CREDIT CARD FRAUD DETECTION USING NEW PREPROCESSING AND HYBRID MACHINE LEARNING TECHNIQUES

ESRAA FAISAL MALIK GASIM

UNIVERSITI SAINS MALAYSIA

2023

CREDIT CARD FRAUD DETECTION USING NEW PREPROCESSING AND HYBRID MACHINE LEARNING TECHNIQUES

by

ESRAA FAISAL MALIK GASIM

Thesis submitted in fulfillment of the requirements
for the degree of
Doctor of Philosophy

July 2023

ACKNOWLEDGEMENT

The Prophet Muhammad ﷺ said:
'He who is not grateful to people, is not grateful to Allah'

Throughout the writing of this dissertation, I have received a great deal of support and assistance. Words cannot express my gratitude to my supervisor Ts. Dr. Khaw Khai Wah, your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. I would also like to thank my co-supervisor Ts. Dr. Chew XinYing, you provided me with the tools I needed to successfully complete my work and finish my dissertation.

I also could not have undertaken this journey without my defense committee members for letting my proposal defense be an enjoyable moment and for your valuable comments and suggestions that helped me sharpen my work and dissertation. Thanks should also go to the librarians, and research participants from the university especially SOM, who impacted and inspired me.

This endeavor would not have been possible without my family, especially my parents, brothers, and husband for their unlimited emotional support. Their belief in me has kept my spirits and motivation high during this process. I am also grateful to my friends who have supported me along the way. Thanks to all of you !

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

ACFE	Association of Certified Fraud Examiner
AdaBoost	Adaptive Boosting
AML	Anti-money Laundering
ARL	Association Rule Learning
AUC-PR	Area Under the Precision–Recall Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
BALO	Binary Ant Lion Optimizer
BBN	Bayesian Belief Network
BG	Bagging Algorithms
BN	Bayesian Network
CART	Classification and Regression Trees
CBoost	Cluster-Based Boosting
CCFDD	Credit Card Fraud Detection Domain
CFGVM	Cost-sensitive Feature Selection General Vector Machine
CIS	Computational Intelligence Society
CLOPE	Clustering with sLOPE
CNN	Condensed Nearest Neighbors Rule
COVID-19	Coronavirus Pandemic
CRISP-DM	Cross-Industry Standard Process for Data Mining
C-RTDL	Cost-Sensitive Learning and Resampling Technique at Data-Level
CSL	Cost-Sensitive Learning
DL	Deep Learning
DS	Decision Stump
DT	Decision Tree

EFB	Exclusive Feature Bundling
ENN	Edited Nearest Neighbors
FB	False Positive
FFS	Fraudulent Financial Statement
FLDA	Fisher Linear Discriminant Analysis
FN	False Negative
FNN	Fuzzy Logic and ANN
FNR	False Negative Rate
FPR	False Positive Rate
GA	Genetic Algorithm
GAN	Generative Adversarial Network
GB	Gradient Boosting
GBDT	Gradient Boosting Decision Tree
GBT	Gradient Boosted Tree
GOSS	Gradient-Based One-Side Sampling
GRU	Gated Recurrent Unit
HAOC	Oversampling Technique and CSL
HHM	Hidden Markov Models
ID3	Quinlan's iterative Dichotomiser 3
IEEE-CIS	IEEE Computational Intelligence Society
iForest	Isolation Forest
IR	Imbalance Ratio
KNN	K-Nearest Neighbors
LDA	Linear Discriminant Analysis
LGBM	Light Gradient Boosting
Lin-LS-SVM	Linear Kernel Least-Squares Support Vector Machine
LOF	Local Outlier Factor

LR	Logistic Regression
LSTM	Long Short-Term Memory Network
MCC	Matthew’s Correlation Coefficient
MCE	Multiple Algorithms Ensemble
MFDA	Modified Fisher Discriminant Analysis
ML	Machine Learning
MLP	Multilayer Perceptron
NaNSMOTE	Minority Oversampling Technique with Natural Neighbors
NB	Naive Bayes
NHI	Taiwan's National Health Insurance
NN	Neural Network
Non-FFS	Non-Fraudulent Financial Statement
OLGBM	Optimized Light Gradient Boosting
OPT	Optimistic Voting
OPWEM	Optimistic, Pessimistic, and Weighted Voting in an Ensemble of Models
OSS	One-Sided Selection
PCA	Principal Component Analysis
PES	Pessimistic Voting
PNN	Probabilistic Neural Network
PWC	PricewaterhouseCoopers
QDA	Quadratic Discriminant Analysis
RAM	Random-Access Memory
RBF	Radial Basis Function
RBM	restricted Boltzmann Machine
RF	Random Forest
RHSBoost	Random Hybrid Sampling Boosting

ROC	Receiver Operating Characteristic
ROS	Random Oversampling
ROSE	Random Oversampling Examples
RT	Random Tree
RTDL	Resampling Technique at the Data-Level
RUS	Random Undersampling
SMO	Sequential Minimal Optimization
SMOTE	Synthetic Minority Oversampling Technique
SMOTE-ENN	Synthetic Minority Oversampling Technique with Edited Nearest Neighbors
SOM	Self-Organizing Maps
SVM	Support Vector Machine
SVM-RFE	Support Vector Machine-Recursive Feature Elimination
TAN	Transductive Adversarial Networks Bayes Classification
TLD	Top Level Domain
TN	True Negative
TP	True Positive
TPR	True Positive Rate
ULB	Université Libre de Bruxelles
US	United States
USD	United States Dollars
WGT	Weighted Voting
XGBoost	eXtreme Gradient Boosting

LIST OF APPENDICES

PENGESANAN PENIPUAN KAD KREDIT MENGGUNAKAN TEKNIK PRA PEMPROSESAN BAHARU DAN PEMBELAJARAN MESIN HIBRID

ABSTRAK

Salah satu masalah cabaran penting dalam domain penipuan kad kredit ialah peningkatan jumlah data yang tidak seimbang. Nisbah kelas majoriti kepada minoriti yang lebih tinggi boleh membawa kepada keputusan yang mengelirukan, kerana algoritma pembelajaran mesin konvensional menganggap pengagihan kelas yang sama. Sumbangan pertama penyelidikan ini adalah untuk membangunkan teknik prapemprosesan baharu yang menggunakan pembelajaran sensitif kos dan teknik pensampelan semula pada peringkat data untuk meningkatkan prestasi set data yang sangat tidak seimbang. Teknik prapemprosesan yang dibangunkan terdiri daripada tiga fasa. Pada fasa pertama, beberapa teknik pensampelan semula pada peringkat data, seperti SMOTE-ENN, SMOTE-TOMEK, SMOTE-OSS, SMOTE-RUS dan ROS-RUS, dengan parameter lalai mereka dibandingkan untuk mencari teknik optimum dengan prestasi tertinggi. Fasa kedua melibatkan penggunaan pembelajaran sensitif kos dengan nisbah yang berbeza untuk menentukan julat nisbah terbaik untuk digunakan dalam fasa tiga. Selepas itu, dalam fasa ketiga, peratusan teknik pensampelan semula pada peringkat data diperhalusi untuk mengelakkan kehilangan maklumat penting atau menghasilkan data sintetik berulang yang boleh menyebabkan overfitting. Selain itu, nisbah pembelajaran sensitif kos diperhalusi untuk menentukan kos salah klasifikasi dalam kelas minoriti. Teknik prapemprosesan baharu yang dibangunkan didapati memberi impak positif dari segi ukuran F1 dan kadar salah klasifikasi berbeza dengan teknik pensampelan semula konvensional. Tambahan pula, kesan negatif jenayah kewangan terhadap institusi kewangan telah berkembang secara mendadak sejak beberapa tahun. Sumbangan kedua kepada penyelidikan ini adalah untuk membangunkan pelbagai

model pembelajaran mesin hibrid untuk meningkatkan pengesanan aktiviti penipuan dalam domain pengesanan penipuan kad kredit. Model hibrid yang dibangunkan terdiri daripada dua fasa. Pertama, algoritma pembelajaran mesin konvensional, iaitu Mesin Vektor Sokongan, Regresi Logistik, Hutan Rawak, Pohon Keputusan, Peningkatan Kecerunan eXtreme, dan LightGBM, digunakan terlebih dahulu untuk mengesan transaksi penipuan. Kedua, model pembelajaran mesin hibrid telah dibina berdasarkan algoritma tunggal terbaik dari fasa pertama. Untuk menentukan algoritma tunggal berprestasi terbaik dalam setiap set data dan oleh itu paling sesuai digunakan sebagai algoritma pertama dalam model hibrid yang dicadangkan, algoritma pembelajaran mesin konvensional telah dibandingkan menggunakan keputusan daripada teknik prapemprosesan baharu yang dicadangkan dari segi F1 -kadar pengukuran dan salah klasifikasi. Hutan Rawak dikenal pasti sebagai model garis dasar yang optimum kerana prestasi unggulnya dalam ukuran F1 dan kadar salah klasifikasi. Selepas itu, lima model pembelajaran Mesin hibrid telah dibangunkan dengan menghibridkan Random Forest dengan algoritma pembelajaran Mesin yang lain, termasuk RF-SVM, RF-LR, RF-DT, RF-XGBoost dan RF-LGBM. Penyelidikan ini mengikuti versi diubah suai bagi metodologi Proses Standard Merentas Industri untuk Perlombongan Data, merangkumi pengumpulan data, penerokaan, prapemprosesan (mengendalikan nilai yang hilang, mengubah ciri kategori, penskalaan ciri, pemilihan ciri dan pensampelan semula), pembangunan model dan penilaian. Prestasi model hibrid yang dicadangkan dinilai menggunakan dua set data: IEEE-CIS dan Credit_Card_Fraud_Detection. Metrik penilaian seperti Kawasan Di Bawah Keluk Ciri Operasi Penerima, F1-measure, ingat semula, ketepatan dan kadar salah klasifikasi digunakan untuk menilai keputusan. Penemuan menunjukkan bahawa model hibrid RF-XGBoost adalah model juara kerana ia memaparkan prestasi tertinggi dari segi ukuran F1 dalam dua set data yang digunakan (0.83516 dan 0.94444, untuk dataset IEEE-CIS dan Credit_Card_Fraud_Detection, masing-

masing). Hasil penyelidikan ini amat dihargai oleh sektor kewangan kerana hasilnya diharapkan dapat membantu organisasi, bank dan institusi kewangan dalam mengenal pasti dengan jelas aktiviti penipuan, oleh itu, mengurangkan kos operasi yang meningkat daripada penggera palsu.

CREDIT CARD FRAUD DETECTION USING NEW PREPROCESSING AND HYBRID MACHINE LEARNING TECHNIQUES

ABSTRACT

One of the significant problems in the credit card fraud domain is the increasing number of imbalanced data. The higher ratio of majority to minority classes can lead to misleading results, as conventional machine learning algorithms assume equal class distribution. The first contribution of this research is to develop a new preprocessing technique that utilizes cost-sensitive learning and resampling techniques at the data-level to improve the performance of highly imbalanced datasets. The developed preprocessing technique consists of three phases. In the first phase, several resampling techniques at the data-level, such as SMOTE-ENN, SMOTE-TOMEK, SMOTE-OSS, SMOTE-RUS, and ROS-RUS with their default parameters, are compared to find the optimum technique with the highest performance. The second phase involves using cost-sensitive learning with different ratios to determine the best range of ratios to be used in phase three. Subsequently, in the third phase, the percentage of resampling techniques at the data-level is fine-tuned to avoid losing crucial information or producing repetitive synthetic data that could cause overfitting. Additionally, the cost-sensitive learning ratio is fine-tuned to determine the misclassification costs in the minority class. The developed new preprocessing technique was found to have a positive impact in terms of F1-measure and misclassification rate in contrast to the conventional resampling techniques. Furthermore, the negative effect of financial crimes on financial institutions has grown dramatically over the years. The second contribution to this research is to develop multiple hybrid machine learning models in order to enhance the detection of fraudulent activities in the credit card fraud detection domain. The developed hybrid models consist of two phases.

Firstly, conventional machine learning algorithms, namely Support Vector Machine, Logistic Regression, Random Forest, Decision Tree, eXtreme Gradient Boosting, and LightGBM, were used first to detect the fraudulent transactions. Secondly, hybrid machine learning models were constructed based on the best single algorithm from the first phase. To determine which single algorithm perform best in each dataset and are thus most suitable for use as the first algorithm in the suggested hybrid models, the conventional machine learning algorithms were compared using the results from the developed new preprocessing technique in terms of F1-measure and misclassification rate. Random Forest was identified as the optimal baseline model due to it is superior performance in F1-measure and misclassification rate. Subsequently, five hybrid Machine learning models were developed by hybridizing Random Forest with other Machine learning algorithms, including RF-SVM, RF-LR, RF-DT, RF-XGBoost, and RF-LGBM. This research follows a modified version of the Cross-Industry Standard Process for Data Mining methodology, encompassing data collection, exploration, preprocessing (handling missing values, transforming categorical features, feature scaling, feature selection, and resampling), model development, and evaluation. The performance of the developed hybrid models is evaluated using two datasets: IEEE-CIS and Credit_Card_Fraud_Detection. Evaluation metrics such as Area Under the Receiver Operating Characteristics Curve, F1-measure, recall, precision, and misclassification rate are employed to assess the results. The findings indicated that the hybrid model RF-XGBoost is the champion model as it displayed the highest performance in terms of F1-measure in the two used datasets (0.83516 and 0.94444, for IEEE-CIS and Credit_Card_Fraud_Detection dataset, respectively). The results of this research are appreciable to the financial sector as the outcome can hopefully assist the organizations, banks, and financial institutions in clearly identifying fraudulent activities, therefore, lowering operational costs rising from false alarms.

CHAPTER 1

INTRODUCTION

1.1 Background of the Research

Black's Law Dictionary defined fraud as "A knowing misrepresentation of the truth or concealment of a material fact to induce another to act to his or her detriment" (Garner, 2004). Financial fraud is the deliberate use of illegal techniques and activities to achieve financial gains (Zhou & Kapoor, 2011). There are two different strategies to defeat fraud, namely, fraud detection and fraud prevention. Both strategies are used to enhance the security system for the organization, however with slight differences. Fraud prevention is a proactive strategy that is used to prevent deception and fraud events from occurring (Sahin et al., 2013). Nevertheless, this method alone is insufficient to prevent fraud as fraudsters can easily surpass the fraud prevention systems. On the other hand, fraud detection is the process of identifying and recording fraud events that occurred and notifying the system administrator (Ngai et al., 2011).

Recently, the rapidly changing world and the evolving financial industry have led to an ease in individual's life, especially in the time of the Coronavirus Pandemic (COVID-19), as numerous services were forced to shift to online platforms including health care, banking, business, education, essential government services and entertainment (Hakak et al., 2020; K. W. F. Ma & McKinnon, 2021). As a result of the widespread availability of the internet and the ease with which web users can conceal their location and identity during online transactions, fraudsters have quickly evolved to benefit from the new fast-moving digital (Faraji, 2022). The United States (US) Federal Trade Commission estimated that United State dollars (USD) 12 million were

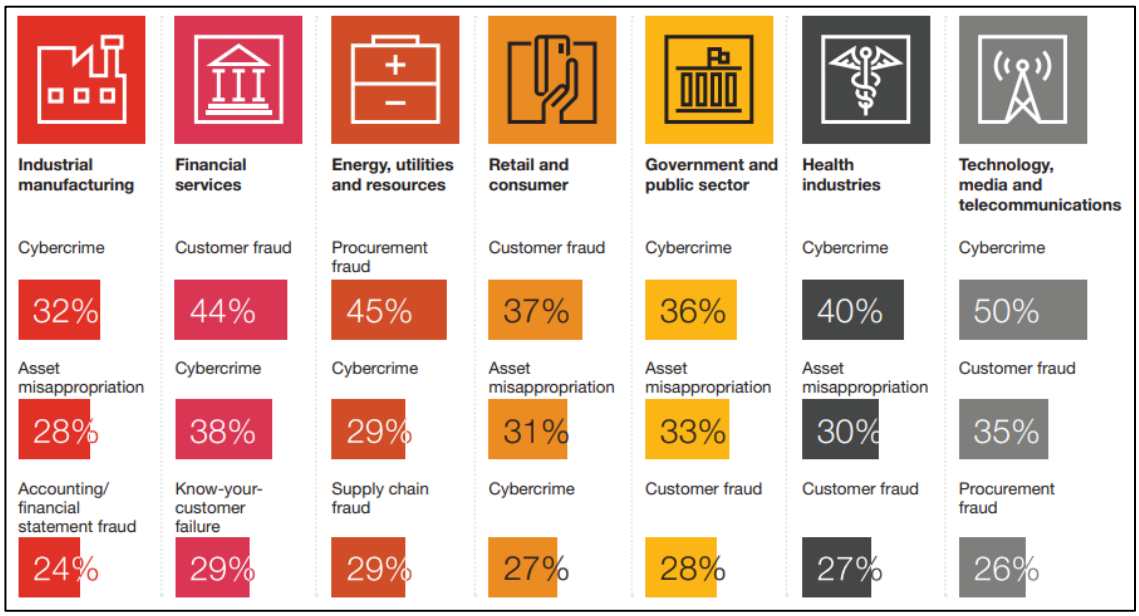
lost from COVID-19 which was associated with fraudulent activities between January 1st and April 14th, 2020 (Witt, 2020).

Although there has been a long-standing interest in fraud detection, it starts to become an international problem affecting both domestic and global markets as many studies reported a huge amount of losses in different countries throughout the years (Abdallah et al., 2016; Ata & Hazim, 2020; Bhattacharyya et al., 2011; Carneiro et al., 2017; de Sá et al., 2018; Kültür & Çağlayan, 2017; Manlangit et al., 2019; Sarno et al., 2015; Zhang et al., 2021). For instance, in the year of 2020, Malaysian organizations continued to experience high levels of fraud, having customer fraud with the most disruptive impact on organizations (20%), followed by bribery and corruption (18%), cybercrime, and asset misappropriation each have 16% of the total fraud (PricewaterhouseCoopers [PWC], 2020a).

Internationally, this issue has risen to the highest level in the world over the last 20 years. According to the Global Economic Crime and Fraud Survey 2022 conducted by PwC, fraudsters have been swift to adapt to the emergence of new platforms and exploit vulnerabilities in the security framework. The survey, which involved 1,296 organizations from 53 countries and regions, revealed that 51% of the surveyed organizations reported incidents of fraud between 2020 to 2022, constituting the highest level documented in PwC's research over a span of two decades (PwC, 2022).

Figure 1.1 illustrates that customer fraud (e.g., identity theft, credit card fraud, and mortgage fraud) indicated the highest occurrence rate for fraud incidents in financial services and retail and consumer, which equaled 44% and 37% worldwide, respectively (PwC, 2022). The problem of fraud caught the governments and financial institutions' concerns not only because of the monetary losses due to fines, penalties,

responses, and remediation but also because these acts have been responsible for the sudden failure of many reputable institutions, causing them brand damage, loss of market position, employee morale, and loss of future opportunities (Ali et al., 2021; Lim et al., 2017; Sule et al., 2019).



(Source: PwC, 2022)

Figure 1.1 Crimes frequency of overall experience

As reported by the (Association of Certified Fraud Examiners [ACFE], 2020), fraud events can occur across a wide range of private and public institutions as well as throughout the economy. Nevertheless, government, public administration, manufacturing, financial services, and banking, were particularly vulnerable to fraudulent activities. However, ACFE pointed out that the high fraud rate in these areas does not necessarily imply that there is more fraud in these industries, rather, it could simply indicate that companies in these industries employ more certified fraud examiners than others.

Based on previous efforts as displayed in Figure 1.2, five main fields are vulnerable to fraud, namely, financial statement fraud, money laundering fraud, bitcoin fraud, fraud insurance and credit card fraud (Abdallah et al., 2016; Al-Hashedi & Magalingam, 2021; Behdad et al., 2012; Bhattacharyya et al., 2011). The results from the most recent survey conducted by Al-Hashedi and Magalingam (2021) have shown that bank fraud is the most focused area in the literature. Furthermore, their results indicated that most machine learning (ML) algorithms have been widely applied to bank fraud, particularly credit card fraud, which accounts for 40% of all papers studied in the survey.

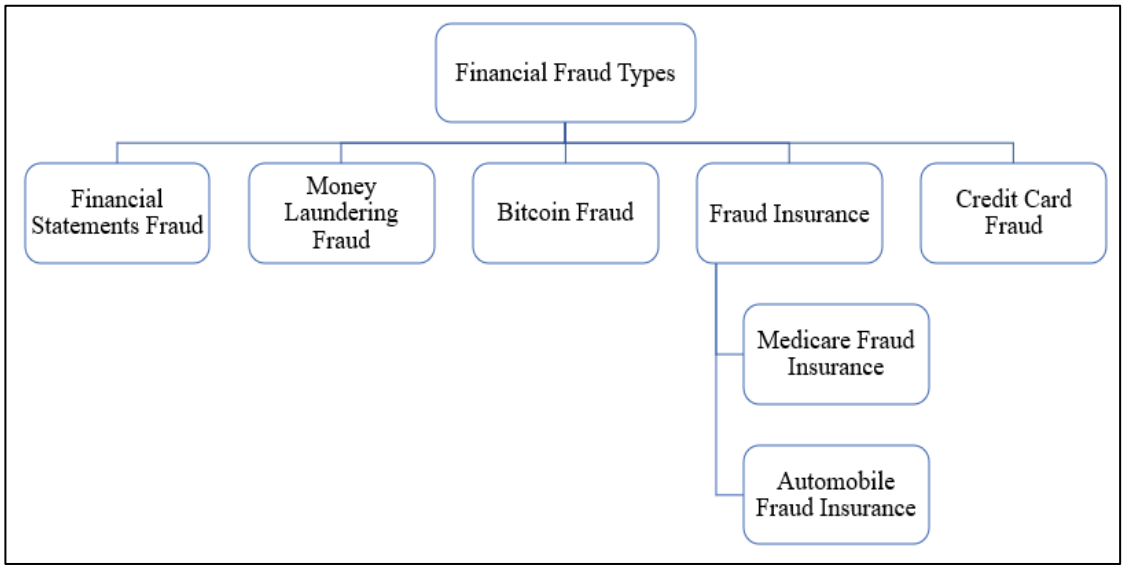


Figure 1.2 Financial fraud types

Credit is a concept used to describe the practice of acquiring and selling products without possessing money. A credit card is a small plastic card that is used to give the consumer the leverage to acquire products and services based on the consumer's promise to pay the money (Raj & Portia, 2011).

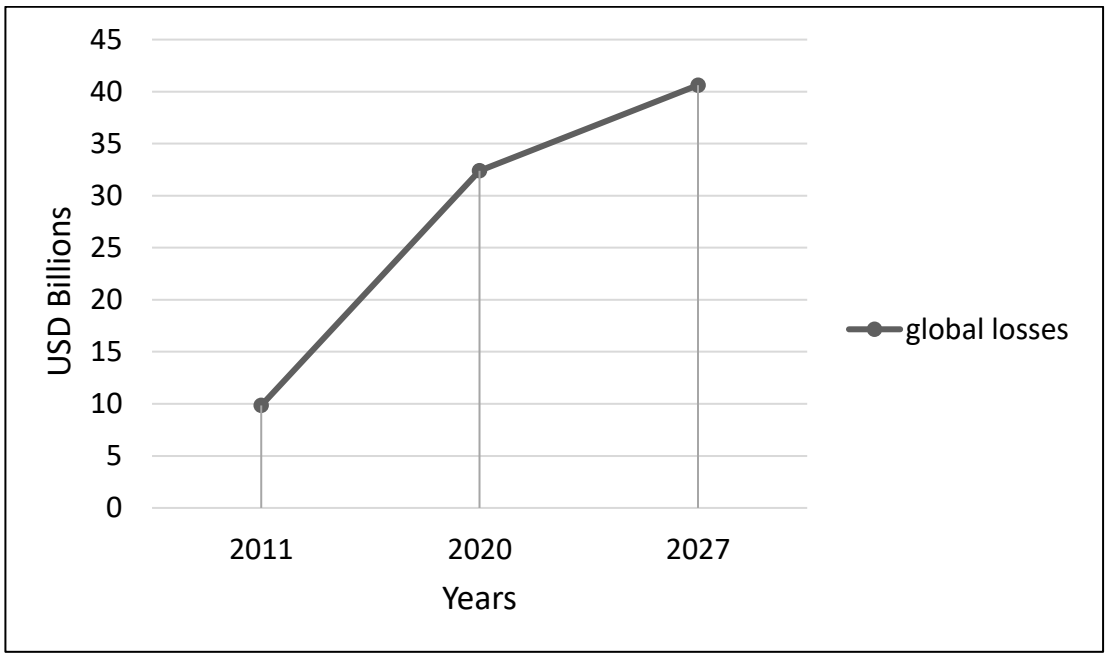
Credit card fraud is a type of identity stealing that involves the fraudulent use of another person's credit card details to charge payments to the account or withdraw funds from it. There are two ways to categorize credit card fraud, behavioral fraud and application fraud, which are also known as online and offline fraud, respectively (Bhattacharyya et al., 2011; Hilal et al., 2022). Application fraud is frequently associated with identity fraud because it typically involves fraudsters attempting to obtain new credit cards from credit companies using other individuals' personal information. Behavioral fraud consists of four different types, including (i) mail theft, (ii) counterfeit cards fraud, (iii) lost/stolen credit card fraud, and (iv) cardholder-not-present fraud (Bhattacharyya et al., 2011; Zhang et al., 2021).

Mail theft fraud occurs when fraudsters intercept individuals' mail to obtain personal banking information or physical credit cards, which they then use to commit application fraud. Lost/stolen credit card fraud happens when fraudsters find or have access to physical cards that have been taken from individuals, either with or without their knowledge. Counterfeit card fraud is committed by fraudsters who create fake credit cards using stolen data. The counterfeit cards are used by the fraudsters to conduct fraudulent transactions, while the victims keep their original cards and use them for legitimate transactions. Counterfeit cards are frequently used only a few times and then abandoned before the victims realize that their credit card information has been stolen.

Cardholder-not-present fraud occurs when transactions are conducted remotely. In this case, only the card information (i.e., holder name, card number, and expiration date) is used to complete the transaction. The difference between counterfeit card fraud and cardholder-not-present fraud is that in the former case, the fraudsters will use the

physical card rather than just the card information (Krawczyk et al., 2014; Zhang et al., 2021).

Nowadays, credit card plays a critical role as a result of the growing industry of online platforms and cashless transactions. Consequently, the number of credit card frauds has drastically increased as fraudsters always find new ways to leverage such events. Since 2011, there has been a dramatic growth in global losses due to payment fraud, escalating from USD 9.84 billion in 2011 to USD 32.39 billion in 2020 (PWC, 2020b). With this growth rate, it is expected that it will eventually be a severe global concern, costing USD 40.62 billion in 2027 as illustratred in Figure 1.3.



(Source : PWC, 2020b)

Figure 1.3 Global losses

Furthermore, there are serious implications associated with undetected credit card fraud, as they have even been exploited for funding terrorism, organized crime, and drug trafficking (Bhattacharyya et al., 2011). To hinder credit card theft and detect such crimes and violations, most banks and financial firms use a rule-based method, in

which an expert will use historical fraud data to define a wide set of rules and static thresholds to flag irregular transactions of rules, and a system will raise an alarm if a new transaction match one of the rules (e.g., a sum greater than USD 10,000) (Kültür & Çağlayan, 2017; Kurshan & Shen, 2021).

The main drawbacks of this manual process encompass expensive operational costs, the lack of flexibility and consistency as well as the fact that it is time-consuming (West & Bhattacharya, 2015). Additionally, it has proven to be unsuccessful as fraudsters quickly discover and circumvent rigid rules (Kurshan & Shen, 2021). Amid these challenges, firms ought to espouse a proactive technology-driven approach to fraud detection, particularly with the new sophisticated criminal techniques that are continuously evolving with technological advancements. The era of technological advancement has aided the financial industry in a better detection of these financial crimes by harnessing the power of ML algorithms that can uncover hidden patterns and, therefore, identify fraudulent financial activities using realistic dataset to simplify decision-making processes. Additionally, it aids in keeping up with the ever-changing sophisticated fraud techniques.

Various studies demonstrated the use of ML algorithms in the credit card fraud detection domain (CCFDD) as it has been the most explored method for fraud detection in the finance sector (Błaszczyński et al., 2021; Fang et al., 2021; Huang et al., 2020; Jain et al., 2016; Krivko, 2010; Kültür & Çağlayan, 2017; Raj & Portia, 2011; Randhawa et al., 2018; Taha & Malebary, 2020). However, due to the severity of the Imbalance Ratio (IR) in the credit card fraud transaction data, conventional ML algorithms are inefficient and exhibit defects of differing severity, particularly when

utilized in the CCFDD (Abdallah et al., 2016; Sahin et al., 2013; Taha & Malebary, 2020; Zhang & Hu, 2014; Zheng et al., 2021).

1.2 Problem Statement

In the real-world, the class distributions are rarely balanced, almost all datasets have a skewed distribution of classes to some extent. These datasets are known as imbalanced datasets. In CCFDD, the dataset tends to be highly imbalanced, which cannot be solved by conventional ML algorithms since these algorithms are based on the assumption of equal class distribution. It has been found that the ML algorithms are highly overwhelmed by the majority class and ignore the most concerned class, minorities, causing inaccurate results (Brownlee, 2020; Kim & Hwang, 2022; Le et al., 2019; Malik et al., 2022). Various approaches have been proposed in the literature to deal with the highly imbalanced datasets problem using resampling techqniue.

Resampling techniques are widely employed in data analysis and ML to address the challenges posed by imbalanced datasets. These techniques aim to alleviate the impact of class imbalance by adjusting the class distribution within the dataset. Nevertheless, most of the studies only focused on a single approach such as resampling techniques at the algorithm-level García et al. (2007), Liang et al. (2022), Mayabadi and Saadatfar (2022), Johnson and Khoshgoftaar (2019), and Krawczyk (2016), resampling techniques at the data-level (RTDL) Hordri et al. (2018), Khaldy and Kambhampati (2018), and Sisodia et al. (2017), Cost-Sensitive Learning (CSL) Cao et al. (2013), Feng et al. (2020), Krawczyk et al. (2014), Sahin et al. (2013), Sun et al. (2007), and Zhang and Hu (2014), and Multiple Algorithms Ensemble (MCE) (Kuncheva, 2014; Roy et al., 2018).

These current approaches have several drawbacks, for instance, resampling techniques at the algorithm-level and MCE work by modifying current algorithms to make them more suitable for imbalanced datasets. However, those techniques need a better understanding of the nature of the used algorithms, as well as the factors that lead to their inability to recognize minority classes (Hu et al., 2016). Furthermore, in RTDL, the optimal class distribution of training data is frequently uncertain. Additionally, an insufficient resampling approach may result in information loss in the majority class when undersampled and overfitting of the minority class when oversampled (Sun et al., 2007).

In addition, the most significant disadvantage of CSL is the lack of understanding of how to set the actual values in the cost matrix, which is rarely understood from data and needs to be provided by an expert. Despite the fact that CSL gives the minority class a higher misclassification cost, it does not provide new information or reduces redundant data for the learning algorithms when used solely (Kotsiantis, Kanellopoulos, & Pintelas, 2006; Roy et al., 2018). Only a few studies have explicitly included the cost of fraud detection in their prediction models (Cao et al., 2013; Ngai et al., 2011). According to Brownlee (2020), Cao et al. (2013), and Hordri et al. (2018) there is no absolute winner among RTDL, Algorithm-Level, MCE, or CSL which remains one of the many challenges with the current approaches.

Most existing fraud detection techniques are based on conventional single ML algorithms, However, fraud detection can pose a challenge for these algorithms for several reasons including, highly imbalanced datasets, continuously evolving data over time (concept drift), lack of real-world datasets due to privacy concerns, overlapping class, noisy data, and misclassification cost issues (Abdallah et al., 2016; Sahin et al.,

2013; Taha & Malebary, 2020). Moreover, conventional single ML algorithms themselves lack shortcomings such as some algorithms having a high ability to overfit, underfit or even provide a very low accuracy regardless of the applied improvements (Awad et al., 2015). Therefore, in an endeavor to overcome the challenges of single ML algorithms, multiple approaches were proposed in the literature with a focus on hybrid models (Carcillo et al., 2021; Jain et al., 2016; Krivko, 2010; Kültür & Çağlayan, 2017; Randhawa et al., 2018; Sarno et al., 2015).

Nevertheless, there appears to be a lack of research on the development of multiple hybrid ML models in the CCFDD. The existing studies mainly focus on the modifications or development of a single hybrid ML model. This gap has not been covered in the literature, implying that not all proposed hybrid ML models are suitable for CCFDD, as researchers proposed a single hybrid ML model utilizing specific ML algorithms without the consideration of other ML algorithms that could have the potential to offer a noticeable improvement to the final hybrid ML model prediction. Thus, due to this gap, a research is required to investigate multiple hybrid ML models hybridization in the CCFDD and conclude a champion hybrid ML model on real-world datasets.

1.3 Research Questions

The following research questions need to be addressed:

i. How to improve the performance of highly imbalanced datasets?

ii. How to enhance the detection of fraudulent activities in the CCFDD?

1.4 Research Objectives

This research aims to develop better detection models for fraudulent activities in the CCFDD. To effectuate the main goal, this research conducts the subsequent objectives, which are:

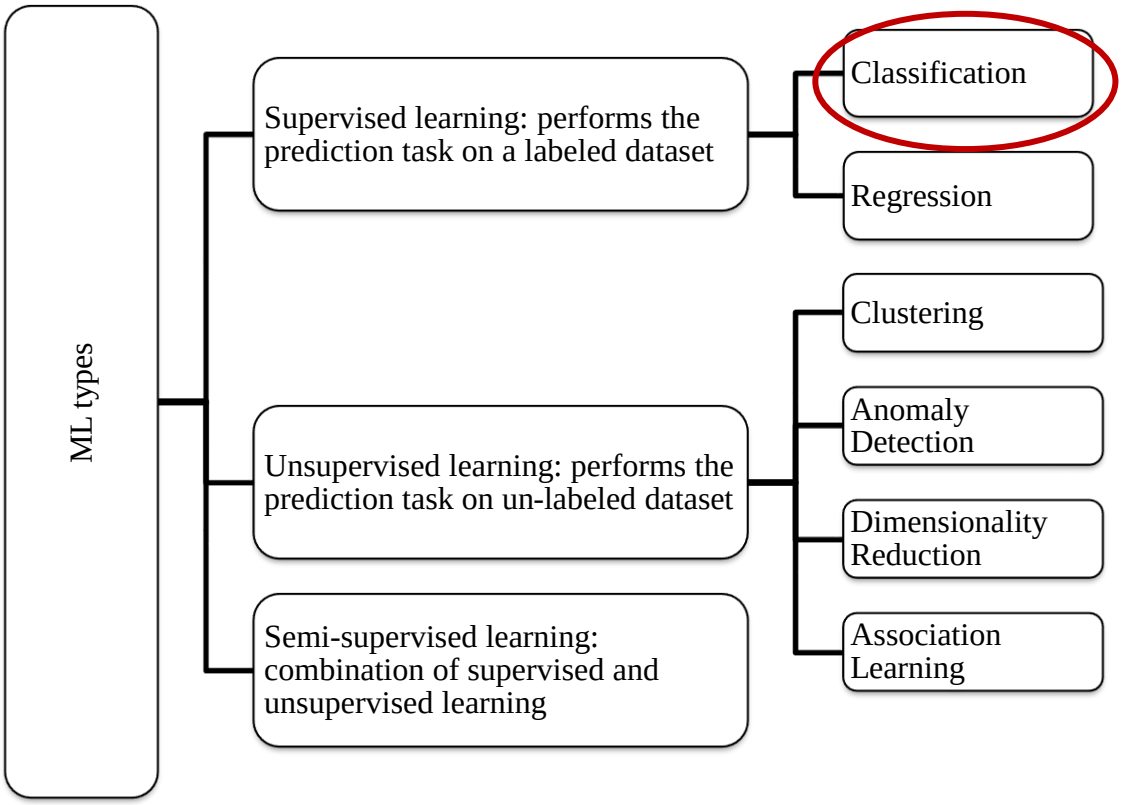
i. To develop a new preprocessing technique using CSL and RTDL to improve the performance of highly imbalanced datasets.
ii. To develop multiple hybrid ML models in order to enhance the detection of fraudulent activities in the CCFDD.

1.5 Scope

As illustrated in Figure 1.4, three major types of ML models were used by researchers and practitioners in the CCFDD (Abdallah et al., 2016; Al-Hashedi & Magalingam, 2021; Carcillo et al., 2021; Carneiro et al., 2017; Kültür & Çağlayan, 2017; Mekterović et al., 2018; Osegi & Jumbo, 2021; Sánchez et al., 2009). Each one is supported by different algorithms, used for specific tasks, and can be employed in a certain type of dataset. In supervised learning, the dataset is labeled into legitimate and fraudulent transactions. While in unsupervised learning, the cardholder's past transactions are used to model the spending behavior of the cardholder. A coming transaction is considered possible fraudulent when it does not match the existing model behavior. On the other hand, semi-supervised learning is a combination of supervised learning and unsupervised learning.

The scope of this research is limited to the classification of fraudulent activities using supervised learning in CCFDD as the nature of most fraud datasets especially credit card datasets is labeled and discrete (e.g., 0 and 1) (Malik et al., 2022).

Furthermore, the CCFDD is presently having the most concerns for both academic and industrial fields as shown in Section 1.1.



(Source: Alpaydin, 2020)

Figure 1.4 ML types

1.5.1 Dataset Characteristics

Understanding the characteristics of the dataset is crucial in selecting appropriate data for credit card fraud detection research. In this section, a description of the differences between balanced and imbalanced datasets and the justification of the selection of the dataset is provided.

A balanced dataset refers to a dataset where the number of instances in each class (fraudulent and non-fraudulent) is roughly equal. This type of dataset enables models to learn from a sufficient number of positive and negative instances, leading to

unbiased performance evaluation. On the other hand, an imbalanced dataset is characterized by a significant disparity in the number of instances between the minority class (fraudulent) and the majority class (non-fraudulent) (Abdallah et al., 2016). Imbalanced datasets are common in credit card fraud detection, as fraudulent transactions are relatively rare compared to legitimate transactions.

The selection of an imbalanced dataset for this research is justified by the need to address the real-world scenario of credit card fraud detection, where the occurrence of fraudulent transactions is low. By focusing on imbalanced datasets, the research aims to develop new preprocessing technique and evaluate the existing ones that specifically tackle the challenges associated with detecting rare fraud cases while maintaining high accuracy in identifying legitimate transactions. This will provide valuable insights into the effectiveness of different approaches and the generalizability of the proposed models in real-world scenarios.

1.6 Significance of Research

This research offers significant contributions to the body of knowledge by attempting to address several gaps. Firstly, the number of highly imbalanced datasets has risen dramatically over the years. Various approaches have been proposed in the literature to deal with the highly imbalanced datasets problem. Nevertheless, most of the studies only focused on a single approach such as resampling techniques at the algorithm-level García et al. (2007), Liang et al. (2022), and Mayabadi and Saadatfar (2022), RTDL Hordri et al. (2018), Khaldy and Kambhampati (2018), and Sisodia et al. (2017), CSL Cao et al. (2013), Feng et al. (2020), Krawczyk et al. (2014), Sahin et al. (2013), Sun et al. (2007) and Zhang and Hu (2014), and MCE (Kuncheva, 2014; Roy et al., 2018). This research develops a new preprocessing technique using CSL and

RTDL to improve the performance of highly imbalanced datasets which is the first contribution of this research. The new technique is known as C-RTDL. This research is one of the first to consider the hybridization of CSL and RTDL, expanding on the limited research on the hybridization of these two common techniques (i.e., CSL and RTDL). The hybridization of CSL and RTDL for imbalanced datasets provides several benefits, including:

- Improved classification performance: Conventional ML algorithms may produce misleading results due to the higher ratio of majority class instances, leading to the poor classification of minority class instances. The developed technique improves classification performance by reducing the impact of the class imbalance problem.

- Reduced misclassification costs: CSL allows for the specification of different misclassification costs for different classes, leading to reduced misclassification costs for the minority class (Brownlee, 2020).

- Avoiding overfitting: Fine-tuning the percentage of CSL ratio and RTDL in the developed technique helps to avoid producing repetitive synthetic data that could cause overfitting.

- Easy to use and implement: The technique is designed to be user-friendly and can be easily implemented in different ML models and domains. This means that researchers and practitioners with limited experience in handling imbalanced datasets can use the technique to improve the performance of their ML algorithms without needing advanced technical skills. The ease of use of the developed technique makes it accessible to a wider audience, thereby

increasing it is potential impact in addressing the problem of imbalanced datasets in CCFDD.

Overall, the hybridization of CSL and RTDL provides an effective solution to the problem of imbalanced datasets and improves the performance ML for imbalanced datasets. Furthermore, over the years single hybrid ML models were considered the most powerful tool in detecting fraudulent transactions. There have been numerous studies in the CCFDD with the idea of hybridization of a single model (Carcillo et al., 2021; Esenogho et al., 2022; Krivko, 2010; Kültür & Çağlayan, 2017; Randhawa et al., 2018; Sarno et al., 2015; Li et al., 2012).

However, there appears to be limited research on the development and the effect of multiple hybrid ML models in the CCFDD. This research proposes multiple hybrid ML models using six ML algorithms, namely, Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), eXtreme Gradient Boosting (XGBoost), and Light Gradient Boosting (LGBM). The performance of these multiple hybrid ML models are investigated, and compared to conventional ML algorithms. As a result, both contributions will help bridge the gap in the literature and pave the way for further analysis and investigation in other domains.

From a practical perspective, it is expected that this research will shed light on the problem of highly imbalanced datasets in the CCFDD and provide a new preprocessing technique CSL and RTDL to help in improving the performance of highly imbalanced datasets regardless of the domain. Additionally, it will generally (not only specific to CCFDD) serve as a basis for developing and exploring multiple hybrid ML models for financial institutions in fraud detection. This research is expected to help decision-makers and practitioners to revisit their current detection programs and

reconsider the use of multiple hybrid ML models in their institutions. Thus, the developed hybrid ML models are expected to aid financial organizations in improving their fraud detection and therefore employ those models in tackling fraudulent activities problems.

1.7 Definition of Key Terms

Algorithm: also named a classifier, it automates the process of ML model that captures the relationship between the descriptive features and the target feature in a dataset (Kelleher et al., 2020).

Class imbalance: a situation in which one class has much fewer samples than the other class (e.g. fraudulent instance than normal instance) (Abdallah et al., 2016).

Cost-sensitive Learning: a type of learning in data mining that takes the misclassification costs (and possibly other types of costs) into consideration (Cao et al., 2013; Ganganwar, 2012).

Credit card: a small plastic card that is used to give the consumer the leverage to acquire products and services based on the consumer's promise to pay the money (Raj & Portia, 2011).

Credit card fraud: when an individual uses another individual's credit card for personal use while the owner of the card, as well as the card issuer, are not aware that the card is being used (Chaudhary et al., 2012; Ramakalyani & Umadevi, 2012).

Data preprocessing: a phase carried out prior to model development to remove noise and modify data to make it suitable for use by ML algorithms (Huang et al., 2020).

Feature selection: the process of selecting a subset of discriminative features for developing strong learning models by removing the redundant and irrelevant features from the data where good performance can be achieved by selecting the right features (Guyon & Elisseeff, 2003; Pes, 2020).

Fraud: a knowing misrepresentation of the truth or concealment of a material fact to induce another to act to his or her detriment (Garner, 2004).

Fraud detection: a set of actions undertaken to detect fraudulent activities that involve money or property gained by deception (Abdallah et al., 2016; Behdad et al., 2012; Ngai et al., 2011).

Fraud prevention: a proactive strategy that is used to prevent deception and fraud events from occurring (Sahin et al., 2013).

Hybrid models: a combination of ML algorithms with each other along with or without optimization techniques to improve the performance of the required tasks (Tsai & Chen, 2010).

Imbalance ratio: the ratio of the sample size of the majority class divided by the minority class (Barua et al., 2012; Zhu et al., 2020).

Machine learning: a subfield of artificial intelligence that allows computers to learn from historical data without being explicitly programmed (Burkov, 2019; Ileberi et al., 2022).

Majority class: the class that contains the major portion of the samples in a dataset (Barua et al., 2012).

Minority class: the class that contains the smallest number of samples in a dataset (Barua et al., 2012).

Overfitting: the model's ability to predict the training data very well but predicts the labels of the data from at least one of the two holdout sets inadequately. It occurs when the algorithm's prediction model is complex that it fits the dataset very closely and becomes sensitive to noise in the data (Kelleher et al., 2020; Burkov, 2019; Huang et al., 2020).

Resampling technique at the data-level: it is used as a preprocessing step to rebalance the dataset or remove the noise before employing ML algorithms (Abdallah et al., 2016; García et al., 2012; Khaldy & Kambhampati, 2018; Roy et al., 2018).

Resampling technique at the algorithm-level: adapts existing ML algorithms to tune them for imbalance dataset problem (Khaldy & Kambhampati, 2018; Roy et al., 2018).

Rule-based method: a method where an expert will use historical fraud data to define a wide set of rules and static thresholds to flag irregular transactions, and a system will raise an alarm if a new transaction match one of the rules (Kültür & Çağlayan, 2017; Kurshan & Shen, 2021).

Semi-supervised learning: a combination of supervised and unsupervised learning where the dataset contains both labeled and un-labeled examples (Burkov, 2019; Alpaydin, 2020).

Supervised learning: general ML learning methods that can exploit training data (i.e., pairs of input data points and the corresponding desired output) to learn an

algorithm that can be used to compute predictions on unseen new data (Aggarwal & Zhai, 2012; Taha & Malebary, 2020).

Training dataset: in ML terms, each row in the dataset is referred to as a training instance, and the overall dataset is referred to as a training dataset (Kelleher et al., 2020).

Underfitting: the model's inability to predict the labels of the data on which it was trained. It occurs when the algorithm's prediction model is too simplistic to represent the underlying relationship in the dataset between the descriptive and target feature (Burkov, 2019).

Unsupervised learning: ML method where algorithms detect hidden patterns in un-labeled transactional data (Alharbi et al., 2022; Zhang & Trubey, 2019).

Validation dataset: a hold-out sampling of the training dataset (Kelleher et al., 2020).

1.8 Organization of The Thesis

This thesis consists of five chapters. This chapter provides a brief background and overview of the fraud problem, it is impact, the domain context, and the different methods used to detect fraud events. Furthermore, this chapter presents the problem statement, research questions, objectives, scope, significance of the research, and definition of key terms.

Chapter 2 provides a comprehensive literature review of previous studies, regarding different aspects associated with this research. In specific, the related works on different approaches for addressing the problem of highly imbalanced datasets and

it is solution, as well as credit card fraud detection using single and hybrid ML models, are critically analyzed and discussed.

Chapter 3 presents detailed experiments procedure and used techniques in this research, mainly in five steps including data collection, data exploration, data cleaning and preprocessing, model development, and evaluation.

Chapter 4 presents the results of the various experiments that were conducted in this research. Additionally, it includes a comparison between the developed new preprocessing technique and the conventional well-known resampling technique. Alongside a comprehensive evaluation of the developed hybrid models in comparison with conventional ML algorithms.

Chapter 5 concludes the work of this research and provides recommendations for future studies and limitations of the current research.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter introduces a comprehensive review of financial fraud detection by ML models in the domain of financial statements fraud, money laundering fraud, bitcoin fraud, fraud insurance, and credit card fraud detection using single and hybrid ML models. This chapter also reviews the problem of highly imbalanced datasets and the existing solutions in the literature. Figure 2.1 shows the literature review organization.

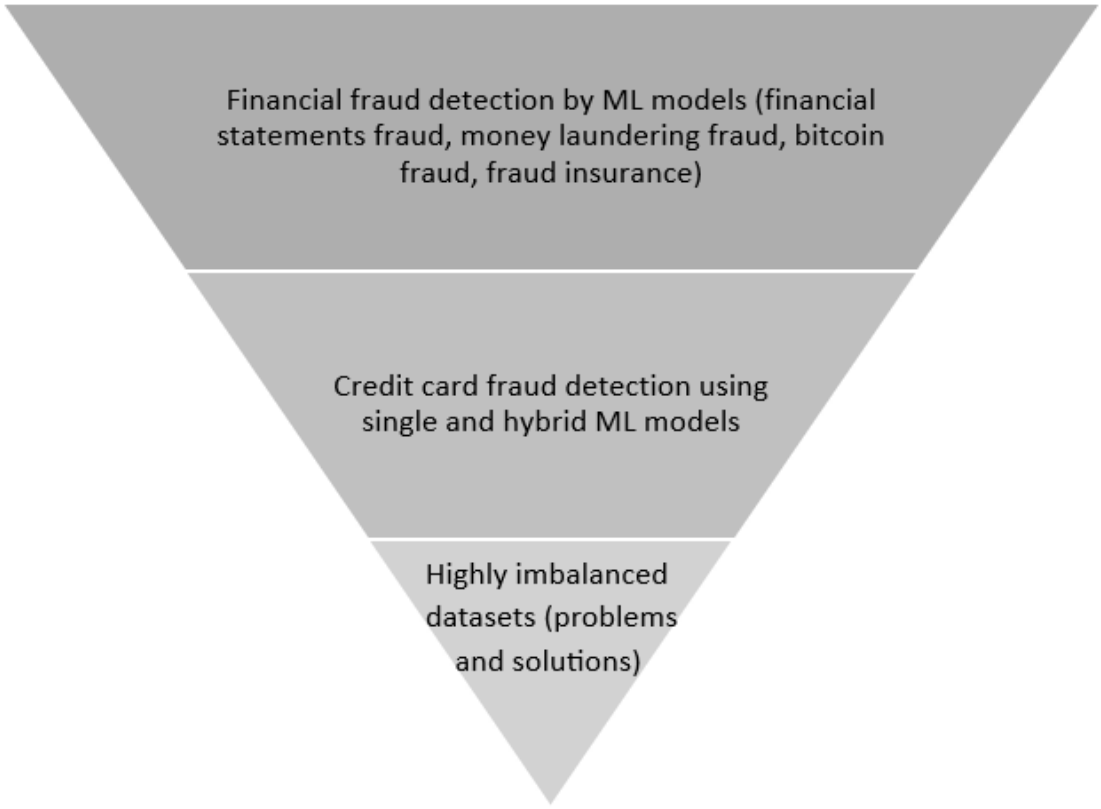


Figure 2.1 Literature review organization

2.2 Financial Fraud Detection Using ML Algorithms

Financial fraud has been intensively studied by researchers and practitioners due to it is importance in diverse critical industries as it has become a sensitive issue recently (Bhattacharyya et al., 2011; Błaszczyński et al., 2021; Carneiro et al., 2017; Hajek & Henriques, 2017; Hooda et al., 2018; Krivko, 2010; Kültür & Çağlayan, 2017; Lin et al., 2015; Randhawa et al., 2018; Sarno et al., 2015; Taha & Malebary, 2020; West & Bhattacharya, 2015). The following section discusses the existing ML algorithms, both single and hybrid models in several fraud areas including financial statement fraud, money laundering fraud, bitcoin fraud, and fraud insurance.

2.2.1 Financial Statements Fraud

Independent auditors oversee the planning and carrying out of audits to ensure that their client's financial statements are free of material misstatements due to fraud (Lin et al., 2003). However, incorrect risk assessment can lead to ineffective or inefficient auditing, which can expose the auditor to regulatory consequences. As a result, many studies have started to use ML algorithms to help auditors assess fraud risk more accurately.

For example, Green and Choi (1997) were the pioneers to use Neural Networks (NN) in fraud detection as they developed an effective NN fraud classification model employing endogenous financial data. Their results support the future use of NNs as fraud risk assessment tools. Green and Choi's findings were replicated by Lin et al. (2015) who explored the discrepancy between expert opinion and empirical outcomes of a prediction model. The used dataset was collected from prosecution and judgment cases against big securities offenses released by the Taiwan Securities and Futures Bureau, in addition to group litigation cases published by the Securities and Futures

Investors Protection Centre between 1998 and 2010. Their results indicated that NN outperforms the other approaches with a classification rate of 91.2%. Moreover, Lin et al. (2003) assessed the risk of fraudulent financial statements by developing an intelligent hybrid model that is a combination between Fuzzy Logic and NN (FNN). In terms of prediction accuracy, the FNN model outperformed the baseline Logit model.

Alternatively, to recognize firms that issue fake financial statements and identify characteristics related to it, Kotsiantis, Koumanakos, Tzelepis and Tampakas (2006) used C4.5, Radial Basis Function (RBF), K2, 3-NN, ripper, and Sequential Minimal Optimization (SMO) as a representative of DT, NN, Bayesian Network (BN), K-nearest Neighbors (KNN), Rule-Learner and SVM, respectively. Thereafter, the authors used a stacking variant approach to combine all these algorithms. The authors suggested that the stacking variant approach outperforms both ensemble and simple methods tested. Additionally, Kirkos et al. (2007) applied NN, BN, and DT using 76 Greek manufacturing companies to differentiate between the fraudulent financial statement (FFS) and non-fraudulent financial statement (non-FFS). Their result indicated that Bayesian Belief Network (BBN) achieved the optimal performance as it manages to properly classify 90.3% of the validation sample in a 10-fold cross-validation method.

Furthermore, the usefulness of the SVM algorithm in detecting FFS was examined by Deng (2009). Deng used a training dataset consisting of 44 FFS and 44 non-FFS from China listed companies between 1999 and 2002. Correspondingly, for the test dataset, 73 FFS and 99 non-FFS from 2003 to 2006 were used. Deng points out that the experimental results correspond with past study findings demonstrating that public financial statement data contains falsification indicators. Using the same dataset as the previously mentioned study, Deng (2010) conducted an investigation into the

development of FFS detection model through the utilization of the Naïve Bayes (NB) algorithm, which involved identifying fraudulent financial statements.

In the same vein, Ravisankar et al. (2011) applied several ML algorithms such as Multilayer Feed Forward Neural Network, SVM, Genetic Programming, Group Method of Data Handling, LR, and Probabilistic Neural Network (PNN) to recognize companies that use FFS. A total of 28 features were selected in the feature selection process using the t-statistic technique. Those algorithms were invoked again using the new subsets. To test the performance of the prediction model, the authors used a dataset with 202 Chinese companies. Their results showed that PNN achieved the highest performance when feature selection techniques were used. However, without feature selection, PNN and Gradient Boosting (GB) outperformed others approximately equally.

Additionally, Hajek and Henriques (2017) conducted a comparative study by employing several ML algorithms such as LR, BBN, NB, DT, NN, SVM, and ensemble methods. The used dataset was collected from various industries, and it consisted of 622 firms, with 311 FFS and 311 non-FFS. Their results indicated that BBN achieves the best prediction performance. More recently, Yao et al. (2018) developed a hybrid model that enhances the detection of financial fraud by integrating feature selection and ML algorithms. In the first phase, the authors applied Principal Component Analysis (PCA) and XGBoost to extract the features, while in the second phase, the fraudulent cases were detected using SVM, DT, RF, LR, and NN. To select the best model, the algorithms were evaluated based on the accuracy and using 120 financial statements that were enclosed by China Securities Regulatory Commission between 2007 and 2016. According to their findings, RF outperformed other algorithms.