using precision, recall, accuracy and confusion matrix and concluded ANN being superior to all other models like SVM, KNN in overall performance. To handle the class imbalance issue, (Asha and KR; 2021) have implemented normalization and under-sampling methods in a combination of ANN and achieved accuracy around 100%. (Wang et al.; 2019) have proposed a novel deep learning approach using a semi-supervised graph embedding neural network (SemiGNN) for the fraud detection task. SemiGNN involves the use of both labelled and unlabelled data for modelling multi-view graphs. These graphs highlight the hidden correlations between the financial transactions. Here, the authors have used social relations of both labelled and unlabelled data to form a multi-view map network. (Wang et al.; 2019) have highlighted the ability of their proposed technique to identify the important feature for a specific task and also obtained higher performance compared to baseline models like XGBoost.

2.2 Class Imbalance: Challenges and Solutions

As highlighted in the above-discussed research work, class imbalance Is one of the crucial challenges while dealing with credit card fraud detection tasks. To handle this challenge, (Brennan; 2012) have proposed two different approaches: data-oriented and algorithm-based. For the algorithm bases approach, the authors have used Naïve Bayes, ID3, KNN, RF algorithms for various samples taken from 3 different datasets. In this approach, they have identified the best classifiers in basis of misclassification cost using probability threshold. The data-oriented approach involved the resampling techniques like oversampling, undersampling and SMOTE. As per (Brennan; 2012), the data-centric approach using oversampling methods yields the best results, whereas the undersampling technique performs poorly. They have identified Random Forest as the best classifier while implementing an algorithm-based approach using F1-score as an evaluation metric. (Dal Pozzolo et al.; 2014) have proposed another approach to handle the class imbalance challenge, using the data mining technique. This proposed method creates a new model each time on the arrival of new data in the system. Here the author has done a critical comparative study between techniques like RF, SVM, NNET, sampling methods like SMOTE, undersampling, etc. and modelling techniques using real-world data. Here, the author has highlighted the importance of updating a model in non-stationary environments to get better results. Here (Dal Pozzolo et al.; 2014) have identified RF as a better performing model than others.

(Thabtah et al.; 2020b) has carried out a critical analysis of the class imbalance issue in the fraud detection domain and discussed different techniques to handle the same. Here the author has studied techniques like undersampling, oversampling, cost-sensitive learning, thresholding methods, SMOTE, etc. and done a comparative analysis. In addition to this, (Thabtah et al.; 2020b) have also tried to identify the effect of degree of skewness on the given classifier. They have carried out this experimental study on the Naïve Bayes algorithm with different degrees of skewness and evaluated the results. In another study, (Zhu et al.; 2020) have proposed another innovative approach of using Weighted extreme learning machines(WELM) to handle data skewness issues. WELM consists of an enhanced neural network having a single hidden layer forward neural network, which makes them quicker than ANN. By assigning different weights to various samples, WELM is quite efficient while working with imbalanced data. Here, the authors have obtained a significant increase in performance over other algorithms using this approach.

Class imbalance is a common challenge occurring while dealing with classification tasks using machine learning, and is not limited to credit card fraud detection only. (Le et al.; 2019) studied the class imbalance challenge associated with bankruptcy prediction task. Here, author have proposed two methodologies for handling the class imbalance challenge. Here, the author has developed a hybrid approach of sensitive learning and oversampling techniques. Initially, oversampling is done over validation set using optimal balancing ratio to get ideal performance. On the other hand, the cost-sensitive learning model-CBoost is implemented for bankruptcy prediction. The data used for the study has a high imbalance ratio of 0.0026. Here the author has highlighted the possibility of model overfitting while using the oversampling technique as it generates copies of minority class for balancing the data. On the other hand, it also increases the data size, which directly affects the processing time and memory requirements for holding the data. SMOTEENN technique has been used for oversampling, followed by the CBoost model for clustering. AUC and Geometric mean are used for the evaluation of various classifiers like Random

Forest, AdaBoost, Bagging etc.

2.3 Generative Adversarial Networks(GAN)

(Ngwenduna and Mbuvha; 2021) have discussed various aspects associated with GAN. In this piece of research, authors have claimed GAN as a more suitable and efficient approach to handle class imbalance challenge as compared to other sampling techniques. According to the authors, GAN is more robust towards over-fitting and over-lapping due to its flexibility and property of understanding hidden structures of data using deep networks. (Ngwenduna and Mbuvha; 2021) have highlighted the efficiency of GAN using various aspects like application areas, architectural design, different variants present to deal with specific properties, challenges accompanying GAN, Etc. They have also discussed the previously done experimental studies and methodologies used for the evaluation of GAN using different metrics. Taking the research study further (Ngwenduna and Mbuvha; 2021) have also done a comparative analysis between GAN and other resampling methods SMOTE and claimed GAN to be more efficient in overall performance. Here the author has also highlighted the challenges like mode collapse, vanishing gradient, training instability associated with GAN and claimed variants of GAN: WGAN and WGAN-GP as more suitable for practical applications to overcome these challenges. Contributing further to the research work, (Creswell et al.; 2018) have reviewed the numerous aspects of GAN. Here authors have studied different GAN architectures like fully connected GAN, convolutional GAN, conditional GAN and discussed advantages and disadvantages associated with each of them.

Unlike the above-reviewed research works, (Gui et al.; 2020) have discussed the GAN in a more mathematical and theoretical approach. Here, the authors have provided deep insights into the training complexities associated with different variants of GAN and done a comparative analysis based on their applications. (Gui et al.; 2020) have proposed 3 different points of view for tackling the challenges associated with the training of GAN; skills, the structure of GAN and the objective of the application. In this study, authors have claimed model score, Fréchet inception distance, inception score and multi-scale structural similarity as suitable metrics for evaluating the performance GAN. To understand the applications of GAN for the banking domain, (Pandey et al.; n.d.) have discussed the suitability and limitations of GAN for banking problems. In this research work, the authors have utilized the WGAN-GP variant of GAN for data augmentation and used the augmented data for the application of machine learning techniques. Here (Pandey et al.; n.d.) have recorded a significant increase of 5% in recall value of XGBoost classifier after trained on augmented data compared to a model trained on original data but observed decrease in precision and F1-score values.

To highlight the applications of GAN to handle class imbalance challenge in fraud detection tasks, (Fiore et al.; 2019) have done an experimental study by making use of GAN for the generation of synthetic samples for minority class in transaction data for balancing the skewed dataset. To back up the effectiveness of the proposed approach, (Fiore et al.; 2019) have compared the proposed methodology with SMOTE technique. In this research, authors have identified the increased rate of false-positive errors as a major limitation of the proposed approach and suggested ensemble methods as a potential solution for the same. According to (Fiore et al.; 2019), the proposed methodology is unsuitable for an unsupervised approach. In another similar study done by (Yilmaz et al.; 2020) have made use of GAN for balancing the data in the cyberattack dataset. Here, the

authors have augmented samples for UGR'16 data and used the same for classification tasks. Here (Yilmaz et al.; 2020) have a significant surge in overall performance when evaluated using accuracy, precision, recall, F1-score as evaluation metrics.

3 Methodology

CRISP-DM (Cross Industry Process for Data Mining) and KDD (Knowledge Discovery in Databases) are the two most popular approaches being used for data mining projects in the industry. SEEMA is another modelling approach that is majorly used for SAS enterprise miners. Unlike KDD modelling technique which involves nine steps, CRISP-DM provides a complete modelling structure in six steps. CRISP-DM has strong advantages like generalizable architecture, adaptability, flexibility which makes it much more suitable for many data science tasks [1]. Due to these advantages, we are adapting CRISP-DM as our research methodology. The steps involved in our methodology are discussed in the following sections. Figure -1 depicts the architectural design and steps for CRISP-DM methodology.

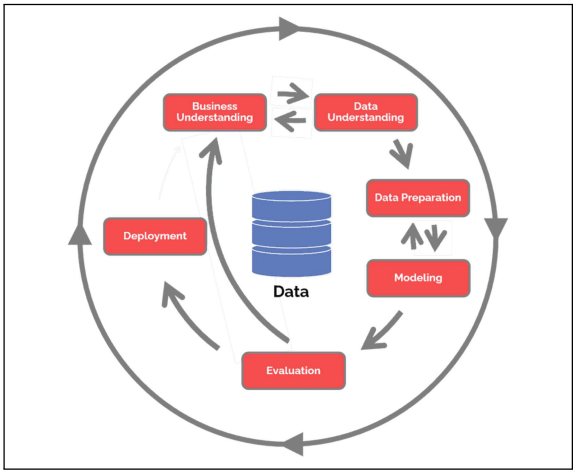


Figure 1: CRISP-DM Methodology, Source:1

3.1 Business Understanding

Understanding the objectives of the project is the first step towards building a successful plan for any project. Our prime goal in this research work is to classify fraudulent credit card transactions from legitimate ones by using supervised machine learning techniques. Thus, the main objective of this research is to build a credit card fraud detection model which would be lightweight, simple, robust, and more immune to class imbalance problems. This model would be useful for banks, credit card providers and other financial institutes for fast, scalable yet reliable credit card fraud detection which would be able to prevent large financial losses occurring by fraud transactions. Class imbalance is a major issue that puts limitations on the performance of machine learning techniques. We are proposing a novel approach for handling the class imbalance challenge by using CT-GAN

[1] https://www.datascience-pm.com/crisp-dm-2/

for data augmentation for balancing our data. This research would provide a significant contribution in the field of credit card fraud detection and would help the stakeholders to reduce the number of frauds.

3.2 Data Understanding

The next step in CRISP-DM methodology is understanding the selected data. Having a proper understanding of the data is critical for building an efficient and reliable credit card fraud detection model. Getting the required data ethically from a reliable source itself is a big challenge. Also, due to high confidentiality concerns regarding sensitive information, getting transaction data for the credit card is a challenging task.

For our research, we are using open-source transaction data for European credit cardholders recorded in 2013. Due to confidentiality issues, the data providers have masked the sensitive information in the form of transformed PCA components. The data contains one categorical feature and 30 other continuous features, including 'Time' and 'Amount' of the given transaction. The dataset can be found at this URL. [2]

3.2.1 Data Exploration

Data exploration is a crucial step towards getting insights into the selected data. Visualization is one of the efficient ways to explore and understand the structures and patterns in data. The quality exploration of data makes it easier to use and navigate through it at the later stage of the project. In our research, we are using various visualizations to explore the data, which are discussed in this section.

- **Data types of all features: Fig- 2**

  In our dataset, the majority of features are in float type. Only our target variable 'Class' is in integer format. We don't have to convert any feature for further processing.

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

Figure 2: Data types of features

[2] https://www.kaggle.com/mlg-ulb/creditcardfraud

- **Analysing missing and duplicate values: Fig- 3**

  In our dataset, there are no missing values present. On the other hand, we have very small percentage of duplicate samples present in the data, which we could remove in data cleaning step.

Dataset Statistics	
Number of Variables	31
Number of Rows	284807
Missing Cells	0
Missing Cells (%)	0.0%
Duplicate Rows	1081
Duplicate Rows (%)	0.4%
Total Size in Memory	67.4 MB
Average Row Size in Memory	248.0 B
Variable Types	Numerical: 30 Categorical: 1

Figure 3: Duplicates and missing values in data

- **Investigating target class distribution: Fig- 4**

  By looking at the distribution of our target feature 'Class' we can see there is large class imbalance present in our data. As shown in the 4 only 0.17% of samples are recorded as 'fraud'. We would be handling this class imbalance in further stages to avoid any negative effects on our machine learning models.

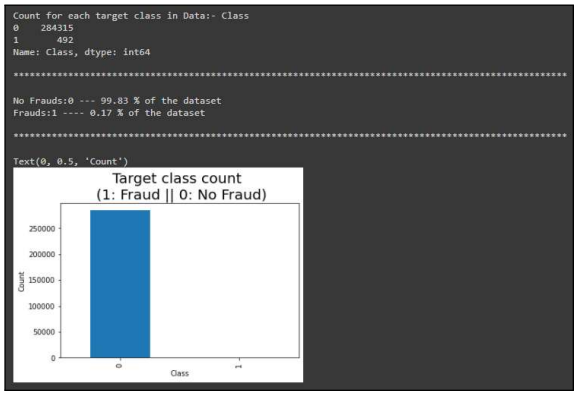
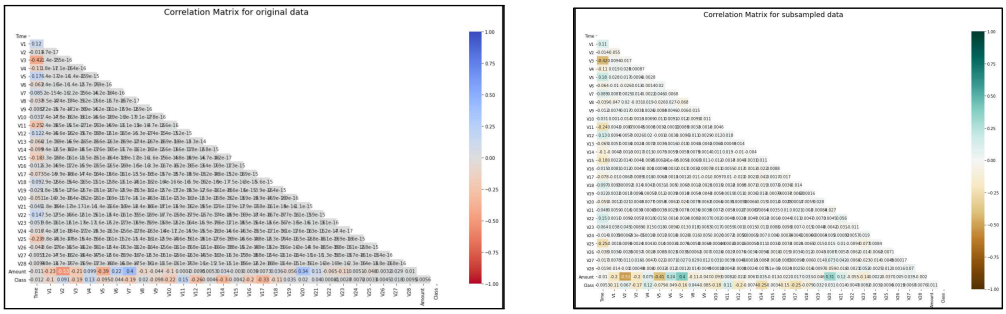


Figure 4: Target feature distribution

- **Correlation matrix: Fig- 5**

  Despite having the significant number of features in our data, not all the features are contributing towards our target feature. Fig- 5 depicts the correlations present between features of data. As the dataset contains large number of samples, to get better understanding of present correlations, we have used correlation matrix for sub-sample of data. With the help of correlation matrix, we have shortlisted significant features towards prediction of our target feature.



(a) Original Data (b) Sub-sampled data

Figure 5: Correlation Matrix

3.3 Data Preparation

To achieve accurate high-quality results out of the machine learning model, the data used for its training and testing should be well-prepared. Data preparation is one of the most critical steps in the process of data mining. There are several aspects like handling missing data, dealing with duplicate values, removing redundant features from data using correlation matrix and feature selection methods, dealing with the unbalanced nature of data, etc. which are involved in the data preparation stage in CRISP-DM [3]. The quality of machine learning performance is highly correlated with the quality of data preparation techniques followed. The low-quality data preparation approach could result in high computational time and cost, poor results of models. Due to all these factors, data preparation is the most challenging and time-consuming stage in the data mining process. In our proposed research, we have applied the following data preparation steps.

- **Handling missing values and duplicates from data:**

  As shown in Fig:- 3 there is a small percentage of duplicates present in our data which we have dropped for our further analysis. On the other hand, we have no missing values present in our data, hence we don't have to take any action on that front.

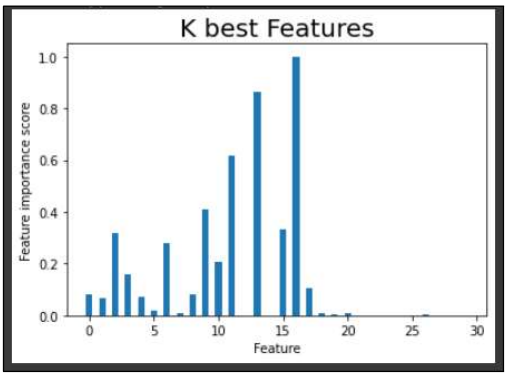
- **Scaling of features in data:**

  By looking at our data, we have identified the scale difference between all other features except Time and Amount. To avoid any negative impact of this scaling gap between features on our model performance, we have scaled the Time and Amount of features from the data using the 'RobustScalar' method.

3.3.1 Feature Selection

After cleaning the data and analysing correlations between the features, feature selection is one of the techniques which further improves the performance of models. This technique is used to get rid of redundant variables, resulting in reduced feature space, which may help to elevate the model's overall performance (Liu and Motoda; 1998). There are several feature selection presents in the field. For our research, we have used the univariate feature selection technique SelectKBest(SKB) which calculates the feature importance

[3] https://machinelearningmastery.com/data-preparation-techniques-for-machine-learning/



(a) Feature scores using SelectKBest method

	V1	V2	V3	V4	V5	V7	V9	V10	V11	V12	V14	V16	V17	V18	Class
0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.239599	0.363787	0.090794	-0.551600	-0.617801	-0.311169	-0.470401	0.207971	0.025791	0
1	1.191857	0.266151	0.166480	0.448154	0.060018	-0.078803	-0.255425	-0.166974	1.612727	1.065235	-0.143772	0.463917	-0.114805	-0.183361	0
2	-1.358354	-1.340163	1.773209	0.379780	-0.503198	0.791461	-1.514654	0.207643	0.624501	0.066084	-0.165946	-2.890083	1.109969	0.121359	0
3	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	0.237609	-1.387024	-0.054852	-0.226487	0.178228	-0.287924	-1.059647	-0.684093	1.965775	0
4	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.592941	0.817739	0.753074	-0.822843	0.538196	-1.119670	-0.451449	-0.237033	-0.038195	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
284802	-11.881118	10.071785	-9.834783	-2.066656	-5.364473	-4.918215	1.914428	4.356170	-1.593105	-2.711941	4.626942	1.107641	1.991691	0.510632	0
284803	-0.732789	-0.055080	2.035030	-0.738589	0.868229	0.024330	0.584800	-0.975926	-0.150189	0.915802	-0.675143	-0.711757	-0.025693	-1.221179	0
284804	1.919565	-0.301254	-3.249640	-0.557828	2.630515	-0.296827	0.432454	-0.484782	0.411614	0.063119	-0.510602	0.140716	0.313502	0.395652	0
284805	-0.240440	0.530483	0.702510	0.689799	-0.377961	-0.686180	0.392087	-0.399126	-1.933849	-0.962886	0.449624	-0.608577	0.509928	1.113981	0
284806	-0.533413	-0.189733	0.703337	-0.506271	-0.012546	1.577006	0.406180	-0.915427	-1.040458	-0.031513	-0.084316	-0.302620	-0.660377	0.167430	0

283726 rows × 15 columns

(b) Reduced data after feature selection

Figure 6: Feature selection:- SelectKBest technique

scores using ANOVA F-values. [4] The selected features are based on various statistical test scores. Here, we can select the top K number of features by fixing the threshold for filtering.(Zulfiker et al.; 2021) We are using the top 14 features having scores above 0.

Figure:-6 shows the scores calculated by the SelectKBest technique for all the features in our data.[5] The scores calculated by this technique shows that there are many redundant variables present in our data that can be dropped. Also, it supports our preliminary selection of predictors using a correlation matrix. After removing all other features, our dataset looks like Figure:6

3.3.2 Handling Class Imbalance

As discussed in the above sections, the class imbalance is one of the major challenges in the domain of fraud detection. Machine learning algorithms are designed to perform best when trained to adequate samples of both classes. The rare nature of fraud transactions within the overall data makes the performance of machine learning models prone to biased results and overfitting. This may result in misclassification for the less represented class samples. To overcome this challenge, there exists several sampling techniques which offer various advantages and disadvantages. These approaches work on either oversampling of the minority class, undersampling of the majority class, or a combination of both. [6]

In this research, we are proposing an innovative approach of using an oversampling technique using Generative Adversarial Networks(GAN) to tackle the class imbalance issue. We would be using the same to generate synthetic data samples for our minority class data(fraud transactions) and train our models on this augmented data. As our selected data for this research has a large class imbalance ratio of **[284315:492]**, undersampling approach won't be efficient. The in-depth working of GAN and its advantages over other sampling techniques are discussed below.

- **Generative Adversarial Networks- Overview**

  Generative Adversarial Networks (GAN) is a conjunction of two deep neural networks working together to elevate each other's performance. The major components

[4] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

[5] https://medium.com/@nmscott14/3-feature-selection-methods-e7ccd6dbf316

[6] http://www.chioka.in/class-imbalance-problem/

involved in the GAN framework are [1] A generator (G) [2] A discriminator (D). Its overall operation includes iterative and concomitant training of both. The generator is used to generate samples from pure noise using latent space, whereas the discriminator is used to receiving samples generated by the generator and verify them with original data samples. In a nutshell, both G and D competes in a zero-sum game where G elevates its performance by learning to generate more genuine samples and D gradually gets unable to make the difference between synthetic and original samples (Ngwenduna and Mbuvha; 2021). Figure:-7 shows the general architecture of GAN.

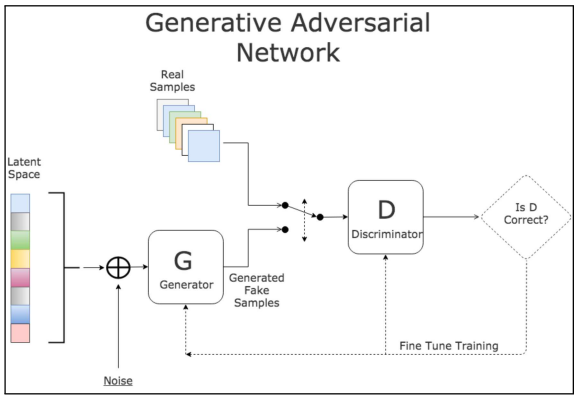


Figure 7: GAN architecture, Source:(Ngwenduna and Mbuvha; 2021)

- **Challenges associated with GAN**

  Model collapse and vanishing gradient are two major challenges associated with the training of GAN which makes use of GAN a more complex and challenging task (Ngwenduna and Mbuvha; 2021).

  - Mode collapse: Mode collapse is a phenomenon when the noise values used by the generator are mapped against like data points, which results in a lack of variations in generated data. Due to this, the generated data is skewed and may result into under-fitting of a model.
  - Vanishing gradient: This issue arises in the training of GAN when the discriminator dominates the training process and leaves no room for improvement of the generator, which results in poor augmentation.

  There are several pieces of research done to overcome the challenges associated with GAN. This process resulted in different evolved variations of GAN, each with elevated performance and advantages. We would be using CT-GAN to overcome these challenges in our research study. The working of CT-GAN is discussed in the following subsection.

- **Conditional tabular GAN (CT-GAN)**

  CT-GAN is an evolved framework developed on top of GAN architecture. It is an open-source library build on python framework [7]. To generate evenly distributed

[7] `https://sdv.dev/SDV/user_guides/single_table/ctgan.html`

samples during the training process from discrete attributes and to retain the real data distribution for augmented samples, CT-GAN uses a conditional generator. The training of conditional GAN requires an input condition to model data based on the same using conditional generator. To enable conditional generator and discriminator to capture and retain all possible correlations between given data, CT-GAN uses fully connected LSTM network structures having two fully connected layers. The conditional generator uses batch-normalization and RELU activation function. To overcome the mode collapse issue, CT-GAN uses the PacGAn framework, with 10 samples in each pac for the discriminator. Also, to get rid of the vanishing gradient challenge, CT-GAN trains the models using the WGAN loss function with gradient penalty.

After generating the synthetic data samples for minority class from our training data, we have merged the generated data with our original training set. The target class distribution for both before and after CT-GAN on training data is shown in Figure:- 8

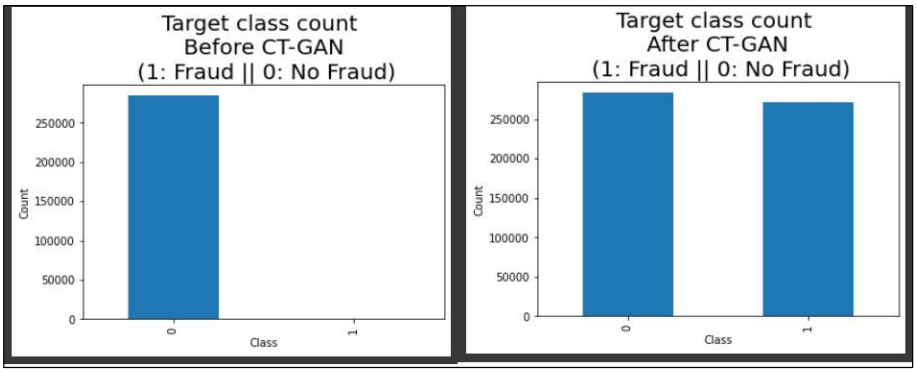


Figure 8: Target feature count before and after CT-GAN

3.4 Modelling Approach

Modelling is a crucial part of the machine learning process in the CRISP-DM methodology. After data preparation steps like feature selection and class balancing, the proposed modes are implemented on processed data. The effect of various preprocessing techniques on proposed machine learning models is evaluated and compared. The detailed working of proposed models is discussed in this section.

- **Logistic Regression:**

  Logistic regression(LR) is a supervised machine learning classifier model. There exist 3 variants of logistic regression model naming binary, multinomial and ordinal. The application of each depends on the type of classification involved in the given task. As we have only two classes present in the target feature, we would be using a binary variant of LR. It works on the fundamentals of the sigmoid function to estimate the dependency between target features and predictors. Logistic regression works on the fundamentals of the linear regression model (Setiawan et al.; 2020). The logistic regression model comes with simple yet very powerful and cost-efficient features which work best when applied to large data (Li et al.; 2020). As the data

used in our research satisfies both the requirements of the binary target class and large data size, logistic regression is a potential model for our task. Figure -9 provides better understanding of logistic regression in visualization. [8]

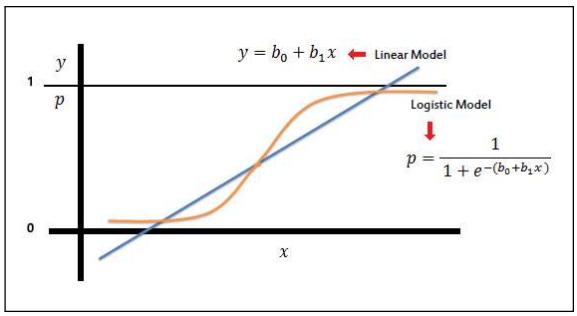


Figure 9: Logistic Regression, Source:8

- **Random Forest:**

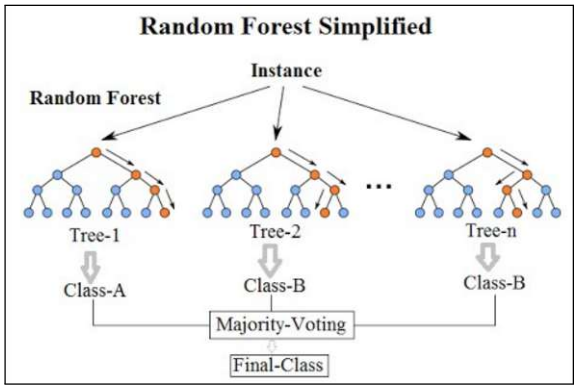


Figure 10: Random Forest, Source:9

The random forest model is an ensemble of many decision trees to solve classification problems. It works on techniques like feature randomness and bagging for building each tree, resulting in the building of uncorrelated forest of trees. Every tree in the forest is based on a basic training sample, and the number of trees in the forest directly affects the results (Khoshgoftaar et al.; 2007). This set of trees has an increased prediction performance compared to individual trees. The random forest has two-step processing as follows: (I) RF creation (II) Calculation using arbitrary forest generated by the classifier in the preliminary phase. Random forest has advantages like robustness towards outliers, ability to handle unbalanced data, high capability to handle large data which makes it a suitable choice for our research work (Sudha and Akila; 2021). Figure -10 visualizes the overall working of the Random forest model. [9]

[8] https://www.saedsayad.com/logistic_regression.htm

[9] https://williamkoehrsen.medium.com/random-forest-simple-explanation-377895a60d2d

- **XGBoost Model:** The XGBoost is one of the most popular and efficient machine learning models suitable for both regression and classification tasks. Since its invention, it has become a state-of-art machine learning model to handle structured data. XGBoost is a decision-tree-based ensemble model which works on gradient boosting methodology (Chen and Guestrin; 2016). In other words, it is an evolved version of decision tree algorithm and unlike decision tree, it has some powerful features like robustness for missing values, regularization to avoid overfitting, parallelized processing, optimized gradient boosting (combined performance of software and hardware optimization technique) which makes it potential choice for our research work. Figure -11 explains the concept of gradient boosting visually.

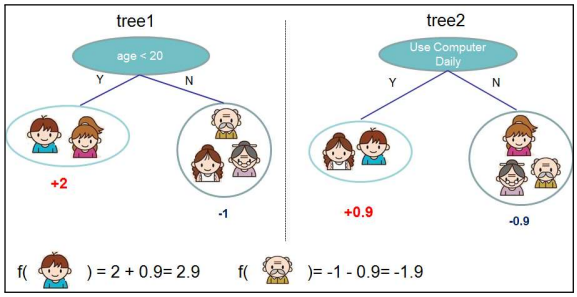


Figure 11: XGBoost Model- Gradient Boosting, Source:(Chen and Guestrin; 2016)

4 Design Specification

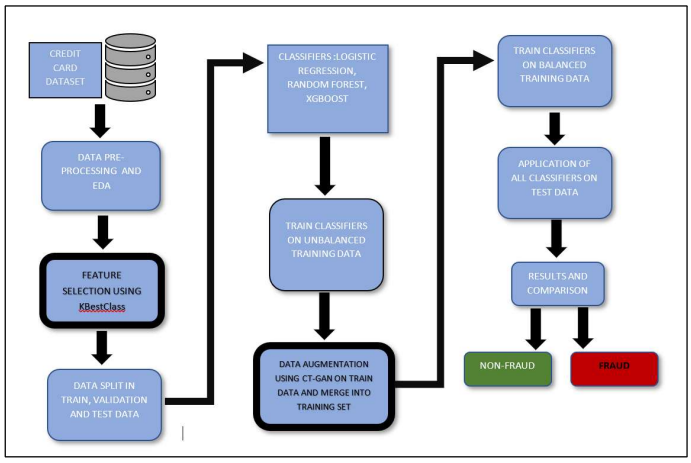


Figure 12: Design Architecture

Figure -12 illustrates the process flow diagram designed for our research. At the early stage, we are fetching data from the source followed by pre-processing and EDA(explanatory data analysis) steps which involves removal of null and duplicate values and finding hidden patterns from the data. In a later step, we are filtering our features using the KBestClass feature selection method to keep only relevant columns for our analysis. After this, we have split our data into train, validation and test dataset followed by training of our

baseline models on unbalanced training data set. At a later stage, we would be doing data augmentation using CT-GAN to generate samples for the minority class. We are merging this augmented data with our training data to balance it. After that, we are proceeding with the training of three classifier models on these balanced training datasets. The performance of classifiers trained on both unbalanced and balanced training sets is tested and evaluated at the final stage using the test dataset. At the final stage, we are evaluating and comparing the results for both case scenarios.

5 Implementation

In this section, the implementation steps followed in the proposed research work are described in detail. Also, it explains the undertaken techniques for the selection of relevant features and processes incorporated for handling datasets using CT-GAN architecture. All implementation of the proposed methodology has been carried out using Python language (v.3.7) and Google Colab has been used as an integrated development environment (IDE). Python has been identified as the best choice for our implementation as it has a wide online support forum, easy yet very powerful features and offers great code readability. Due to its high availability packages for data handling and pre-processing, python has been a prime choice for machine learning projects.

The data we are using for our research is publicly hosted in CSV format [10]. It consists of credit card transactions since 2013 for European cardholders. Due to confidentiality concerns, all the sensitive features are transformed in form of PCA components. The dataset contains 31 features in total, including the target variable class, which signifies whether the given sample transaction is a fraudulent or legitimate one. We have imported the data into pandas data frame [11] using Python. After cleaning and scaling the data, we have analysed the data using visualizations to understand the patterns and correlations present in the data. After analysing correlations of all features with the target variable, we have identified the significant features which are highly correlated to the target variable. To cross verify our choice of selected features, we have used the KBestClass feature selection technique from the feature_selection library of sklearn package. [12]. Upon obtaining the final dataset, we have divided our dataset into train, validation, and test sets for further usage.

While exploring our data, we have observed a huge class imbalance for our target variable. To achieve reliable results with machine learning models and to avoid overfitting, we have used the tabular data modelling technique CT-GAN [13] for generating samples for minority class and merging them to our original data to balance the same. CT-GAN is an open-source python package developed at MIT, consisting of a Conditional Generative adversarial network specifically designed for modelling tabular data. After that, we have applied various classifier models on our balanced data for the classification of given samples into fraud or non-fraud. We have used Random Forest, Logistic regression and XGBoost classifiers in our approach, which are available in the python sklearn library. We are using a balanced dataset consisting of augmented data samples only for the training of our models. For validating the training results, we have used a validation

[10] https://www.kaggle.com/mlg-ulb/creditcardfraud

[11] https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html

[12] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

[13] https://sdv.dev/CTGAN/

set of data to fine-tune our models using hyperparameter tuning. We are testing our models on a testing set of data only after fine-tuning our models. We are comparing the performance of all classifier models on both unbalanced and balanced data using two different case studies. The evaluation metrics used in the research consist of specificity, sensitivity, F1- score, AUC-ROC score, Geometric Mean. [14]. We are using a confusion matrix for obtaining these parameters. After testing our fine-tuned models on testing data and evaluating the results using evaluation metrics, we have observed that random forest outperformed the other two models, obtaining the highest accuracy and AUC score with minimal errors for both unbalanced and balanced approaches. Logistic regression stands the lowest on performance metrics, but showed a grate increase in AUC score when trained on balanced data. The detailed evaluation and performance comparison is explained out in the following sections.

6 Evaluation

The main objective of this research work is to make combined use of supervised machine learning techniques and GAN and evaluate the results whether our proposed approach elevates the model performances compared to other state-of-art approaches or not. To do the comparative study, we have carried out two experiments as following: [1] Evaluating model performances by training on unbalanced data. [2] Evaluating the model performances by training on balanced data augmented by GAN. To form a common ground for evaluating the performances of our models we have selected the metrics like recall, specificity, F1-score, AUC score, AUC-ROC curve and the geometric mean of recall and specificity. Due to the imbalanced nature of our data, we cannot evaluate the performance of models based on their accuracy. (Adepoju et al.; 2019) To derive our above-selected metrics, we have used the confusion matrix, as explained below.[15]

- **True Positive(TP) -** It shows that the given model has accurately identified actual non-fraud(true)cases as non-fraud(positive)
- **False Positive(FP) −** It shows that the model has inaccurately identified actual fraud(False) cases as non-fraud(positive).
- **False Negative(FN) −** It shows that the model has inaccurately identified actual non-fraud (False) cases as fraud(negative).
- **True Negative(TN) −** It shows that the model has accurately identified actual fraud (True) cases as fraud(negative).

Precision/specificity signifies the percentage of transactions that are classified as fraud and are frauds. On the other hand, recall/sensitivity values measures the percentage of actual fraud transactions that are correctly classified. F1-score is a harmonic mean between precision and recall. and should be close to 1 for better classification (Ngwenduna and Mbuvha; 2021). The geometric mean is an aggregate of both specificity and sensitivity. We have selected this metric for our model evaluation as it is suitable for unbalanced data (Tharwat; 2020). Due to the imbalanced nature of our data, recall and AUC score are the most crucial evaluation metric in our research.

[14] https://www.ritchieng.com/machine-learning-evaluate-classification-model/

[15] https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

To calculate above-mentioned evaluation metrics using confusion matrix, we have used following equations.

$$\textbf{Recall/Sensitivity} = \frac{TP}{TP + FN}$$

$$\textbf{Precision/Specificity} = \frac{TP}{TP + FP}$$

$$\textbf{F1-Score} = \frac{2 * Recall * Precision}{Recall + Precision}$$

$$\textbf{Geometric Mean(GM)} = \sqrt{Sensitivity * Specificity}$$

After the application of various data preparation methods discussed in the above sections, we have obtained our final dataset for the application of our selected models. To examine the effectiveness of our proposed technique of using CT-GAN for handling the data imbalance challenge, we have carried out two separate case studies for all the algorithms and evaluated them using the above-discussed evaluation metrics. Both case study setups are discussed in the following sections.

6.1 Case Study 1: Model performances before CT-GAN

In our first experiment, we have trained a base model for each classifier using unbalanced training data after obtaining final data using feature selection and data split. The Table 1 summarizes the performance metrics for each classifier. The lower recall values in Table 1 signifies that despite giving higher accuracy our model is failing to predict the fraud transactions. This signifies the high rate of overfitting of our models due to class imbalance in data. The same can be cross verified by lower values of AUC scores. In this case study, among all 3 classifiers, Random Forest outperforms the other two in terms of all metrics with the recall of 75%, whereas logistic regression performed the worst with the lowest recall value of 57%. To test our hypothesis, 'whether data balancing using CT-GAN elevates the classifier performance or not' we have further applied the same classifiers on balanced data in another case study.

Table 1: Model Performances on Unbalanced Data.

Classifier	Recall	F1-Score	AUC Score	Geometric Mean
Random Forest	75%	84%	87%	85%
XGBoost	72%	81%	86%	82%
Logistic Regression	57%	69%	78%	70%

6.2 Case Study 2: Model performances after CT-GAN

In our second case study, we have tested our classifiers on balanced data using CT-GAN and summarized the results in Table 2. It is evident from the observed values in the given table as after training our models on balanced data, their performance has

improved significantly. Among all classifiers, Random Forest tops the list with 100% recall value after being trained on balanced data as well. This high recall value signifies that Random Forest has not missed a single fraud transaction among all data. Also, compared to 6.1 Logistic Regression, the classifier also performed very well with a recall value of 81%. Overall, the significant surge in model performance signifies the effectiveness of our proposed approach. We will examine the effect of CT-GAN on all classifies in-depth in the following section.

Table 2: Model Performances on Balanced Data.

Classifier	**Recall**	**F1-Score**	**AUC Score**	**Geometric Mean**
Random Forest+CTGAN	100%	100%	100%	100%
XGBoost+CTGAN	84%	84%	92%	84%
Logistic Regression+CTGAN	81%	81%	91%	81%

6.3 Discussion

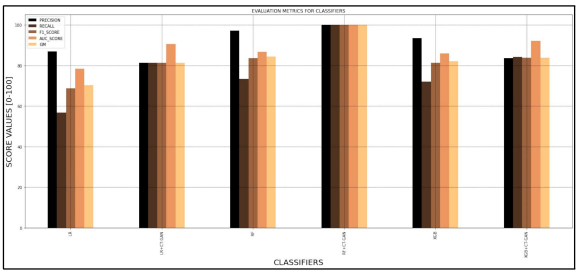
Given the high weightage of fraud transactions in our classification task, we cannot afford our model to miss-classify fraud transactions as non-fraud. To ensure this, we have considered recall as a crucial evaluation metric in our research. As recall signifies the percentage of how accurately our model is detecting the fraud transactions, having a high recall value indicates the higher performance of our models. Also, F1-score signifies the overall performance of our models as it provides the balance between precision and recall scores. The higher F1-score value for a model signifies that the model is predicting both fraud and non-fraud transactions with minimum errors.

After analysing the results summarised in Table:-1 and Table:-2 we can state that the Random Forest model outperforms all other classifiers in all cases and proved to be a better classifier even when applied on imbalanced data. By implementing CT-GAN for the generation of synthetic data samples of the minority class, we have balanced our data and trained our models on the same. By studying the results, we can state that pour proposed data augmentation technique elevated the model performances in all terms. To get reliable results out of our models, we have used the augmented data for training only and tested the trained models by using real samples of original data.

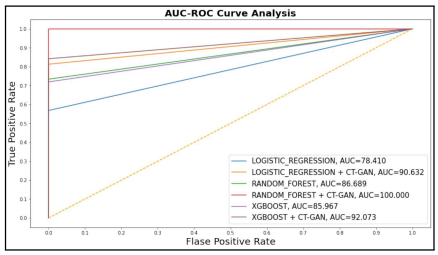
Table 3: Percentage Increase in Model Performances After Data Balancing

Classifier	**Recall**	**F1-Score**	**AUC Score**	**Geometric Mean**
Random Forest	25%	16%	13%	15%
Logistic Regression	24%	12%	13%	11%
XGBoost	12%	3%	6%	2%

Table:-3 summarises the percentage increase in evaluation parameter values after application of CT-GAN. Out of all classifiers, Random Forest and Logistic Regression have observed major improvement in recall and F1-score values. Random forest and Logistic Regression has recorded an overall 25% increase in recall score, which signifies that they have made 25% fewer errors while classifying fraud transactions after being trained on balanced data. On the other hand, the XGBoost classifier recorded the lowest increase



(a) Evaluation Metrics Summary



(b) AUC-ROC Curve for All Classifiers

Figure 13: Evaluation Metrics Before and After CT-GAN

in recall value after the application of CT-GAN. In terms of Geometric Mean, Random Forest is followed by the XGBoost model, having a Gmean value of 84%. AUC-ROC is another important evaluation metric which visually indicates the performance of the given model. It is a graph plotted between True positive and false positive values predicted by the model. The area under the curve signifies the probability of model being correct while differentiating fraud transactions from non-fraud transactions. The larger area under the curve is an indicator of a better performing model. Figure:-13 summarises the evaluation metrics for all the classifiers visually. By looking at the overall results, evaluation parameters and AUC-ROC curve, we can suggest that the Random Forest classifier is a more reliable model for the fraud detection task. Additionally, Random Forest + CT-GAN architecture provides the most accurate fraud detection with no errors and could be verified by applying other feature engineering methods.

7 Conclusion and Future Work

As can be seen from the above-studied literature, the issue of credit card fraud detection has piqued attention in the last few years, and a variety of studies have been employed in the pursuit of optimal prediction performance using machine learning. As discussed earlier, class imbalance is a stumbling block in the field of credit card fraud detection. We utilized a hybrid combination of a data modelling method-CT-GAN and feature selection technique-SelectKBest in this study to handle the class imbalance challenge while using machine learning models and developed 3 models using this approach. In terms of recall value and F1-score values, Random Forest combined with SelectKbest and CT-GAN outperforms all other classifiers(LR, XGB), giving 100% recall and F1-score values. Also, after being trained on augmented data using CT-GAN all the classifiers have shown a significant surge in overall prediction performance with the drop in errors. The data used for this research consisted of all numerical features. In future, this approach can be tested on categorical data to verify its reliability. The proposed methodology can be implemented to solve other classification problems from other domains.

Acknowledgement

I'd want to express my heartfelt thanks to my supervisor, Dr. Bharathi Chakravarthi, for giving regular feedback and helpful ideas during the research project's implementation phase. I'd also like to express my gratitude to my family and friends for their unwavering support and encouragement, without which this study would not have been possible.

References

Adepoju, O., Wosowei, J., Jaiman, H. et al. (2019). Comparative evaluation of credit card fraud detection using machine learning techniques, *2019 Global Conference for Advancement in Technology (GCAT)*, IEEE, pp. 1–6.

Al Smadi, B. and Min, M. (2020). A critical review of credit card fraud detection techniques, pp. 0732–0736.

Asha, R. and KR, S. K. (2021). Credit card fraud detection using artificial neural network, *Global Transitions Proceedings* .

Bhattacharyya, S., Jha, S., Tharakunnel, K. and Westland, J. C. (2011). Data mining for credit card fraud: A comparative study, *Decision Support Systems* **50**(3): 602–613. On quantitative methods for detection of financial fraud.
URL: *https://www.sciencedirect.com/science/article/pii/S0167923610001326*

Brennan, P. (2012). A comprehensive survey of methods for overcoming the class imbalance problem in fraud detection, *Institute of technology Blanchardstown Dublin, Ireland* .

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system, p. 785–794.
URL: *https://doi.org/10.1145/2939672.2939785*

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B. and Bharath, A. A. (2018). Generative adversarial networks: An overview, *IEEE Signal Processing Magazine* **35**(1): 53–65.

Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S. and Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective, *Expert systems with applications* **41**(10): 4915–4928.

Fiore, U., De Santis, A., Perla, F., Zanetti, P. and Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection, *Information Sciences* **479**: 448–455.
URL: *https://www.sciencedirect.com/science/article/pii/S0020025517311519*

Gui, J., Sun, Z., Wen, Y., Tao, D. and Ye, J. (2020). A review on generative adversarial networks: Algorithms, theory, and applications, *arXiv preprint arXiv:2001.06937* .

John, H. and Naaz, S. (2019). Credit card fraud detection using local outlier factor and isolation forest, *Int. J. Comput. Sci. Eng* **7**(4): 1060–1064.

Khatri, S., Arora, A. and Agrawal, A. P. (2020). Supervised machine learning algorithms for credit card fraud detection: a comparison, pp. 680–683.

Khoshgoftaar, T. M., Golawala, M. and Van Hulse, J. (2007). An empirical study of learning from imbalanced data using random forest, **2**: 310–317.

Le, T., Vo, M. T., Vo, B., Lee, M. Y. and Baik, S. W. (2019). A hybrid approach using oversampling technique and cost-sensitive learning for bankruptcy prediction, *Complexity* **2019**.

Li, Z., Liu, G. and Jiang, C. (2020). Deep representation learning with full center loss for credit card fraud detection, *IEEE Transactions on Computational Social Systems* **7**(2): 569–579.

Liu, H. and Motoda, H. (1998). *Feature extraction, construction and selection: A data mining perspective*, Vol. 453, Springer Science & Business Media.

Lucas, Y., Portier, P.-E., Laporte, L., He-Guelton, L., Caelen, O., Granitzer, M. and Calabretto, S. (2020). Towards automated feature engineering for credit card fraud detection using multi-perspective hmms, *Future Generation Computer Systems* **102**: 393–402.

Ngwenduna, K. S. and Mbuvha, R. (2021). Alleviating class imbalance in actuarial applications using generative adversarial networks, *Risks* **9**(3): 49.

Oblé, F. and Bontempi, G. (2019). Deep-learning domain adaptation techniques for credit cards fraud detection, **1**: 78–88.

Pandey, A., Bhatt, D. and Bhowmik, T. (n.d.). Limitations and applicability of gans in banking domain.

Sadineni, P. K. (2020). Detection of fraudulent transactions in credit card using machine learning algorithms, pp. 659–660.

Setiawan, Q. S., Rustam, Z., Hartini, S., Wibowo, V. V. P. and Aurelia, J. E. (2020). Comparing decision tree and logistic regression for pancreatic cancer classification, pp. 623–627.

Sudha, C. and Akila, D. (2021). Credit card fraud detection system based on operational transaction features using svm and random forest classifiers, pp. 133–138.

Thabtah, F., Hammoud, S., Kamalov, F. and Gonsalves, A. (2020a). Data imbalance in classification: Experimental evaluation, *Information Sciences* **513**: 429–441.

Thabtah, F., Hammoud, S., Kamalov, F. and Gonsalves, A. (2020b). Data imbalance in classification: Experimental evaluation, *Information Sciences* **513**: 429–441.
URL: *https://www.sciencedirect.com/science/article/pii/S0020025519310497*

Tharwat, A. (2020). Classification assessment methods. appl comput inf, *Press. Google Scholar* .

Wang, D., Lin, J., Cui, P., Jia, Q., Wang, Z., Fang, Y., Yu, Q., Zhou, J., Yang, S. and Qi, Y. (2019). A semi-supervised graph attentive network for financial fraud detection, pp. 598–607.

Yilmaz, I., Masum, R. and Siraj, A. (2020). Addressing imbalanced data problem with generative adversarial network for intrusion detection, pp. 25–30.

Zhu, H., Liu, G., Zhou, M., Xie, Y., Abusorrah, A. and Kang, Q. (2020). Optimizing weighted extreme learning machines for imbalanced classification and application to credit card fraud detection, *Neurocomputing* **407**: 50–62.

Zulfiker, M. S., Kabir, N., Biswas, A. A., Nazneen, T. and Uddin, M. S. (2021). An in-depth analysis of machine learning approaches to predict depression, *Current Research in Behavioral Sciences* **2**: 100044.