Addis Ababa University

College of Natural and Computational Science

School of Information Science

Internal Core System User Fraud Prediction Using Machine Learning: The Case of Commercial Bank of Ethiopia

By:

Abayneh Aklilu (ID- GSE/3323/12)

Advisor:

Michael Melese (Ph.D.)

Completion of the requisites for the Master of Science in Information Science degree.

January, 2024

Addis Abeba

Ethiopia

Addis Ababa University

College of Natural and Computational Science

School of Information Science

Internal Core System User Fraud Prediction Using Machine Learning:

The Case of the Commercial Bank of Ethiopia

By:

Abayneh Aklilu (ID- GSE/3323/12)

Advisor:

Michael Melese (Ph.D.)

Members of the Examining Board

Name	Title	Signature	Date
Michael Melesse (Ph.D.)	**Advisor**	_____	_____
_____	**Examiner**	_____	_____
_____	**Examiner**	_____	_____

January, 2024

Declaration

This thesis has not previously been accepted for any degree and is not being concurrently submitted in candidature for any degree in any university.

I declare that this thesis entitled “Internal Core System User Fraud Prediction Using Machine Learning: In Case of Commercial Bank of Ethiopia” is a result of my investigation, throughout this study, I received guidance and support from my research advisor. I have duly acknowledged external sources through citations, and providing specific references. A comprehensive list of references is included for further clarity.

Signature: _____

Abayneh Aklilu Betemariam

I have given my approval as a university advisor for the submission of this thesis for examination.

Advisor ‘s Signature: _____

Michael Melese (Ph.D.)

Acknowledgment

First, thanks to JESUS and his mother St. Mary for giving me the courage to do this research study. I would like to express my sincere gratitude to my advisor, Dr. **Michael Melese**, for their guidance and support throughout this project. Their expertise and patience were invaluable, and I am grateful for their willingness to go above and beyond to help me succeed.

I would like to thank all CBE core system admins and Security operation admins, particularly for assisting with application log collection. I would not have been able to complete this project without their help.

Finally, I would like to thank my family and friends for their support and encouragement. They have always been there for me, and I am truly grateful for their love and support.

ABSTRACT

Internal user fraud identification is critical to the operation and growth of any business. Identification of fraudulent users can help businesses understand the reasons for internal fraud and plan the bank's strategies accordingly to boost business growth. Internal core system users, such as CBE employees and authorized personnel with privileged access to core banking systems, pose a risk for internal fraud. This type of fraud involves unauthorized access, misuse of privileges for malicious purposes like data theft or unauthorized transactions, and manipulation of internal users through social engineering. Detecting such fraud can be challenging due to insiders' deep understanding of the systems they exploit. The goal of this research is to develop a machine learning model that can accurately predict internal core system users' fraud from the Commercial Bank of Ethiopia Internal core system users. For this study, a total of 7,754 datasets with twelve attributes have been used. To determine the best classifier, the model's overall accuracy was used as the evaluation metric. To this end, supervised machine learning techniques Logistic Regression, Random Forest, Support Vector Machine, and K-Nearest Neighbor were applied to predict internal core system user fraud in the situation of the Commercial Bank of Ethiopia.

Based on the previous literature, those have been widely used classifier algorithms for fraud prediction. In this study, attribute selection was conducted through the utilization of the correlation matrix and feature importance. The process involved identifying the variables that demonstrate the strongest correlation with the outcome variable and those that have the most significant predictive capability. Furthermore, the selected algorithm's efficiency was evaluated and compared after balancing the data using the SMOTE technique. The best overall classifier is Random Forest (RF) with an accuracy of almost 71.63%, a precision of 96.90%, and recalls also 72.96%; then Logistic Regression (LR), with an accuracy of almost 57.96%, a precision of 97.10%, and recall 58.21%, K-Nearest Neighbor (KNN) with an accuracy of almost 47.52%, a precision of 97.63%, and recall 46.80%, and Support Vector Machine (SVM) with an accuracy of almost 54.74%, a precision of 97.04%, and recall also 54.58% respectively.

Keywords: internal core system user fraud prediction, internal core system user fraud, machine learning

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

Abbreviation	Definition

ACFE	Association of Certified Fraud Examiners
AI	Artificial Intelligence
CBE	Commercial Bank of Ethiopia
CRM	Customer Relationship Management
CSM	Customer Service Manager
DSR	Design Science Research
DSRM	Design Science Research Method
EFCC	Economic and Financial Crime Commission
FN	False Negative
FP	False Positive
ICS	Internal Control System
ICSU	Internal Core System Users
IIA	Institute of Internal Auditors
KNN	K-Nearest Neighbor
KPMG	Klynveld Peat Marwick Goerdeler.
LR	Logistic Regression
ML	Machin Learning
RF	Random Forest
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
WEKA	Waikato Environment for Knowledge Analysis

CHAPTER ONE
INTRODUCTION

1.1. BACKGROUND

In today's competitive global business environment change and complexity are common and incorporated into day-to-day business activity [1]. When working in an environment of constant change and increasing complexity, companies should be competitive, productive, and profitable to stay in the business. On the other side in the competitive and complex environment fraud risks are challenged in technological, management control, policies, and procedures issues/gaps.

The modern economy relies heavily on the banking industry, which plays a crucial role in conducting financial transactions and safeguarding the public's assets and trust [2]. With the majority of financial operations now carried out electronically, ensuring the security and integrity of the banking system is of utmost importance. However, digitalization has led to the emergence of new and complex challenges, including the growing risk of financial fraud.

The banking sector faces an ongoing and ever-changing threat from a variety of financial fraud types and it requires implementing different mechanisms and adopting various projects to prevent and mitigate internal fraud risks starting from the front end (business activities) to the back end (IT administrators) for monitoring and controlling [3]. Instances of fraud include but are not limited to, insider fraud, identity theft, cyberattacks, and account takeovers. These fraudulent actions lead to significant monetary losses, and result in trust deterioration, reputation damage, and concerns regarding privacy and security of data. Researchers proposed several fraud prevention techniques and a range of strategies consisting of fraud policies, cellphone hotlines, worker reference checks, fraud vulnerability reviews, seller contract critiques and sanctions, analytical critiques (financial ratio analysis), password protection, firewalls, digital evaluation, and other forms of software technology, and discovery sampling [4].

The banking sector encompasses elements, approaches, and entities contributing to the efficiency of financial markets and facilitating accessibility to capital and financial services [5]. The significance of enhancing the banking sector to stimulate monetary growth is crucial, as it fosters economic development, nurtures private sectors, increases liquidity to harness local savings, fosters competition among banks, and enhances the diversity of financial institutions [6]. Fraud is a major

difficulty for the whole banking industry [7]. The public, relying on banking services, anticipates operational accountability, fairness, and transparency on a daily basis for efficient intermediation facilitated by robust internal controls. This indicates that the positive contribution of internal control systems is crucial for the development, productivity, and long-term viability of financial institutions. [5].

Fraud is an unethical practice that gathers significant attention from regulators, auditors, and the public, driven by the escalating instances of corporate failures and the misuse of corporate reputation [8]. The majority of fraudulent activities occur within organizations rather than external transactions. Contrary to popular belief, 68% of fraud cases take place internally, involving both employers and employees, while the remaining cases involve external individuals within the value chain [8].

Fraud has persisted across historical periods, manifesting in various forms. The evolution of the banking sector, coupled with technological advancements and the widespread adoption of the Internet, has contributed to the expansion of bank fraud [9].

Employees who are users of the Internal Core System (ICSU) within the bank possess understanding of the systems and have access to classified and confidential information. This, coupled with technological advancements, creates a potential scenario where these individuals could engage in fraudulent activities [10]. With the introduction of external pressures and rationalization, they may become involved in fraud cartels, contributing to the misappropriation of millions from the banks.

In the banking industry one of the main concerns today is internal fraud. Within the Bank domain, internal fraud refers to fraudulent activities or misconduct carried out by individuals who work in a bank or financial institution. Typically, these individuals are employees, managers, or other insiders who use their positions and knowledge of the bank's systems, processes, and controls to commit fraudulent acts for personal gain [10]. As a result, CBE may experience a potential loss of income.

For a bank to progress its business strategy, it is crucial to have a clear understanding of its Internal Core System Users (ICSU). Hence, it is essential to identify those ICSUs who are involved in fraudulent activities, as it not only helps in safeguarding the bank and its users but also assists in

gathering business intelligence [11]. To address this issue, businesses have started using predictive modeling techniques to aid in detecting these Internal Core System Users.

Fraud detection relies on traditional methods to pinpoint potential fraud risks [12], but using rules-based systems can pose challenges due to their inflexibility and tendency to generate false alarms, which can undermine trust in the system. Moreover, users may grow frustrated with repeated inconveniences or delays, and a focus on minimizing false alarms can sometimes exacerbate the issue. In contrast, artificial intelligence (AI) provides a promising solution for predicting fraudulent activities within the ICSU (Internal Control Systems and User) domain. AI possesses the capability to analyze increasingly intricate data over time and uncover relationships between various ICSU behaviors that might elude human perception [13]. By leveraging AI, a diverse array of data sources can be harnessed to examine complex behavioral patterns and evaluate individual risk levels based on historical activities.

Machine learning is a subfield of artificial intelligence that enables machines to learn from the data they are trained on. In the past decade, various machine-learning techniques have been utilized for predicting fraud [14]. Machine learning algorithms utilize data analytics and predictive modeling to unveil hidden patterns and anomalies. Banks can utilize this technology to proactively identify and mitigate internal fraud risks, thereby strengthening their defenses against insider threats.

Researchers are exploring advanced methods to detect fraudulent employees at an early stage. In the past, various machine learning techniques such as Random Forest, Decision tree, Bagging, and so forth have been utilized to predict fraudulent ICSUs. Over the last two decades, machine learning has become a fundamental aspect of information technology and an integral part of our daily lives, though often unnoticed. As data continues to grow, it is likely that intelligent data analysis will become even more crucial for technological progress [15].

Machine learning, as its essence, denotes computer programs that can learn independently, without the need for human intervention. This concept originated from Alan Turing's 1950 paper "Computing Machinery and Intelligence," which introduced the concept of a "Learning Machine" that could deceive a human into believing it was authentic [16]. Currently, machine learning is a broadly employed concept that covers a range of programs used in big data analytics and data mining. Ultimately, machine learning algorithms are the driving force behind many predictive programs, such as spam filters, product recommenders, and fraud detectors.

In machine learning, the learning process involves extracting knowledge from data without the need for human intervention [17]. By building complex concepts from simpler ones, the learning process effectively reduces complexity. Machine learning has made significant progress in addressing challenges that the artificial intelligence community has been attempting to resolve for an extended period of time.

Running a business using traditional approaches can be challenging, but leveraging machine learning methods can offer significant advantages, especially when creating a fraud detection model tailored to the data and specific characteristics of the problem. This facilitates the recognition of fraudulent patterns within ICSUs [18].

Despite the significance and urgency of the matter, ICSU has notably lacked research dedicated to machine learning algorithms for fraud prediction within the banking industry. Existing research mainly concentrates on developing generic fraud detection models or addressing external threats, thereby ignoring the unique challenges and peculiarities presented by insiders.

This research begins on a mission to address a pressing concern within the banking sector, namely the need for more advanced fraud prediction methods targeted at internal core system users. By binding the capabilities of machine learning algorithms, efforts to enhance security, protect customer assets, and strengthen the trust and integrity that are the lifeblood of the banking industry in today's digital age. A company's ability to prevent system breaches and maintain the stability and security of the financial system is gauged by the implementation of a strong internal control system. Consequently, the use of machine learning algorithms is essential for analyzing datasets, identifying patterns, and making predictions.

1.2. MOTIVATION OF THE STUDY

The finance sector plays a critical role in modern economies, ensuring trust and stability. However, in today's digital world, the industry is facing a growing threat fraud. Fraudulent activities, such as insider fraud and cyberattacks, can lead to significant financial losses and challenge people's trust in financial institutions. Consequently, there is an urgent requirement for innovative solutions to safeguard the integrity and security of the banking system. Preventing and detecting fraud committed by internal core system users is one of the most complex aspects of this issue. Employees of banks and authorized personnel who have access to internal systems hold a unique position

within the institution. They are responsible for managing daily operations, but their knowledge and access also make them potential insider threats. Since these individuals can manipulate internal systems, it is especially difficult to detect their activities. As banking goes digital, fraudsters have more opportunities. Preventing internal fraud in the industry now requires proactive and adaptable measures using the state-of-the-art technology rather than a traditional method.

The motivation behind this study is rooted in the imperative need to protect the banking industry against the threats of internal fraud. By Utilizing the potential of machine learning, the desire to equip financial institutions with proactive tools to preserve their integrity, security, and reputation become vital.

1.3. PROBLEM STATEMENT

The banking industry is now growing fast and gaining more and more positions in the country's economy of Ethiopia [19]. However, by establishing more banking services, the chances and rate of fraud occurring in the banking industry would also increase. Fraud is a universal problem that affects every business, industry, and organization [20]. It leads to losses that continue to cause a significant problem for industries, even though significant progress is made in fraud prevention and detection technologies [21].

The global financial system relies significantly on the banking sector, a crucial player in facilitating monetary transactions and ensuring the security of assets and trust for millions of individuals and businesses. In today's digital age, ensuring the safety and reliability of internal core system users, which are critical for core banking operations, is of utmost importance [22]. Nonetheless, the threat of fraudulent activities carried out by insiders, which is widespread and ever-changing, poses a significant challenge to the industry's reputation and stability.

CBE employees and authorized personnel who have privileged access to core banking systems are considered as internal core system users. They have an advantage and can use their expertise and access to exploit systems for illegal purposes. Internal fraud is a type of fraudulent activity that includes:

- Unauthorized Access,

- Identifying employees or authorized users who misuse their privileges for malicious purposes, such as stealing sensitive data, intellectual property, or conducting unauthorized transactions,
- Recognizing patterns of behavior that may indicate fraudulent use of someone else's identity to gain access to systems or information.
- Predicting and preventing unauthorized access or exfiltration of sensitive data from internal systems, such as customer records, financial data, or intellectual property.
- Identifying attempts by fraudsters to manipulate internal users into disclosing sensitive information or performing unauthorized actions through social engineering techniques.
- Recognizing patterns associated with account takeover attempts, where fraudsters gain control of legitimate user accounts to carry out fraudulent activities.

which can be hard to detect by traditional methods, as insiders have an in-depth knowledge of the systems, they control [23].

Refined and flexible fraud prediction solutions powered by machine learning algorithms are urgently required to tackle this critical issue [24]. These algorithms possess the ability to analyze large datasets, detect subtle patterns, and forecast fraudulent behaviors among CBE internal core system users with greater accuracy and speed than conventional rule-based systems.

The banking industry faces several challenges when it comes to adopting and integrating machine learning algorithms for internal core system users' fraud prediction, despite the potential benefits [25]. Addressing challenges such as feature identification, data quality maintenance, privacy concerns, model explainability, regulatory compliance, and adapting to evolving fraud tactics is crucial. Insider fraud prevention and detection are primary concerns for CBE, necessitating tailored machine learning models.

The research aims to identify the key feature(s) that significantly influence the prediction of fraud among internal core system users. By analyzing various user-related variables, the study seeks to determine which specific feature(s) play the most pivotal role in accurately predicting instances of fraud within the system. Understanding the most determinant feature(s) will not only enhance the effectiveness of fraud detection mechanisms but also contribute to the development of more targeted and efficient fraud prevention strategies.

In the context of internal core system users' fraud detection, the effectiveness of various machine learning algorithms remains uncertain. Despite advancements in machine learning techniques, it remains unclear which algorithm is most suitable for accurately predicting instances of fraud perpetrated by internal core system users. This knowledge gap hinders the development of robust fraud detection systems tailored to the unique characteristics of internal core systems. Therefore, there is a pressing need to systematically evaluate and compare different machine learning algorithms to determine which one(s) demonstrate superior performance in predicting fraud among internal core system users.

The problem at hand affects to evaluating the effectiveness of a proposed machine learning model in predicting instances of fraud perpetrated by internal core system users within a financial institution, referred to hereafter as CBE. Given the rising concerns surrounding insider fraud and the critical importance of early detection and prevention, it becomes imperative to assess the predictive capabilities of the suggested machine learning approach. Thus, the research endeavors to quantify the extent to which the proposed model accurately identifies fraudulent activities within CBE's internal systems, with the ultimate aim of enhancing fraud detection mechanisms and fortifying the institution's resilience against potential threats posed by insider fraud.

This study aimed to bridge existing gaps by leveraging machine learning for predictive modeling. Solving these issues will enhance CBE system reliability and safety, safeguard client financial welfare, and uphold global economic stability objectives. In the attempt to solve the above stated problems, the following research questions are formulated to be answered by this research work;

RQ1. What is the most determinant feature to predict fraud for internal core system users?

RQ2. Which machine learning algorithm is suitable to the prediction of internal core system users' fraud?

RQ3. To what extent the suggested machine learning model predict fraud among internal core system users within CBE?

1.4. OBJECTIVES OF STUDY

1.4.1. GENERAL OBJECTIVE

The general objective to identify and develop the most effective model capable of predicting fraudulent internal core system users

1.4.2. Specific Objective

- To identify the single most influential feature among a set of variables in predicting fraud among internal core system users, thereby providing insights into the primary indicator of fraudulent activity within the system.
- To systematically evaluate and compare the performance of various machine learning algorithms in predicting fraud among internal core system users, with the aim of identifying the most effective algorithm for accurately detecting fraudulent activities within the system.
- To identify the contributing factors of internal core system users' fraud behavior in banking by applying machine learning.
- To evaluate the accuracy, effectiveness, and reliability of the suggested machine learning model in predicting fraudulent activities perpetrated by internal core system users within CBE, thereby assessing its practical applicability and potential for enhancing fraud detection and prevention strategies.
- To apply the selected tools and machine learning techniques to build models for better capabilities and improved predicting performance.
- To design the general architecture of internal core system users’ fraud prediction.
- To evaluate the machine learning model and compare performance with related works and supervised machine learning algorithms to predict the internal core system user's fraud.
- To offer concluding insights and suggest avenues for future research endeavors in this particular field.

1.5. Significance of study

This research aims to predict internal core system users' fraud. The model is intended to aid in:

- By identifying the single most influential feature among a set of variables in predicting fraud among internal core system users, the research can significantly enhance fraud detection

capabilities. Understanding the primary indicator of fraudulent activity allows for more targeted monitoring and detection efforts.

- Tightening the logical control/access control by amending the user matrix based on their profile.
- Tightening the logical control/access control by amending the user matrix based on their profile.
- Improving the administrative policy. For instance:
    - The models can provide transparency into the decision-making process, helping CBE demonstrate compliance to regulators and auditors.
    - Utilizing machine learning can aid in monitoring and documenting compliance, ultimately minimizing the possibility of breaching regulatory requirements.
    - Machine learning models provide alerts and findings that allow CBE to adjust policies in real-time.
    - Machine learning models can be used to improve policies using identified weakness and vulnerabilities.
    - Use of machine learning model allows for a continuous feedback loop.
- To apply/take reactive measures to the fraudulent users instead of proactive measures.

The model strengthens the internal control system and getting a positive and significant impact on fraud detection and takes proactive measures in the Commercial Bank of Ethiopia.

Finally, this study will be of great help to anyone wishing to conduct further research on fraud prevention in Ethiopian commercial banks or other related fields.

1.6. Scope and limitation of the study

The scope of this study is to design and develop a reliable machine-learning model that accurately predicts fraudulent activities conducted by internal core system users within CBE. To achieve this, the study will analyze historical user data, including transactional patterns, access logs, and behavioral attributes, to pinpoint potential indicators of fraudulent actions. The research will consider several machine learning algorithms, such as logistic regression, random forest, k-nearest neighbor, and support vector machine, to establish a predictive framework that can detect unusual activities in real-time.

The study's main constraint is that it only focused on the Addis Ababa kirkos district of CBE through the simple random sampling technique. Although this technique is an easy way to obtain sample data for CBE internal core system users, it may not be representative of other CBE districts. Therefore, the user may vary from one district to another. Another limitation of this study is that it did not include essential features like users' mobility. This is because these features have not been incorporated into the CBE database.

1.7. Organization of the study

The thesis is structured into five chapters, with Chapter One serving as an introduction that outlines the motivation, problem statement, objectives, research questions, significance of the study, and the scope and limitations. In Chapter Two, the extensive literature review and related works are the main topics of discussion. Previous studies in the banking sector and internal core system user fraud of CBE are examined, with a particular emphasis on the use of machine learning applications to predict internal core system user fraud. In Chapter Three, the research methodology, the design and process part are examined. The Peffers et al. design science methodology is followed, and a detailed approach to solving the study problem is explained. Chapter Four, titled Experiment, Result, and Discussion, specifies the implementation in detail, results, and subsequent discussions. It delves into model selection, justification, hypothesis evaluation, and identification of the best model. The chapter concludes by analyzing the research problem based on the outcomes, assessing the hypothesis, and acknowledging the strengths and limitations of the thesis. Finally, Chapter Five, Conclusion and Recommendation, synthesizes the research, outlines its contribution to the research question, and suggests avenues for future research in a related domain.

CHAPTER TWO
LITERATURE REVIEW

2.1. Overview

This chapter presents an overview of the critical concepts related to fraud, fraud committed by internal core system users, fraud prediction approach, the banking sector, and machine learning algorithms that help banks in detecting such frauds at an early stage. It also provides a review of several relevant papers to facilitate the continued exploration of the research issue and explains how machine learning algorithms identify internal core system users' fraud before it occurs. The chapter concludes by outlining the gaps in the current research body and setting forth the study's objectives.

2.2. Fraud

Fraud, in its essence, involves deliberate deception or misrepresentation for personal gain or to cause harm to others. It can manifest in various forms, from financial fraud where individuals or organizations manipulate financial records or transactions for illicit gains, to identity theft where someone assumes another person's identity for fraudulent purposes [23]. Regardless of the method, fraud undermines trust and integrity in both personal and professional interactions, and it can have far-reaching consequences for victims and society as a whole.

One of the key elements of fraud is deception. Perpetrators often employ tactics to conceal their true intentions or to mislead others into believing false information. This deception can take many forms, such as falsifying documents, providing misleading information, or using sophisticated techniques like phishing emails or social engineering to trick individuals into divulging sensitive information [24].

Another important aspect of fraud is the element of harm caused to victims. Whether it's financial losses, damage to reputation, or emotional distress, fraud can have devastating effects on individuals, businesses, and even entire communities. Victims of fraud may experience financial ruin, struggle to recover their stolen identities, or face challenges in rebuilding trust with others.

Moreover, fraud is not limited to specific industries or sectors. It can occur in various contexts, including healthcare, insurance, banking, e-commerce, and even within governmental institutions. As technology advances, new avenues for fraudulent activities emerge, such as cyber fraud and cryptocurrency scams, posing additional challenges for law enforcement agencies and regulators.

Preventing and combating fraud requires a multi-faceted approach involving education, awareness, enforcement, and technological innovation. Individuals and organizations must remain vigilant, employing best practices for securing sensitive information, verifying the legitimacy of transactions, and staying informed about emerging fraud schemes. Additionally, collaboration between the public and private sectors is essential to develop effective strategies and policies to detect, investigate, and prosecute fraudulent activities.

2.3. Fraud prediction

Fraud prediction encompasses a range of methodologies aimed at identifying and mitigating fraudulent activities before they occur. One approach involves the use of statistical modeling and machine learning algorithms to analyze historical data and detect patterns indicative of fraudulent behavior [24]. These models can be trained on past instances of fraud, learning to recognize common characteristics and anomalies associated with fraudulent transactions or activities.

Another approach is the use of anomaly detection techniques, which focus on identifying deviations from normal behavior within a dataset [25]. By establishing a baseline of typical behavior, anomalies that may indicate fraudulent activity can be flagged for further investigation. This approach is particularly effective in detecting previously unseen or evolving forms of fraud.

Furthermore, predictive analytics techniques, such as decision trees, neural networks, and ensemble methods, are commonly employed to forecast the likelihood of fraudulent events based on a combination of variables and features. By leveraging large volumes of data and sophisticated algorithms, these models can provide insights into potential fraud risks and help prioritize resources for prevention and intervention efforts.

Moreover, the integration of advanced technologies like artificial intelligence, natural language processing, and network analysis enables more nuanced and real-time fraud detection capabilities. These technologies empower organizations to analyze unstructured data sources, detect subtle patterns in human behavior, and uncover complex fraud schemes that may span multiple channels or domains.

Overall, fraud prediction approaches leverage a combination of statistical, machine learning, and domain-specific techniques to anticipate and prevent fraudulent activities proactively. By continuously refining and adapting these methodologies to evolving fraud threats, organizations can stay ahead of fraudsters and safeguard their assets, reputation, and stakeholders' trust.

2.4. Internal Users' Fraud

In the realm of cybersecurity, while external threats are often the focus, there is a more concealed danger that exists within organizations from internal user fraud [24]. This deceptive threat comprises of actions carried out by individuals who have been granted authorized access such as employees or trusted associates. These individuals manipulate their positions for malicious purposes. The spectrum of internal user fraud activities includes financial malpractice, data theft, unauthorized access, and unintentional negligence [25].

Financial fraud perpetrated by internal actors can take many forms such as embezzlement, account manipulation, or fraudulent transactions leading to significant financial losses for organizations [26]. The theft of sensitive data and intellectual property is another aspect where disgruntled employees may compromise valuable information or even sabotage systems. Unauthorized access and misuse of credentials pose a significant risk, potentially resulting in data exposure or compromise.

Internal users or insiders are individuals who possess both authorized and legitimate access to an organization's resources, including but not limited to its corporate networks, applications, and data [27]. In the field of information security behavioral research, the concept of "internal" typically pertains to individuals who possess authorization to access corporate information systems and possess knowledge regarding the organization's processes.

The three components that make up internal fraud are the "fraud triangle": justification for the deception, perceived pressure, and perceived opportunity. There are three components to every act of fraud, regardless of whether it is committed on behalf of or against a business. As we see in figure 2.1 the three elements in the fraud triangle are interactive, for instance, the greater the perceived opportunity or the more intense the pressure, the less rationalization it takes for someone to commit fraud [28].

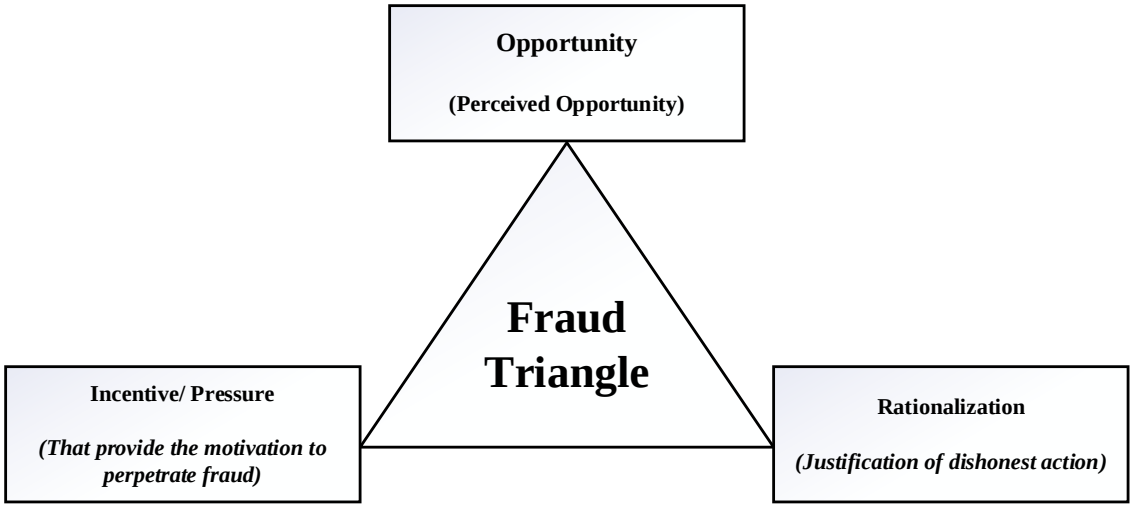


FIGURE 2.1. FRAUD TRIANGLE [28]

Instances of internal user fraud may transpire during periods of financial strain, where employees, feeling the pressure of economic challenges, may succumb to the temptation of financial malfeasance. Disgruntlement and dissatisfaction with work conditions or interpersonal issues can also contribute, as employees may resort to acts of sabotage or data theft as a form of retaliation. Additionally, lax or inadequate internal controls, such as weak access management and monitoring systems, create opportunities for fraud to go undetected. Social engineering tactics, including phishing attacks, can exploit human vulnerabilities, leading unwitting employees to inadvertently participate in fraudulent activities. In essence, internal user fraud often emerges when a combination of personal, organizational, and technical vulnerabilities converges, highlighting the need for comprehensive measures to mitigate these risks effectively [28].

Fraud committed by internal users can manifest in a variety of ways, often due to a combination of factors within an organization. Instances of internal user fraud may arise during times of financial hardship, where employees, feeling the pressure of economic challenges, may be tempted to engage in financial misconduct. Employees may also resort to acts of sabotage or data theft as a form of retaliation when they are unhappy with their work conditions or have interpersonal issues. Inadequate internal controls, such as weak access management and monitoring systems, can also create opportunities for fraud to go undetected. Additionally, social engineering tactics, including phishing attacks, can exploit human vulnerabilities, leading unsuspecting employees to participate in fraudulent activities. Essentially, internal user fraud often occurs when personal, organizational, and technical vulnerabilities intersect, emphasizing the necessity of all-encompassing actions to effectively mitigate these risks [29].

2.4.1. Internal Core system users' fraud prediction

Organizations are constantly threatened by internal fraud committed by their own employees in the modern digital landscape [30]. A significant risk is posed by internal core system users, who have authorized access to sensitive financial and operational data, as they can manipulate or misappropriate resources without being detected. As a result, predicting the likelihood of internal fraud among these users is now a critical component of enterprise risk management.

There are various methods that can be utilized to predict internal fraud among internal core system users. Traditional techniques rely on statistical analysis of historical fraud data to identify correlations, and patterns It could point to possible fraud [31]. However, these methods may fail to capture the complexities of individual user behavior and the intricate interactions within an organizational environment.

The emergence of artificial intelligence (AI) and machine learning has presented new opportunities for fraud prediction. AI algorithms can analyze large amounts of data, including transaction logs, user access patterns, and anomalies in behavior to detect subtle indicators of fraudulent intent. These algorithms can continually learn and adapt, improving their accuracy as more data becomes available [32].

Real-time insights into user activity can be provided by behavioral monitoring systems, which can alert security personnel to suspicious transactions or deviations from normal behavior, in addition to predictive analytics. When predictive modeling is combined with real-time monitoring, organizations can establish a strong defense against internal fraud [32].

It's critical for organizations to predict internal core system users' fraud if they want to strengthen their cybersecurity defenses. By using advanced analytics and machine learning algorithms, businesses can be proactive in identifying patterns and anomalies in core system user behavior. Fraudulent activities can be detected by looking for unusual access patterns, typical transaction volumes, or unexpected data queries. Machine learning models, trained on historical data, are key in detecting deviations from the established norms, allowing organizations to predict potential fraud before it becomes a bigger issue [33]. This predictive approach enables security teams to take preemptive measures, including tightening access controls, launching investigations, or introducing additional security measures. As the cyber threat landscape evolves, the ability to predict and prevent internal core system users' fraud becomes a

cornerstone of safeguarding sensitive information and maintaining the integrity of an organization's core operations.

2.4.2. Internal fraud prediction approaches

Predicting internal fraud necessitates tailored approaches due to the intricacies of insider threats, involving individuals within an organization. Unlike external fraud, internal fraud involves privileged access, making detection challenging. Techniques like user behavior analytics and anomaly detection analyze deviations from normal behavior, enabling proactive intervention. Advanced monitoring systems like network traffic analysis and endpoint detection surveil for suspicious activities. Integration of data analytics and machine learning enhances predictive capabilities by identifying hidden patterns in various data sources. Fostering a culture of transparency and accountability, alongside encouraging reporting of suspicious activities, complements these technical measures. In essence, a proactive, multi-faceted approach combining behavioral analysis, advanced monitoring, data analytics, and organizational culture strengthens defenses against internal fraud.

2.5. Banking Industry

The banking industry is a fundamental part of the global economy, playing a diverse role in enabling financial transactions, advancing economic growth, and providing a range of financial services to both individuals and businesses. With a complicated network of establishments, including conventional banks and internet-based financial platforms, this industry serves as a vital intermediary between lenders and borrowers. As a custodian of funds, banks offer a wide range of services such as savings accounts, loans, investment products, and electronic payment systems. The banking sector has undergone a new era with the constant evolution of technology, leading to digital innovations that transform how customers interact with financial institutions, resulting in online banking, mobile payments, and advanced data analytics [34].

Over the past few years, the banking sector has been confronted with a constantly changing landscape characterized by the introduction of new technology, regulatory modifications, and evolving customer expectations. The regulatory frameworks, which were put in place to ensure stability and safeguard consumers, have undergone significant changes following the global financial crisis. To maintain their resilience in the ever-evolving environment, banks are adapting to new compliance regulations and enhancing their risk management practices [35]. Additionally, the emergence of fintech companies has

promoted cooperation and competition in the sector, compelling traditional banks to adopt digital transformation and provide innovative solutions to keep up with the changing demands of their customers.

As technology advances, regulations change, and the global economy grows, the banking industry is set to transform its future trajectory. Banks must find a way to balance tradition and innovation, while also prioritizing digitalization, reinforcing cybersecurity protocols, and building trust with their customers [36]. All of these factors will play a crucial role in determining the industry's place in the financial landscape of the future.

2.5.1. The Banking Industry in Ethiopia

Ethiopia's economy heavily depends on the banking industry, which plays a critical role in promoting financial inclusion and supporting economic development. Over the years, the industry has undergone significant changes, marked by the introduction of innovative financial services and a greater emphasis on extending banking services to different demographic groups [37]. The banking industry in Ethiopia is comprised of both public and private banks, with the Ethiopia's National Bank serving as the regulatory body responsible for ensuring monetary policy and maintaining the stability of the financial system.

In recent times, Ethiopia has witnessed a gradual shift towards digital banking, with the adoption of mobile banking services and the establishment of online platforms. The primary objective of this digital transformation is to increase financial accessibility for a larger portion of the population, especially those in rural areas. As the Ethiopian government continues to prioritize economic growth and financial inclusion, the banking sector is expected to play a crucial role in channeling funds to critical sectors, supporting entrepreneurship, and contributing to the overall stability and development of the nation's economy [38].

Ethiopia's banking sector encounters several difficulties, which include the requirement for additional infrastructure development, regulatory modifications, and resolving problems connected to non-performing loans. In order to ensure the continued growth and prosperity of the industry, it is essential to navigate these challenges effectively while taking advantage of the opportunities presented by a booming economy [39].

2.6. Internal Core system Fraud Prediction in the Banking Industry

In today's ever-changing banking landscape, the use of advanced technology is paramount. Therefore, predicting and preventing internal core system fraud has become a crucial focus for financial institutions. These internal core systems are the backbone of a bank's operations and include essential functions such as transaction processing, customer data management, and financial record-keeping. Because of the complexity and sensitivity of these systems, there is an urgent need for a proactive approach to detect potentially fraudulent activities [40]. Banks can use sophisticated fraud prediction models powered by artificial intelligence and utilizing machine learning to examine or analyze patterns in their core systems, detect anomalies, and prevent potential financial and reputational risks.

The implementation of robust fraud prediction mechanisms enhances the security posture of banks and contributes to regulatory compliance. As financial regulations continue to evolve, the ability to forecast and preemptively address internal core system fraud is according to the sector's commitment to safeguarding customer assets and maintaining the integrity of financial transactions [41]. Furthermore, these predictive analytics systems are crucial in fostering customer trust by ensuring the confidentiality and reliability of core banking operations. Striking a balance between innovation and security is key as banks navigate the intricate landscape of digital finance, and robust fraud prediction measures stand as a cornerstone in achieving this equilibrium [41].

The objective of internal core system user fraud prediction is to develop and implement systems, tools, and processes that can accurately identify and predict instances of fraudulent activities perpetrated by users within the organization's core systems, assuming the data related to each internal core system user in the transactions. The internal core system user fraud prediction issue is ordinarily portrayed in three main stages, namely, the training stage, test stage, and prediction stage [37]. In the training phase, the contribution to the internal core system user fraud problem is from the historical data such as personal and/or business internal core system users' data, which has been gotten and retained by the CBE's service providers. Moreover, in the training stage, the labels are structured in the list of internal core system users' records. In the test stage, the prepared model with the highest accuracy is tested to predict the internal core system users' fraud records from the real dataset which does not contain any internal core system user fraud label. Lastly, in the prediction stage, the issue is classified as predictive modeling. Internal core system user fraud prediction helps CBEs to protect their financial assets, safeguard their reputation, comply with regulations, improve operational efficiency, detect anomalies early, and

continuously adapt to evolving fraud threats. These benefits contribute to overall organizational resilience and sustainability in an increasingly complex and interconnected business environment [42].

2.7. Machine learning

Machine learning techniques are essential tools in various domains, allowing computers to learn patterns and make predictions or decisions without explicit programming. The success of machine learning models heavily relies on data quality and quantity [43]. Four general categories can be used to categorize machine learning: semi-supervised learning, supervised learning, reinforcement learning, and unsupervised learning.

Unsupervised learning, characterized by its ability to discern patterns and structures within data without the need for labeled information, encompasses tasks such as clustering and dimensionality reduction. Despite its capacity to work with less structured data, the significance of a substantial data volume remains imperative for meaningful insights [44].

Semi-supervised learning stands at the intersection of labeled and unlabeled data, employing algorithms that leverage both types. This approach involves combining a sizable amount of unlabeled data with a smaller set of labeled data, representing a bridge between the realms of supervised and unsupervised learning [45].

Reinforcement learning takes a different approach, where an agent learns decision-making through interaction with an environment. The data in reinforcement learning includes states, actions, rewards, and next states. By receiving feedback in the form of rewards, the agent refines its strategies, with successful models requiring extensive interaction with the environment to learn effective policies [44].

Supervised learning, the foundation of machine learning, entails training algorithms with labeled datasets containing input-output pairs. This process allows algorithms to comprehend underlying patterns and relationships within the data, enabling accurate predictions or classifications on new, unseen data. The labeled training data's quality and representativeness are pivotal for the success of supervised learning [46].

Supervised learning finds widespread application in classification and regression tasks, as illustrated in Figure 2.2. The algorithm, guided by the patterns it discovers, learns to classify incoming data into specified classes or labels and predict continuous outputs. The efficacy of supervised learning hinges on the quality of the labeled training data, making it a valuable tool across diverse domains, including natural language processing, medical diagnosis, image recognition, and speech recognition [46].

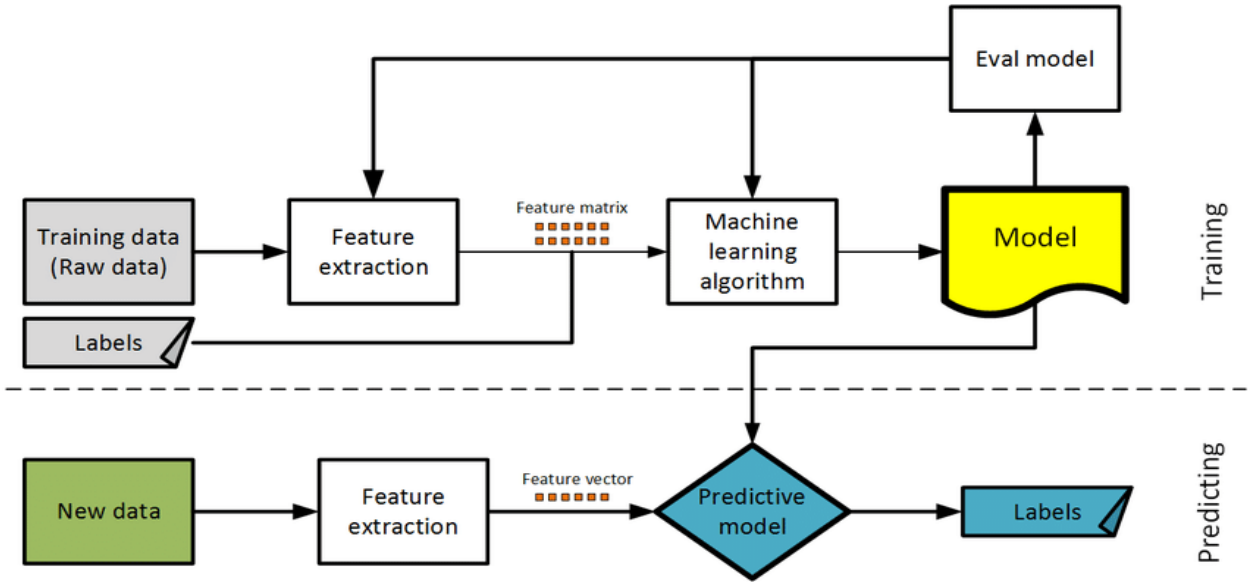


FIGURE 2.2. THE SUPERVISED MACHINE-LEARNING MODEL

The data quality used to train machine learning methods is crucial for their performance, regardless of the type of learning. Having high-quality, diverse, and representative data is essential. As machine learning evolves, it becomes more apparent that reliable data acquisition, preprocessing, and management practices are crucial in building accurate and generalizable models.

2.7.1. Machine Learning techniques

In machine learning the "no free lunch" theorem is a concept that suggests no single algorithm is universally optimal for all problems, particularly in supervised learning, where predictive modeling is involved, this theorem was introduced by david wolpert in 1996 and has significant implications for the field of machine learning [15]. Supervised ml techniques were employed with the proposed algorithms. Several machine learning techniques have been applied to internal user fraud prediction models in the past. Several machine learning techniques have been applied to internal user fraud prediction scenarios in the past.

There are various machine-learning techniques utilized to improve the accuracy and effectiveness of fraud detection systems when predicting internal core system user fraud. Logistic regression (lr) is frequently deployed due to its ease of use, interpretation, and efficiency in binary classification tasks. In handling high-dimensional data and capturing non-linear relationships through kernel functions, support vector machines (svm) are a popular choice. Random forest (rf), an ensemble approach, which combines several decision trees, is a powerful choice that is excellent at capturing complex patterns. K-nearest neighbors (knn) is valued for its simplicity and effectiveness in detecting local patterns based on the majority class of neighboring data points [47]. These techniques are often complemented by ensemble approaches, feature importance analysis, and rigorous data preprocessing, and collectively contribute to creating a comprehensive fraud detection system for internal core system users. Maintaining the efficacy of these machine-learning techniques in dynamic environments requires regular model monitoring, hyperparameter tuning, and adaptation to evolving fraud patterns.

2.7.1.1. Logistic Regression (LR)

Logistic regression is a dominant machine learning algorithm that belongs to the category of supervised learning approaches [44]. Due to the dichotomous structure of the dependent variable, there are only two viable classes. In simple words, the dependent variable is binary having data coded as either 1 (stands for not fraud) or 0 (stands for fraud).

According to the problem, this research needs a model that can classify things into two categories or predict a yes/no or 1/0 type output variable. A common model used for predicting binary outcomes is Logistic Regression.

A binary classification procedure that is a part of the generalized linear regression model is called logistic regression. More than two class problems can also be resolved with it. The internal core system user's fraud prediction can be modeled using logistic regression.

For instance, the "profile" may be one of the data's variables. The user's "gender" is another variable. The results of the logistic regression function will explain how the chance of fraudulent users is determined by the gender and/or profile of the users.

The Logistic Regression function is presented below, along with the Sigmoid Function.

$$P\left(Y = \frac{1}{X}\right) = 1 / (1 + e^{\wedge} (-1 \times (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \ldots + \beta_n X_n)))$$

where:

β_0 to β_n are various coefficients

X_0 to X_n are the independent variables impacting the dependent variable

And P (Y =1 | X) is the probability of a positive outcome.

Notice the exponent in the function. This is where linear regression plays in:

$$(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \ldots + + \beta_n X_n)$$

The sigmoid function used in logistic regression shows a transition between levels with output values ranging from 0 to 1. This function's characteristics aid in predicting binary outcomes. Based on the variable values, the output can be either at level 0 or 1, signifying the probability of the user committing fraud or not.

The following assumptions are made for Logistic regression [48]:

- **The dependent variable is binary.** This means that it can only take on two values, such as "yes" or "no", "fraud" or "notfraud".
- **The independent variables are not perfectly correlated.** This means that they should not be too highly correlated with each other. If they are, it can make it difficult for the logistic regression model to learn the relationship between the independent variables and the dependent variable.
- **The independent variables are linearly related to the log odds of the dependent variable.**
- **The residuals are normally distributed.** This indicates that the model's errors are dispersed randomly about zero. If this assumption is not met, then the standard errors of the coefficients may not be accurate.
- **The sample size is large enough.** The sample size should be large enough to ensure that the estimates of the coefficients are accurate. A general rule of thumb is that the sample size should be at least 10 times the number of independent variables.

2.7.1.2. Support vector machine (SVM)

Support Vector Machines (SVMs) are used to solve classification problems effectively on large datasets. SVMs utilize linear learning techniques and the interesting theory of kernel-induced spaces to achieve this [49]. Finding a hyperplane that can distinguishably classify data points in an N-dimensional space is the goal of the support vector machine method. In order to separate the two classes of data points, a number of hyperplanes could be chosen. But finding a plane with the highest margin—that is, the maximum separation between data points of both classes—is our main goal. We can raise the degree of

confidence in the classification of upcoming data points by optimizing the margin distance. Consequently, the best hyperplane is chosen from several options according to some optimization criteria (typically training set performance).

The function (hyperplane) that provides the maximum minimum distance to the examples is found by support vector machine techniques; we refer to this distance as a margin, and the examples that are closest to the margins are referred to as support vectors [50]. Because the solution solely depends on the support vectors, Figure 2.3 below illustrates how all of the points placed on the margin lines are support vectors, and the width of the margin equals the distance between these margin lines.

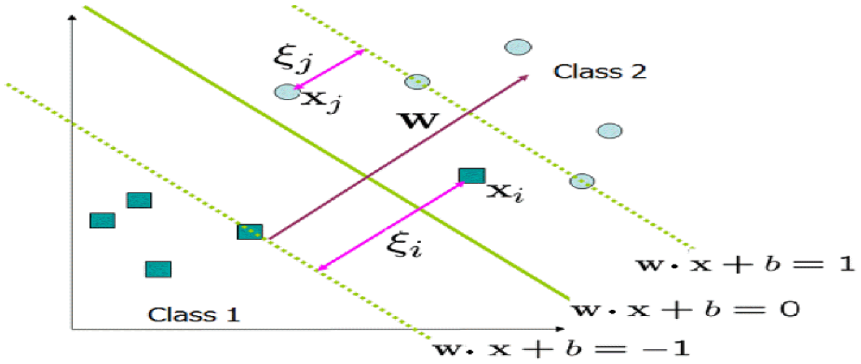


FIGURE 2. 3 CLASSIFICATIONS FOR SUPPORT VECTOR MACHINE [50]

Any hyperplane, as demonstrated in Figure 2.3 above, can be expressed as the collection of points X satisfying the equation w = b, where b is the hyperplane's offset from the original point along the direction of w, and w is a normal vector perpendicular to the hyperplane. We provide labels as, $\boldsymbol{y}$-$\boldsymbol{i}$. € {1, -1} given labels of data points X for two classes (class 1 and class 2). In the meantime, we categorize data X into class 1 or class 2 based on a pair of (,-$\boldsymbol{T}.\boldsymbol{x}$, b) and the sign of the function f (X) = sin (,-$\boldsymbol{T}.\boldsymbol{x}$+$\boldsymbol{b}$). Equations 2.1 and 2.2 can thus be used to represent the linear separability of the data X in these two classes.

$$\boldsymbol{f(xi) = xi \cdot w + b \cdot x + 1, yi = +1}. \text{ f or } \boldsymbol{y_i = +1} \text{ ----------(2.1)}$$

$$\boldsymbol{x_i}.\mathbf{w} + \mathbf{bw} - \mathbf{1}, \text{ f or } \boldsymbol{y_i} = -\mathbf{1} \text{ ----------(2.2)}$$

The above two equations can be combined and form the following equation 2.3.

Furthermore, $\mathbf{r} = (\boldsymbol{w^T x} + \boldsymbol{b}) / ||\mathbf{w}||$ can be used to calculate the distance between a data point and the separator hyperplane, $\boldsymbol{w^T x} + \boldsymbol{b} = 0$. The data points that are closest to the hyperplane are referred to as support vectors. The margin of the separator, or just 2/||w||, is the distance between support vectors, as

shown in Figure 2.3. Equation 2.3 can be used to formulate the quadratic optimization issue and solve linear SVM [51].

$$\underset{w,b}{\textbf{Minimize}} \left(\frac{1}{2}\|w\|^2\right) 9$$

$$\textbf{subject to } y(w^T x + b) \geq 1 \quad \text{---------- (2.3)}$$

Once the optimal separating hyperplane is identified for linearly separable data, only the data points that lie on its margin are considered as support vector points. Other data points are ignored and the SVM solution is then shown as a linear combination of these support vectors. Because the SVM learning method usually selects a limited number of support vectors, the complexity of the SVM model is therefore independent of the number of features in the training data. Other data points are ignored and the SVM solution is then shown as a linear combination of these support vectors. Because the SVM learning method usually selects a limited number of support vectors, the complexity of the SVM model is therefore independent of the number of features in the training data. Hence, SVMs are a suitable choice for learning tasks that involve features that have large number of in comparison to the number of training instances.

2.7.1.3. Random forest algorithm

The process of random forests involves distributing a sampling vector with equal values to every tree in the forest. This is followed by combining the predictions of each tree based on these randomly selected sampling vectors. The generalized inaccuracy of the forest and its tree depends on the strength and correlation of each individual tree [52].

Many decision trees are generated by the random forest algorithm; the number of trees constructed is a parameter that can be chosen during the learning phase. A decision tree is created by using a random subset of features taken from a training set that has been sampled with replacement. The final output of the algorithm is determined by the combined votes of every individual tree. Every decision tree classifies samples of input data using a tree method, and each tree is subsequently used to classify testing data. Every tree decides what to call whatever testing data it receives; this decision is called a vote. Ultimately, the voting is totaled by the forest, which =-then returns the class with the highest popularity as the testing data's categorization result [53]. This scenario is demonstrated in Figure 2.4 graphically.

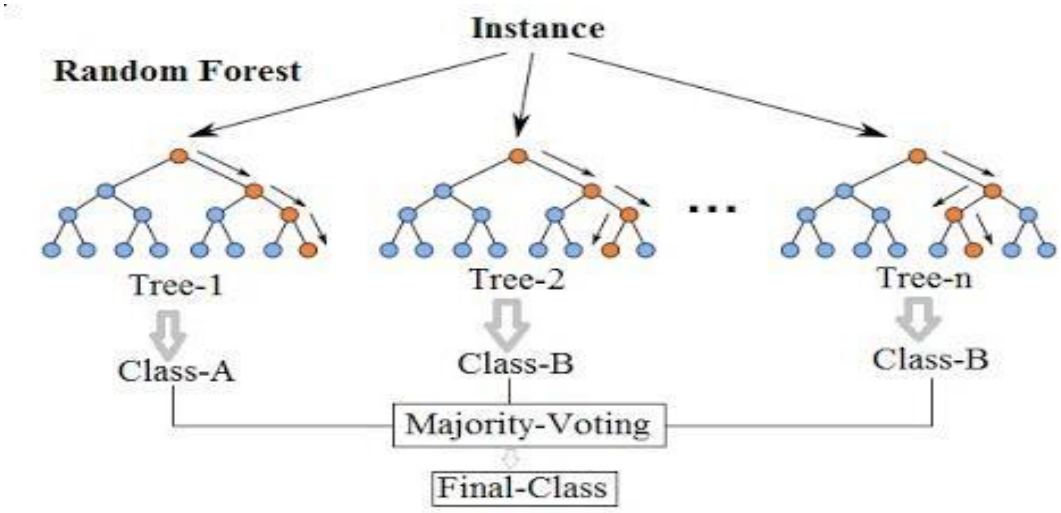


FIGURE 2. 4 RANDOM FOREST ENSEMBLE CLASSIFIER [53].

The model has the option to choose at random from a portion of the observations in the training data set when creating a single decision tree. In addition, when constructing the decision tree, each node takes into account a limited subset of the relevant variables. This minimizes the computation to a great extent. It can handle big datasets with multiple dimensions and is also appropriate when there are a lot of input variables and few observations. RF algorithms identify outliers and create trends [53]. The performance of a forest and its individual trees is determined by the strength of their correlations and generalization error.

The following stages are taken by the **RF** Algorithm in order to complete its categorization.

1. Choose T trees to be grown.
2. Choose the m variables that will be used to divide each node. m ~ M, M (input variables)
3. Grow T trees.

   When growing each tree do

   - Generate a tree by starting with a sample of size n from Sn which is obtained through the bootstrap method with replacement.
   - Choose m random variables at each node and use them to get the optimal split.
4. At every node, random variables are considered and utilized to obtain the most favorable split.
5. Gather votes from each tree in the forest to classify point X, and then use majority voting to determine the class label.

2.7.1.4. K-Nearest Neighbor (KNN)

One of the most basic machine learning techniques is the K-Nearest Neighbor (KNN) classifier. The main objective of the KNN algorithm is to forecast the categorization of a novel data point by utilizing a database that has sorted the data points into clear and separate categories. To examine the nearest examples of an object using any distance metric and a majority vote to decide its classification. The number of neighbors that will be used for voting depends on the value of K. The item will belong to the same class as its closest example if K=1. As K grows, we have to categorize an instance according to how similar it is to all K examples that were stated [54]. Though, a general guideline is to set K ≤ $\sqrt{n}$, where n is the total number of datasets.

Generally speaking, the paper starts with a dataset in which each data point has a predetermined type assigned to it. The objective is to use this knowledge to forecast a new data point's class based on the previous observations' classifications. We call this process the categorization problem. But first, a method for comparing items must be established before assessing how similar two observations are. Using a chosen distance metric, the closest examples of an object can be analyzed to determine its classification through majority voting. The number of neighbors utilized for voting is determined by the value of K, with the item being classified under the same category as its nearest example if K equals one. As K value increases, an instance's classification is based on its similarity to all K examples that were previously categorized.

The other problem here is figuring out which observations in the database are similar enough to the new observation so that the classification of those observations may be taken into account when classifying the new observation [55]. Euclidean distance is one of the most often utilized metrics. The formula provides the Euclidian distance between two instances (X1, X2, X3,…Xn) and (U1, U2, U3,….Un):

$$\sqrt{(X_1 - U_1) + (X_2 - U_2) + \cdots + (X_n - U_n)} \quad \text{----------(2.1)}$$

The predictors of instance #1 are X1, X2, X3, and Xn, while the predictors of instance #2 are U1, U2, U3, and Un.

TABLE 2.1 MACHINE LEARNING CLASSIFICATION ALGORITHMS.

Classification Algorithm	Advantage	Disadvantage

1. Logistic Regression (LR)	Ease of implementation, interpretability, and efficient to train.	The assumption of a direct correlation between the independent and dependent variables makes them inadequate for resolving non-linear issues.
2. Support Vector Machine (SVM)	Kernel-based algorithms are highly effective in high-dimensional spaces, providing great accuracy, power, and flexibility. They work best when there is a clear margin of separation, and they have many applications.	This algorithm may not perform well with large datasets and is more complex to program. Additionally, it may struggle when the data contains a high level of noise, resulting in overlapping target classes.
3. Random Forest (RF)	This algorithm does not suffer from overfitting and can be utilized for feature engineering, allowing for the identification of the most important features from a given set. It performs exceptionally well on large datasets, offering high flexibility and accuracy. Additionally, there is no need for extensive preparation of the input data.	This algorithm is known for its complexity and demands significant computational resources. It can be time-consuming and requires careful selection of the number of trees to achieve optimal performance.
4. K-Nearest Neighbor (KNN)	This algorithm is straightforward to comprehend and simple to implement. It requires minimal training time and performs well with multicast datasets. It possesses good predictive power and demonstrates strong performance in practical applications.	This algorithm may encounter a computationally expensive testing phase and can be affected by skewed class distributions. Furthermore, when working with high-dimensional data, its accuracy may drop, and in order to get the best results, a parameter k must be defined.

2.8. Related Works

Previous studies have focused on predicting internal user fraud using machine learning algorithms applied to user data. This section highlights a few of these studies.

The authors in [56] conducted a systematic literature review of 93 articles on financial fraud detection using machine learning. They found that support vector machines (SVMs) and artificial neural networks (ANNs) are the most popular machine learning algorithms for fraud detection. SVMs are supervised learning algorithms that can classify data into two or more categories. The writers found that credit card fraud is the most frequent fraud handled by machine learning techniques. Credit card fraud happens when someone steals a credit card and uses it to buy things they're not supposed to. The authors suggest that machine learning can detect credit card fraud by identifying unusual patterns in credit card transactions. The authors conclude that machine learning is a promising tool for fraud detection in the banking sector. However, they also note that there are some challenges to using machine learning for fraud detection, such as the need for large amounts of data and the potential for false positives.

The paper provides a valuable overview of the use of machine learning for fraud detection in the banking sector. The authors' findings suggest that machine learning is a promising tool for detecting fraud, but some challenges need to be addressed before it can be widely adopted. Credit card fraud is the most common type of fraud addressed using machine learning techniques.

In paper [57], suggested a machine learning-based solution for anomaly detection. The system models a typical data set using a hidden Markov model and uses a decision tree classifier as two discrete methods for automated learning and training (HMM). These methodologies rely on the examination of large amounts of data. A number of metrics are used to assess the classification performance of the system, including accuracy, precision, false positive rate (FPR), and true positive rate (TPR). The system was able to achieve a classification accuracy of 93.54% and a false positive rate of 4.09% when compared to other single-level machine learning algorithms.

In another research conducted in [58], a predictive model was developed using data mining methods to detect subscription fraud in the context of Ethio Telecom. To identify and describe cases of subscription fraud specifically for Ethio Telecom, the present study analyzed 25,000 data records empirically, using data mining methods. Four different categorization techniques were applied by the researcher: J48,

PART, Random Forest, and Multilayer Perceptron of Artificial Neural Network. The data was processed using WEKA software. The research focused solely on prepaid customs Call Detail Record (CDR) data. The study suggests that future research should go beyond prepaid subscriptions and investigate other types of telecommunication fraud, particularly postpaid subscriptions. The current study is limited to the prepaid subscription category.

In the paper, [59] proposed that the authors combine machine learning techniques and data analytics to detect financial transaction fraud. There are four stages in the framework: data preprocessing, feature selection, model training, and model evaluation. The authors have trained the model using various machine learning algorithms like logistic regression, decision trees, and support vector machines (SVM). They have also evaluated the performance of the models using performance metrics such as accuracy, precision, recall, and f1-score. In addition, the paper includes a case study where the proposed framework is applied to a genuine dataset of credit card transactions. The results demonstrate that the proposed framework can accurately detect fraudulent transactions with high precision The study analyzed several models and found that random forest had the most accurate rate of 96% when predicting and detecting fraudulent credit card transactions. Therefore, the study suggests that random forest is the most suitable machine learning algorithm for predicting and detecting fraud in credit card transactions.

The article [60], discusses the increasing problem of credit card fraud and the limitations of rule-based approaches to detect and prevent this type of fraud. It highlights the need for more intelligent and predictive methods such as Data Mining and Machine Learning algorithms. The article reviews and evaluates various algorithms like Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), naïve Bayesian, k-nearest Neighbor (k-NN), Decision Tree, and Frequent Pattern Mining in detecting fraudulent transactions. The effectiveness of different supervised machine learning algorithms for detecting fraud k-NN, Naïve Bayesian, and Decision Tree algorithms are not effective in detecting new frauds, they are easy to understand for non-AI experts. On the other hand, ANNs and SVMs are more capable of learning and adapting to new fraud patterns. Although these methods may be complex, they provide valuable insights to the banking industry to make better business decisions. The challenge faced by the author is the absence of a standard transaction dataset, which makes it difficult to compare various Machine Learning techniques. Furthermore, due to the lack of complete information on the dataset's fields or dimensions, it becomes more challenging to justify why certain approaches are superior to others.

In this paper, [61] protecting computer networks and systems presents a considerable challenge, primarily because of the existence of insider threats. Recent progress in modern machine learning algorithms have enabled the resolution of complex issues by uncovering latent patterns and creating data models. Utilizing the Deep Feature Synthesis algorithm, behavioral features were derived from historical data. Through the application of advanced machine learning algorithms encompassing both anomaly detection and classification models, insider threats were identified with a precision rate of 91%. The experimentation drew upon the CERT insider threats dataset, a publicly available resource. The testing of the SMOTE balancing technique aimed to mitigate the impact of an imbalanced dataset, revealing enhanced recall and accuracy at the cost of precision. Notably, the feature extraction process combined with the SVM model produced better outcomes, surpassing all other machine learning models and achieving a flawless 100% accuracy for the classification model.

The literature review above makes clear that no study has been done to predict internal core system user fraud using a machine learning method. Most of the research focused on financial fraud detection using machine learning algorithm and proposed framework on fraud detection using data analytics and machine learning algorithms. Instead of suggesting farmwork, the goal of this study is to create a prediction model that can recognize fraudulent activity and its associated behaviors, making it useful for making decisions.

Here you have to provide a summary of the related work. In addition to this, you have to provide challenges and research gaps identified from the literature review and related work.

TABLE 2.2 SUMMARY OF SELECTED RELATED PAPERS

Reference	**Area**	**Business**	**Techniques Applied**	**Results**
(Ali, A., Abd Razak, S., Othman, S.H., Eisa, A. Financial Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H., Saif, and T.A.E. , 2022)	Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review	Banking industry	Naïve Bayes, Decision tree, Hidden Markov's model, KNN, LR, RF, SVM, and ANN	SVM and ANN

(Ali Moradi Vartouni, Saeed Sedighian Kashi, and Mohammad Teshnehlab, (2018)	Anomaly detection system is on machine learning for Web log files	Banking	decision tree, hidden Markov model (HMM)	hidden Markov model (HMM)
(T. Haddish, Diss. AAU, 2013.)	Constructing a Predictive Model for Subscription Fraud Detection Using Data Mining Techniques,	Telecom	J48, PART, Random Forest, and ANN	random forest
(Jonathan K, Kassim T., and Wilhemina A. 2023)	A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions	Banking industry	logistic regression, random forest, and decision trees	Random forest (RF) is the most appropriate
(Kha Shing Lim, Lam Hong Lee, and Yee-Wai Sim (Vol 21, Issue 9/Page 31-40/2021))	A Review of Machine-Learning Algorithms for Fraud Detection in Credit Card Transaction	Banking industry	ArtificialNeural Networks, k-Nearest Neighbour(k-NN), Support Vector Machines (SVMs), naïve Bayesian, , and Decision Tree	ANNs and SVMs are more capable
(Bushra B., and Najwa A. Appl. Sci. 2023, 13(1), 259; https://doi.org/10.3390/app13010259)	Insider Threat Detection Using Machine Learning Approach	Financial sectors	SVM, NN, AdaBoost, Random Forest	The SVM model obtained the highest overall performance compared to the other algorithms

2.9. Summary of Related Works

The topic of Internal core system users' fraud prediction has discussed various internal users' fraud algorithms, which are relevant to the current internal employees that have access to the core system. Different methods have been proposed to classify multiple internal users' fraud problems and each approach has numerous advantages and limitations. While the performance of existing approaches has been improved substantially, there is still abundant room for further progress on the current topic.

Some of the researchers have not handled the class imbalance problem. Additionally, most of the above-listed research was done in developed countries and they have used archived data, especially taken from Kaggle and also which cannot easily describe the real-world problem. Most of the above-listed studies have used a small number of datasets except the firs one, which may not describe the overall problem and it may not develop and select a better model. Local researcher [58], studies used a small number of references, these also incorporated strongly related attributes in different columns, papers not used normalization, the number of instances has small, the models were vulnerable to over-fitting.

To address the limitations and research gaps presented in this section, this research used 7,754 instances with 12 attributes. This study has focused on covering the data preparation steps, and handling class imbalance using SMOTE technique, and also this study has applied four machine learning algorithms, which are Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbour (KNN algorithms are used in internal core system users' fraud prediction are applied and the algorithm with the best predictive performance is to be selected for model performance optimization. Though the aforementioned researches have a substantial contribution to showing the directions in conducting this research, this area (internal core system users' fraud prediction) is not been studied much locally. So, the study fills the knowledge gap as to how churners can be predicted from existing historical data.

CHAPTER THREE
RESEARCH METHODOLOGY

3.1. Overview

In this chapter, the research design, dataset description, correlation analysis, data preprocessing techniques, architectural design, and evaluation metrics of the research are incorporated. This study also follows the design science methodology and each step is carried out and explained in detail in this chapter. Python programming language was used to carry out the experiments of the research.

3.2. Research Design

The Design Science Research (DSR) method is a problem-solving approach that aims to enhance human knowledge through the creation of innovative artifacts. In simple terms, DSR aims to enhance the bodies of knowledge in science and technology by creating innovative solutions that solve issues and enhance the environments in which they are used [62]. The study's general approach is a design science, which involves gathering and organizing critical system user data. This data includes user information such as gender, age, Emp ID, user profile, the action performed, the user operating machine or computer, the location of the action performed, and the user's years of service, etc. The paper refrains from using users' identification data, such as names and addresses, etc., which have been compiled and made available for reporting and administrative purposes and used as input data for the model.

The outcomes of Design Science Research (DSR) encompass not only newly created artifacts but also a comprehensive understanding of design theories that explain how these artifacts can either boost or hinder the relevant application contexts. The primary objective of this methodology phase is to provide a concise overview of DSR concepts to facilitate a deeper comprehension of subsequent chapters that feature DSR case studies. DSR functions as a paradigm for problem-solving and has its roots in engineering and artificial science [63]. Its aim is to enhance human knowledge by generating innovative artifacts and design knowledge that offer inventive solutions to real-world issues [64]. Consequently, this research paradigm has gained considerable attention over the past two decades, owing to its potential to foster innovation capabilities among organizations and contribute to the much-needed sustainability transformation of society [65].

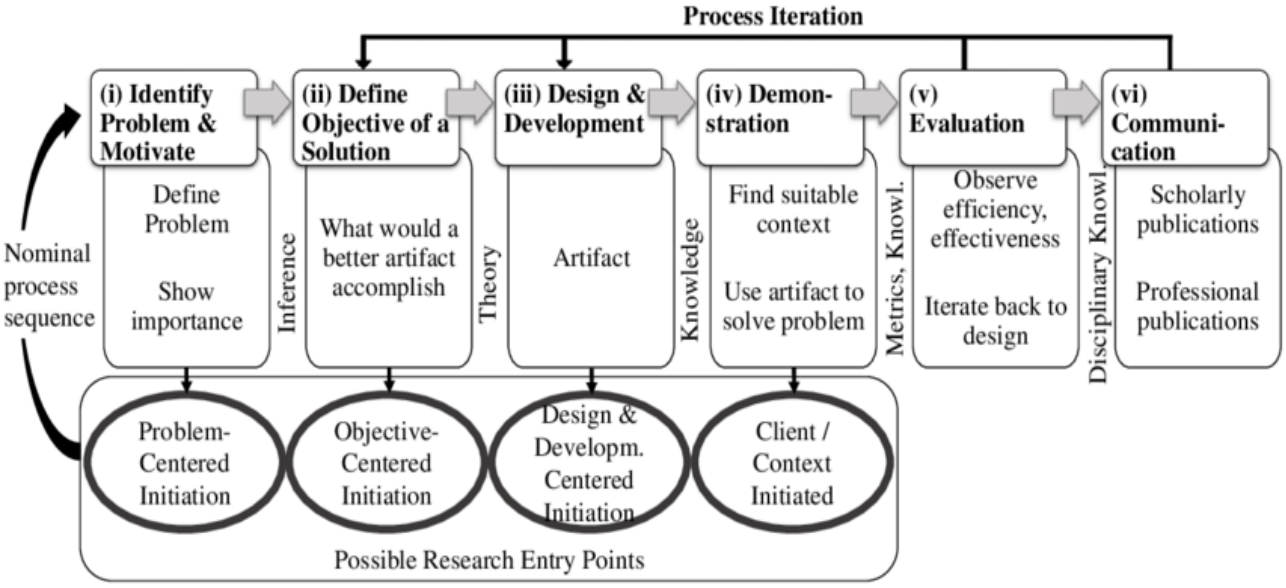


FIGURE 3.1 THE GENERAL RESEARCH DESIGN & METHODOLOGY [66]

The creation of design artifacts through design science research is acknowledged to vary in process across different studies. A six-step methodology for performing design science research is included in the developed process. These steps are problem identification and motivation, solution objectives, design and development, demonstration, assessment, and communication [66].

3.2.1. Problem Identification and Motivation

The main point of the step is the identification of the problem sets the stage for developing a solution, and the motivation provides the rationale for addressing the identified problem. The first chapter of the thesis report has explained this by defining the issue and verifying the need for a resolution.

3.2.2. Objectives of a Solution

The second step in the DSR process model involves determining the objectives of the proposed solution, while the first step involves identifying the problem in general. As per [66], these objectives can be either quantitative or qualitative. Quantitative objectives refer to the terms in which a desirable solution would be better than the current ones. On the other hand, qualitative objectives involve a description of how a new artifact is expected to support solutions to problems not hitherto addressed. Based on the researcher's statement of the problem in Chapter One, this study has qualitative objectives. As per [67], the objectives should be inferred rationally from the problem specification. Therefore, the objective of the internal user fraud prediction model is to assist the Commercial Bank of Ethiopia (CBE) in

manipulating datasets efficiently. This involves designing and developing a predictive model for internal core system user fraud in CBE through machine learning approaches. The research model will help to understand the behavior and wants of internal users so that remedial actions can be taken by the management. It also assists in minimizing the effort for fraudulent activities. Furthermore, it helps maximize profit and the overall success of the bank. Accordingly, building a dataset, including data collection, data pre-processing, data splitting, and data sampling, is incorporated in this phase.

3.2.2.1. Data Collection

The objective of this study is to develop an optimal model for the prediction of internal core system user fraud. To achieve this, a dataset from the Commercial Bank of Ethiopia (CBE) is required to build a suitable model. Data plays a crucial role in machine learning algorithms [68]. Therefore, data collection is a critical aspect of this study. As this research focuses on the CBE, which has numerous users operating on the core system, the data collection has been carried out taking the whole Data get from CBE IS data management by a simple random sampling method from the Kirkos district. The initial dataset has been collected from possible sources at the CBE. The bank has a computerized system that stores all the users' profile data. To execute machine learning research, proper planning and a profound understanding of data are required. This is the second step in the DSR process, which entails gathering data, examining data, assessing the quality of the data, and drawing conclusions from the data to develop a hypothesis.

The CBE internal core system users' dataset before pre-processing, i.e., before feature selection and resampling, contained 12 variables. It contained information on 7,754 users of CBE. The target variable was 'fraud,' a Boolean variable that indicates whether a user action was fraud or non-fraud. A few of the variables are discussed in detail:

Table 3.1 Sample of the Dataset from Header

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Emp_ID	Sex	Age	Referer	Profile	Natic	Lev	Input_	Action_	Action_E	Operat	Servi	fraud
2	3206	Male	20	FT	Teller	ET	1	yes	view	same	single	0	No
3	8197	Male	35	FT	Manage	ET	5	no	no actic	same	single	12	No
4	6183	Female	26	TT	Auditor	ET	3	yes	no actic	same	single	8	No
5	1386	Male	22	FT	Cheker	ET	2	yes	validate	same	multipl	3	Yes
6	7897	Male	24	FT	Maker	ET	2	yes	reverse	same	single	3	Yes
7	9546	Female	26	FT	Teller	ET	1	no	reverse	same	single	5	No
8	5298	Male	35	FT	Teller	ET	1	yes	reverse	same	single	0	No
9	8120	Male	23	FT	Teller	ET	1	yes	view	same	single	0	No
10	4427	Female	22	FT	Teller	ET	1	no	no actic	same	single	0	No

The ‘fraud’ is the target variable with values as ‘1’ for those who are fraudulent user with CBE and ‘0’ for those who are not fraudulent of CBE. Internal fraudulent users refer to individuals or entities within in CBE who engage in fraudulent activities for personal gain or other malicious purposes. These users have authorized access to the CBE’s systems, data, or resources due to their status as employees, contractors, or other affiliated parties. On the other hand, internal non-fraudulent users refer to individuals or entities within a system, CBE, or network who are legitimate and do not engage in fraudulent activities.

3.2.2.2. Data Preparation

The success of machine learning is heavily dependent on the quality and the quantity of the data beside the pre-processing [69]. This phase is considered one of the most critical stages in the process of machine learning. During this stage, data is collected from the source, described, and explored to gain a better understanding. However, the raw data can't be used directly for modeling purposes due to several reasons. Firstly, the data may contain errors, outliers, or missing values that need to be addressed before proceeding. Secondly, the importance of each attribute and domain may differ based on the intended purpose, requiring data selection. Additionally, there may be a need to derive new attributes or tables from the existing data. Furthermore, integrating data from multiple tables may become necessary. Lastly, the original data format and structure may not be suitable for modeling tools and techniques. As a result, a number of tasks are involved in the data preparation phase to transform the raw data into a final dataset that is appropriate for modeling algorithms. This involves various tasks such as data

cleaning, outlier removal, missing value imputation, new attribute construction, and data transformation. To prepare the final dataset for the experiment, a number of data pre-processing processes were carried out based on the comprehension of data from the previous phase. The process included handling missing values, encoding, normalizing the data, feature extraction, standardizing the data, data splitting, and applying the SMOTE Technique in this study.

3.2.2.2.1. Handling Missing Values

Managing missing values is crucial as several datasets with missing information are not supported by machine learning techniques. The dataset provided for analysis contains a considerable number of incomplete values, which might be due to a variety of factors. For instance, failure to collect data can lead to missing data. It may be necessary to eliminate variables from the final dataset if more than 60% of their values are missing [70]. Continuous variables that can take any value between their minimum and maximum values, and have missing values ranging from 2% to 30%, have had their values imputed using mean values, which is a reasonable estimate for randomly selected observations from a normal distribution. Missing data can occur due to various reasons and can lead to several problems, such as reduced statistical power, which can lead to an incorrect evaluation of the hypothesis. Missing data can also decrease the representativeness of the sample and complicate the data analysis. Certain algorithms are unable to handle missing data. The maximum likelihood and last observation carried forward strategies were used to impute the missing values in categorical variables that contain labels. These are the most popular approaches for impute missing values [71]. In maximum likelihood, the missing values were imputed with the values that occurred most often. In the last observation carried forward technique, the previous observation was imputed in the missing value. During the analysis of the dataset, it was observed that there were missing values present in the categorical variable. The mode technique was utilized for imputing the missing values in the categorical variables. Python's Gender, Reference, and Action_Performed methods were used to impute missing values in categorical variables using mode values. Age and Service Year, two continuous variables with missing values ranging from 2 percent to 30 percent, were handled by mean values. For randomly chosen data points from a normal distribution, the mean provided a reasonable approximation [72].

3.2.2.2.2. Encoding

The efficacy of machine learning models is depending on various factors, such as the model and its hyperparameters, as well as how we process and provide diverse data types to the model. Since machine learning algorithms rely on mathematical equations, they can only process numeric data. Consequently, categorical variables must be converted into numeric form. To achieve this, Sklearn's LabelEncoder function is employed to transform these values into numerical values. Categorical data types are frequently found in datasets, and encoding is the act of turning them into numerical data. In this particular dataset, there were nine categorical variables or nominal values, which were stored as text values.

In this study, the Label Encoding method was used to encode the categorical variables. Every category was assigned a numerical value, and this technique neither adds more columns nor slows down the learning curve. The variables Gender, Reference, Profile, Nationality, Input_Action, Action_Performed, Action_Branch, Operating_Machine, and Fraud were all encoded using a label encoder.

3.2.2.2.3. Normalizing Data

Normalization is a crucial technique in data pre-processing for developing machine learning models. Its primary objective is to transform the data into a standard scale, which makes it easy to compare and analyze. To identify skewed data variables, the skewness and kurtosis were measured for each numerical column. If the values were found to be outside the +/-2 range, the corresponding variable was considered skewed. The histogram was also used to determine the data's normal distribution. The normalization process was performed using the MinMaxScaler() function from the sklearn library.

The continuous variables were first evaluated to determine their skewness and kurtosis. If the variables had skewness and kurtosis values outside the [0, 1] range, they were regarded as not normally distributed. To determine if the variables were regularly distributed or not, the histogram was also utilized. The sklearn MixMaxScaler() function was applied to specific variables in the CBE dataset to normalize the data. The use of this function ensures that the variables are brought to a common scale, which is essential in developing machine learning models.

3.2.2.2.4. Feature Selection

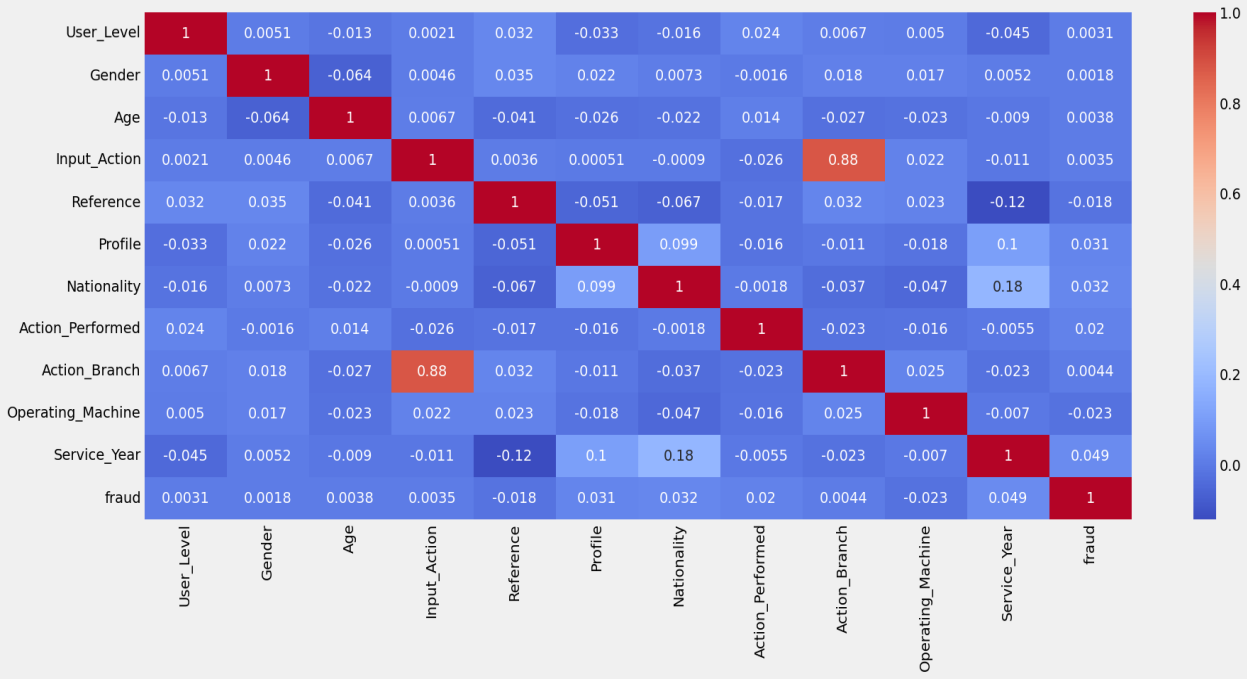
The primary objective of this technique is to determine significant features that can be utilized in model development. It is imperative to note that feature selection is an essential step in ensuring optimal model

performance, as it enables the identification of key variables that contribute most significantly to the model's output. As such, the utilization of feature selection techniques in the model construction process is highly recommended to ensure the production of accurate and reliable results.

Correlation Analysis

In order to look into the relationship between the independent and dependent variables and to find any possible multicollinearity among the independent features, the gathered data was put through a correlation analysis. The 'Spearman' correlation technique was utilized to identify the correlation because the study consists of both continuous and categorical variables. The resulting correlation heatmap and matrix in table 3.2 were developed.

TABLE 3.2 CORRELATION HEATMAP OF THE VARIABLES BEFORE FEATURE SELECTION



When selecting features, it is crucial to identify correlations, as demonstrated in Table 3.2. The correlation heatmap reveals that the variable Operating_Machine has the highest correlation but is negatively correlated with the target variable 'fraud'. Meanwhile, other variables, such as Action_Branch and Input_Action, exhibit high correlations with other independent features.

Nationality, on the other hand, is simply an identifier and not a descriptor. To avoid multicollinearity, Action_Branch, Input_Action, and Nationality were removed prior to modeling. In this study, feature selection was conducted with the use of a heatmap and a correlation matrix. This approach determined the correlation between dependent and independent variables, as well as the correlation between dependent variables themselves. Correlation measures the strength of dependence between two variables, and if the correlation exceeds 0.5, the variables are not considered in the model, as it would affect the accuracy of the model [73]. The size of correlation with interpretation is displayed in the table 3.3.

TABLE 3.3 CORRELATION TABLE [73]

Size of correlation	Interpretation
0.90 to 1.00 (- 0.90 to -1.00)	Very High Positive (Negative) Correlation
0.70 to 0.90 (- 0.70 to - 0.90)	High Positive (Negative) Correlation
0.50 to 0.70 (- 0.50 to - 0.70)	Moderate Positive (Negative) Correlation
0.30 to 0.50 (- 0.30 to - 0.50)	Low Positive (Negative) Correlation
0.00 to 0.30 (0.00 to - 0.30)	Negligible Correlation

Feature selection is a crucial step in improving the accuracy of machine learning models. It not only enhances the speed of model training but also reduces model complexity. To identify the most predictive features for feature selection in predicting the target, a Tree-based classifier was implemented based on a thorough literature review. Predicting the best features is a successful application of ensemble learning. Feature selection, sometimes known as variable selection, attribute selection, or variable subset selection, is a crucial step in the creation of models in the fields of statistics and machine learning. In this study, the correlation method was employed to identify the best-predicting features for detecting internal core system user fraud. By analyzing the correlation between the independent and dependent variables using a correlation matrix, we not only detected the multicollinearity issue but also identified the most relevant feature in predicting the output. In summary, the use of feature selection techniques is vital to building robust machine learning models. By identifying the most relevant features, we can improve model accuracy, reduce

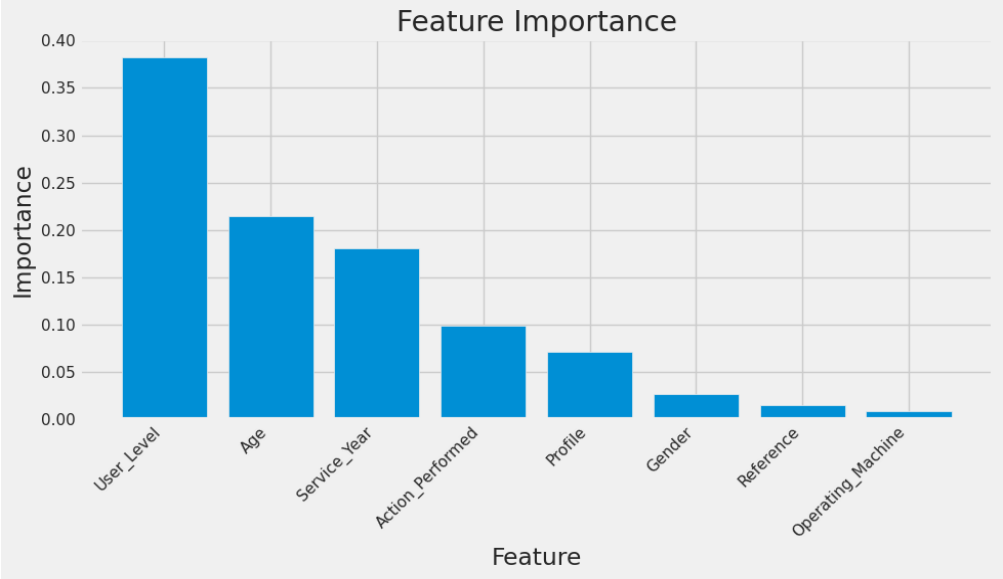


FIGURE 3.2 FEATURE IMPORTANCE

Based on the graph shown above, it can be inferred that User_Level, Age, Service_Year, Action_Performed, Profile, Gender, Reference and Operating_Machine are the most crucial factors to consider when predicting the target.

3.2.2.3. Data Splitting

The dataset that was obtained was divided into training and testing datasets after missing values were addressed, and data was normalized.

Based on previous research [74], the final dataset was divided into 80% training and 20% testing datasets. To create models and determine their accuracy, the models were created 40 times using a for loop and the sklearn train test split function after the data was divided at random. The accuracy scores were stored in a list. The supervised machine learning models were compared using their accuracy scores, and the best model was selected to predict internal core system user fraud. The same data split was used for all models to fit the model and calculate the accuracy score, ensuring that the accuracy results were not affected by the data split.

3.2.2.4. Data Sampling

Class imbalance is a common issue in many real-world applications, where one class (usually the existence value) has a majority of labeled examples, while the other class (usually the important one, such as fraud) has fewer labeled examples. This problem exists in various application domains.

In this study, the minority class (fraudulent users) was crucial, and a decision was made on how to handle the class imbalance problem before experimenting. The study found a class imbalance ratio of approximately 22:1, meaning that for every fraudulent user, there were 22 non-fraudulent users. The non-fraudulent users accounted for 95.6% of the total members, while the fraudulent users accounted for 4.4% of the total dataset. To handle the class imbalance problem in the data pre-processing phase, two data-level methods were considered: random under-sampling and random oversampling. However, in this study, the class imbalance issue was addressed using the SMOTE approach.

3.2.2.4.1. SMOTE Technique

On the training dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to address the problem of class imbalance (which comprised 80 percent of the total dataset). SMOTE generates synthetic samples from the minority class, rather than simply duplicating existing data [75]. SMOTE efficiently balances the data by taking similar records from the minority class and making random adjustments to each column one at a time. The resulting synthetic samples were added only to the minority class records. As shown in figure 3.3 and distribution table for the 'fraud' variable, class imbalance is a common data problem.

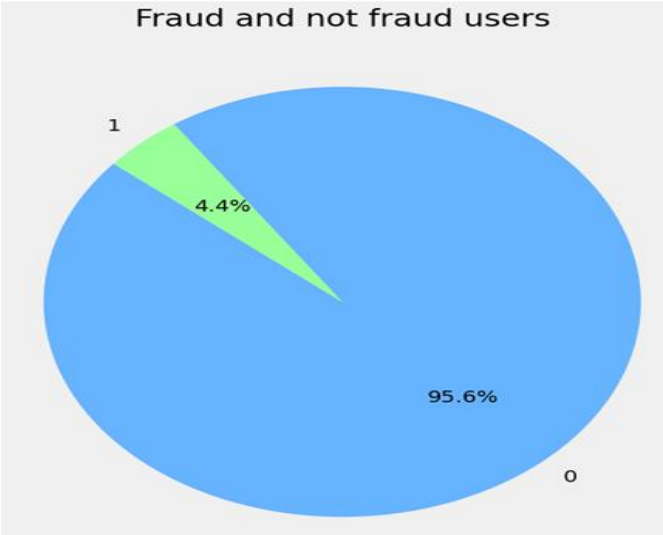


FIGURE 3.3 ALLOCATION OF INTERNAL USERS' FRAUD AND NOT FRAUD BEFORE RESAMPLING

In this study, the class imbalance issue was tackled by utilizing the Synthetic Minority Over-Sampling Technique (SMOTE), which has been previously researched and proven effective [75]. SMOTE generates new minority instances by synthesizing them between the existing minority instances. By

using this technique, we were able to balance the data. The dataset we worked with had a class imbalance ratio of approximately 22:1, meaning that for every fraudulent user, there were 22 non-fraudulent users.

TABLE 3.4 TARGET VARIABLE COUNT

fraud	
Yes (1)	**No (0)**
4.4%	**95.6%**
341.17	**7,412.82**

3.2.3. Design and Development

The process of creating an artifact [66] involves identifying the necessary functionalities and architecture, followed by the actual creation of the artifact. According to [68], these artifacts can take the form of constructs, models, methods, or instantiations. In this particular study, the artifact being created is an internal core system user fraud predictive model, which has been designed and formed based on the objectives previously mentioned.

To accomplish this, a dataset from the CBE is utilized, and the information gathered is incorporated into the set of patterns and principles being developed. The result is the proposed internal core system user fraud predictive model for the Commercial Bank of Ethiopia (CBE), which is designed to address the firm's internal user control issues.

3.2.3.1. Proposed Architecture

This study presents models and their implementation for predicting internal core system user fraud using different machine learning algorithms. The scenario used for this study is CBE internal core system user fraud. The study attempts to find different patterns of internal core system user fraud in the given training datasets by applying machine learning techniques. The several phases in the machine learning cycle are described in the machine learning architecture. It covers the main procedures needed to convert unprocessed data into training sets that a system may use to make defensible conclusions. The comprehensive architectural layout is depicted in Figure 3.4.

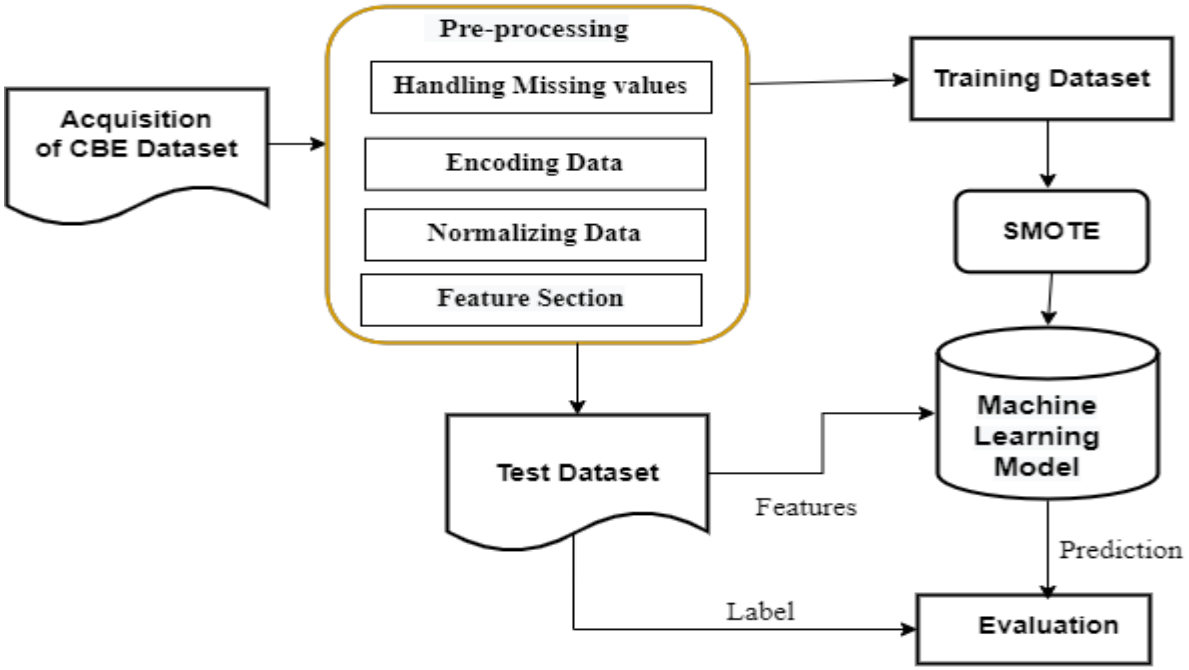


FIGURE 3.4 PROPOSED ARCHITECTURE OF INTERNAL CORE SYSTEM USER FRAUD PREDICTION

The architecture of internal core system user, as depicted in Figure 3.4, involves collecting data from the Commercial Bank of Ethiopia. The collected data undergoes pre-processing, which includes handling missing values, encoding, normalization, and feature selection. After pre-processing, the dataset is split into training and test datasets, with 80% of the data used for training and 20% for testing. The training dataset is then subjected to the SOMTE technique to balance the imbalanced data, followed by SMOTE to build a machine-learning model. Finally, the best model is selected based on the evaluation of the test dataset.

3.2.4. Demonstration

The utilization of the internal core system user fraud prediction model, which is the artifact, is showcased in this activity to address one or more of the problem's occurrences.

3.2.5. Evaluation Methods

Proper model assessment, algorithm selection, and model selection are essential in scholarly machine learning research as well as various industrial applications [65]. The primary goal of evaluating a model is to estimate its generalization error, which measures its effectiveness on previously unseen data. A successful machine learning model should not only perform well on training data but also maintain high performance on new data. Therefore, it is important to ensure the model's ability to retain its performance when exposed to new data before deploying it into a production environment [65]. Evaluating the model's efficacy and effectiveness is critical to demonstrate its capabilities. This iterative approach helps improve the proposed solution to address the identified problem and enhance the overall quality of the solution. In this study various classification models were constructed and assessed through the implementation of distinct sets of training and testing data. The experimental results obtained from the classification models were subjected to analysis and evaluation of their respective performance accuracies, utilizing the confusion matrix. One of the prevalent and favored techniques in use is the recommended approach. The selection of a preferred methodology for assessing output generated by distinct algorithms in the context of datasets varying in size from small to moderate is under consideration.

Confusion matrix

The application of the confusion matrix is commonly employed for the evaluation of a model's efficacy in addressing binary classification tasks for a predetermined dataset. False positive (fp) and false negative (fn) are the classifications for incorrectly classified cases, whereas true positive (tp) and true negative (tn) are the categories for accurately identified occurrences

TABLE 3.5 CONFUSION MATRIX

		Predicted value	
		0	**1**
Actual value	**0**	**True negative (TN)**	**False positive (FP)**
	1	**False negative (FN)**	**True positive (TP)**

- **Accuracy:** This measure determines how many samples are correctly classified compared to the total number of samples. If the distribution of classes is uneven, accuracy may not be the best metric to use. When there is a large imbalance in the number of instances between classes, it is important to achieve balance to gain a better understanding of the subject matter. Recall and precision are often recommended alongside accuracy in academic discussions.
- **Precision:** Figuring out what proportion of all the positive predictions were actually positive is a crucial task. Precision alone isn't enough to accurately identify how many positive instances were mistakenly classified as negative. Therefore, it's necessary to calculate the True Negative Rate (TNR) matrix to address this issue.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall:** The percentage of predicted positive cases out of the total positive is the same as the true positive rate (TPR). A higher sensitivity value results in more true positives and fewer false negatives. Conversely, a lower sensitivity value indicates lower sensitivity, leading to fewer true positives and more false negatives. When a good thing is identified correctly more often, and bad things are occasionally identified as good, it is considered a positive outcome.

$$Recall = \frac{TP}{TP + FN}$$

- **F1 Score:** This particular kind of average takes into account both precision and recall, acknowledging both false positives and false negatives. As a result, it is particularly effective when applied to datasets where certain elements have more instances than others.

$$F1\,score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

CHAPTER FOUR

EXPERIMENT RESULTS AND DISCUSSION

4.1. Overview

This section explains the implementation of the experiment, following the steps outlined in Chapter Three. It addresses the experimental setup, methodological approach, resultant findings, and the analytical insights derived from the conducted experiments. The primary objective of this investigation was to formulate and compare various machine learning methodologies employing a comprehensive CBE Internal Core System Users dataset to predict instances of Users' fraud. The models instantiated encompass Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN with a subsequent comparative analysis of their respective outcomes. Finally, this chapter presents the research's advantages.

4.2. Experimental Setup

Various techniques and tools are utilized for the implementation of the Internal core system user fraud prediction model. The proposed system has been implemented on a laptop with the following configuration: Windows 11 operating system, 11th Gen Intel(R) Core (TM) i5-1135G7 @2.40GHz, and 16GB RAM. Python was utilized as the programming language together with the TensorFlow and Keras anaconda environment libraries to implement the algorithms using software tools. These tools satisfy all the requirements and are utilized in Python.

Tensorflow:

TensorFlow is a library that was created by Google and is available for free as an open-source software. It is currently the most widely used and fastest deep-learning library, according to a report by [76]. Users may carry out data preparation, model development, training, and estimating using TensorFlow's design. Tensors, which are ndimensional arrays, are used to represent all types of data in TensorFlow computations. Additionally, during preparation, TensorFlow uses a graphical framework to visually describe the computational sequence.

Anaconda:

The data preprocessing and major implementation and evaluation mechanisms of this study were carried out using the PYTHON programming language, with the anaconda version of "conda 4.11.0". Python

was chosen as the best programming language for prediction purposes due to its simplicity in coding, high-level language, extensive standard library, extensible language, less code, and numerous libraries. Anaconda, a free and open-source Python distribution designed for data science and deep learning applications, was utilized to develop the model. Its goal is to make package management and deployment easier. The coding part was composed using various IDEs such as Jupyter Notebook and Spyder, with Jupyter Notebook being used for implementation. It is a simple and easy-to-use tool that runs on a web browser.

Scikit-learn:

The machine learning model is created using the Python programming language's Scikit-learn module. The library is constructed on top of SciPy (Scientific Python) and incorporates various other libraries, including:

- **Numpy** package for multidimensional arrays.
- **Matplotlib** is a feature-rich framework for 2D and 3D graphical plotting.
- **Pandas** data analysis and structure, etc.

Keras:

Working on top of TensorFlow, Theano5, or Microsoft Cognitive Toolkit, the high-level neural network API named is developed in Python (CNTK). It is simple to use and readily extensible using Python. Constructing a model is simple, and the API comes with pre-trained CNN models that are ready for experimentation, including VGG16 and InceptionV3. It can support a combination of RNN and CNN or a mix of the two [76].

In this stage, a machine-learning model for predicting internal core system user fraud was developed using the pre-processed data. The study's primary objective was to develop supervised machine learning models for comparative analysis and determine the best prediction model. Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and K-nearest neighbor (KNN) algorithms were used to predict internal core system user fraud. Furthermore, Google Colab with GPU hardware accelerator 1.63GB was utilized.

4.3. Experimental Setting

This study conducted four experiments, which were based on earlier research and the linked publications mentioned above. Following the application of various parameters, the optimal hyper-parameters were as follows:

Logistic Regression Parameter (Solver 'lbfgs')

- The choice of solver 'lbfgs' is supported by the scikit-learn documentation for logistic regression.
- 'lbfgs' is suitable for small datasets and is well-suited for multiclass problems. It often converges faster and more reliably compared to other solvers for logistic regression.

Random Forest Parameters:

- Criterion = "gini": This is one of the two criteria supported by scikit-learn for splitting decision trees. It measures the impurity of the nodes.
- n_estimators = 100: This parameter determines the number of trees in the forest. A higher number can lead to better performance, but it increases computational cost.
- max_depth = 5: This parameter controls the maximum depth of the individual trees. A shallow tree helps to prevent overfitting and reduces computational complexity.
- These parameter choices are common defaults and are supported by the scikit-learn documentation. The values strike a balance between model complexity and performance.

Support Vector Machine (Kernel 'linear'):

- The linear kernel is a well-established kernel function supported by scikit-learn for linear SVMs.
- The linear kernel is suitable when the data is linearly separable or when the number of features is large. It often performs well and is computationally efficient.

K-Nearest Neighbor (KNN):

- n_neighbors = 505: The number of neighbors chosen is quite high and might lead to smoothing out of decision boundaries.
- it's essential to choose an odd number to avoid ties in the voting process. It's generally recommended to perform hyperparameter tuning to find the optimal number of neighbors based on cross-validation.
- These parameters are crucial for fine-tuning the KNN algorithm to achieve optimal performance on a given dataset. By systematically experimenting with different parameter settings, can identify the

configuration that yields the best results in terms of accuracy, generalization, and computational efficiency.

4.4. Experiment

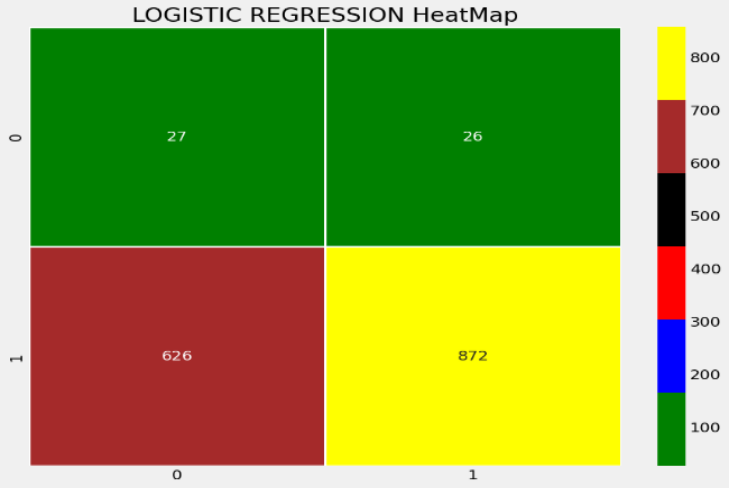
To choose the ideal model for the provided dataset, four supervised machine learning models were applied.

4.4.1. Experiment 1: Logistic Regression (LR) Classification

The Logistic Regression model was created using Python's sklearn.linear model class and its Logistic Regression function. Numerous supervised, and unsupervised learning methods are available through the open-source Scikit Learn, machine learning package for Python. Prior to training and testing the model, all unimportant factors were eliminated and only significant variables were incorporated. All supervised machine learning algorithms employed the same characteristics with the parameters set to default.

The logistic regression results obtained show in the table 4.1.

TABLE 4.1 CONFUSION MATRIX FOR LOGISTIC REGRESSION



The Logistic Regression classification performance metrics, shown in table 4.1, resulted in an average accuracy of 57.96%, precision of 97.10%, recall of 58.21%, and f1 score of 72.79%.

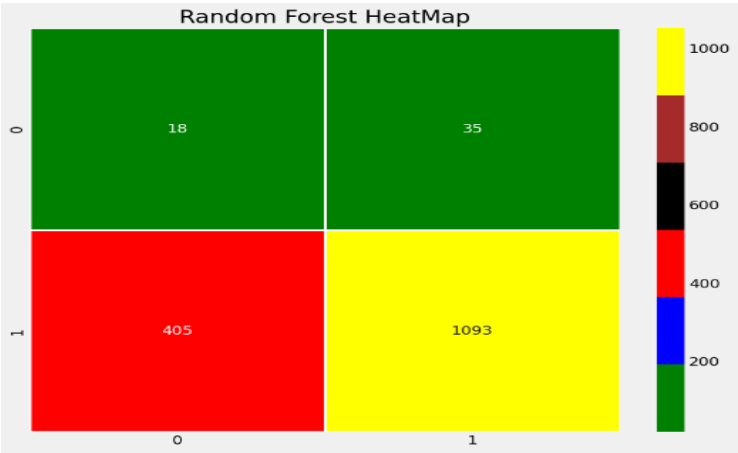
The results of this investigation can be explained by a linear association between the target label and the related attributes. Logistic regression, often known as linear regression in the context of regression problems, has difficulty making accurate predictions when there is no discernible link between the

features and the target label, especially when the training dataset is small. This results from the decision surface of logistic regression's inherent linearity, which makes it inappropriate to use when modeling complex data.

4.4.2. Experiment 2: Random Forest (RF) Classification

The implementation of the Random Forest algorithm in Python involved the use of the sklearn.ensemble class. Ensemble learning is utilized by this algorithm, and this technique relies on multiple machine learning models for achieving accurate predictions on a given dataset. For this particular study, the default function to evaluate the quality of split for this study was selected as the 'gini' criterion. Additionally, it was decided to create a Random Forest by combining 100 decision trees, each chosen randomly. The maximum depth of the trees was set at 10. The class_weight parameter was set to 'balanced', indicating the weight assigned to each class. In Random Forest, class weight is typically based on how frequently the class appears in the data.

TABLE 4.2 CONFUSION MATRIX RANDOM FOREST



The Random Forest classification performance measures as presented in Table 4.2 show that an average of 71.63 percent accuracy, 96.90 percent precision, 72.96 percent recall, and 83.24 percent f1 score are achieved.

The outcome could be explained by the dataset's complexity, which would make it impossible for the Random Forest algorithm to forecast targets in a really accurate manner.

4.4.3. Experiment 3: Support Vector Machine (SVM) Classification

The SVM model in this study was created using the Python's SVM function from sklearn. To achieve efficiency, the model utilized the 'linear' kernel parameter. Choosing an appropriate kernel function is vital for the Support Vector Machine model. For this particular investigation, the kernel function was set as 'linear.' This choice proved beneficial for vast datasets since it can separate classes linearly with a single line. The processing speed is the main benefit of using the linear kernel.

The results were obtained from the Support Vector Machine.

TABLE 4.3 CONFUSION MATRIX FOR SUPPORT VECTOR MACHINE

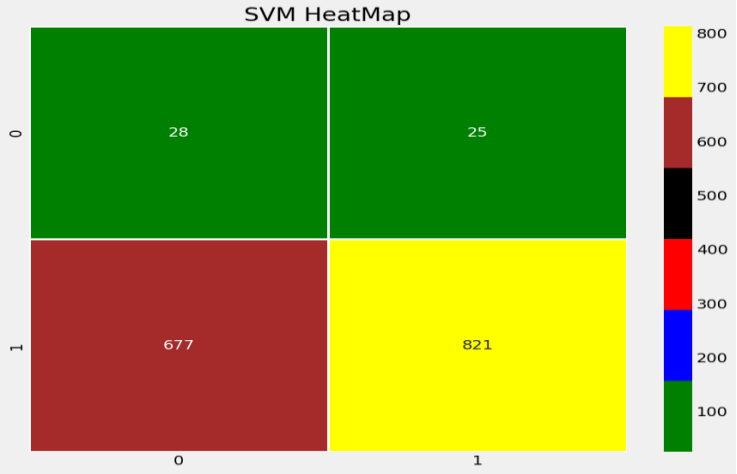


Table 4.3 outlines the average performance attributes of the Support Vector Machine (SVM) classification, which include 54.74% accuracy, 97.04% precision, 54.58% recall, and 70.05% f1 score.

SVM, a widely used machine learning algorithm, exhibits poor performance when dealing with a large number of datasets. In cases where the data is noisier, indicating that there is a lot of overlap between target classes, there is a higher probability of inaccurate predictions, even when using the training set. This suggests that SVM may not be suitable for applications that require high accuracy in such scenarios.

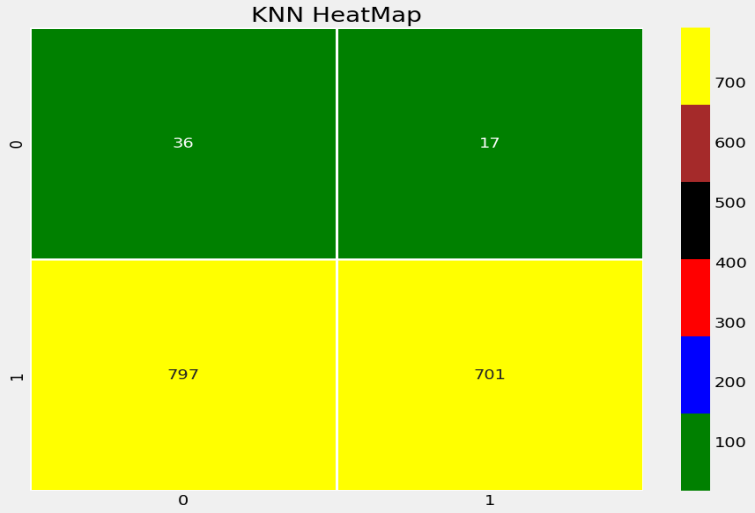
4.4.4. Experiment 4: K-Nearest Neighbor (KNN) Classification

The KNN model was constructed by employing the KNeighborsClassifier function imported from the Python package, sklearn. The K Nearest Neighbor model's most critical parameter, the n-neighbors function, was applied. The KNN model used 6,203 datasets for training with n neighbors determined using the square root of the training datasets, also known as the rule of thumb [54]. The value of K was

set to 79. The simplicity of the KNN algorithm has made it popular. One of the primary advantages of determining the ideal value of K is that it allows us to know how many of the nearest neighbors will be involved in the majority vote process.

The following results were obtained from the K-Nearest Neighbor.

TABLE 4.4 CONFUSION MATRIX FOR K-NEAREST NEIGHBOR



The results of the K-Nearest Neighbor (KNN) classification performance are presented in Table 4.4, wherein an average accuracy of 47.52%, precision of 97.63%, recall of 46.80%, and f1 score of 63.27% are obtained. However, it is crucial to remember that the accuracy of KNN may decrease when dealing with a high number of instances or complex data, as well as when defining the parameter k. In such cases, the predictive capability of KNN may be compromised, leading to lower accuracy of target predictions.

TABLE 4.5 SUPERVISED MACHINE LEARNING MODELS RESULT

Evaluation Metric	Accuracy	Precision	Recall	f1-score	Remark
Logistic Regression	**57.96%**	**97.10%**	**58.21%**	**72.79%**	**2**
Random Forest	**71.63%**	**96.90%**	**72.96%**	**83.24%**	**1**
Support Vector Machine	**54.74%**	**97.04%**	**54.58%**	**70.05%**	**4**
K-Nearest Neighbor	**47.52%**	**97.63%**	**46.80%**	**63.27%**	**3**

Results were collected from each experiment, as indicated in Table 4.5 selected supervised machine learning techniques.

4.5. Answers to the Research Questions

In addressing **Research Question #1**, this study focuses on identifying the pivotal factors and features influencing fraud prediction within internal core system users. The fundamental variables considered for constructing a robust fraud prediction model encompass User_Level, Age, Service_Year, Action_Performed, Profile, Gender, Reference, Operating_Machine, and Fraud. The combination of these components serves as the basis for constructing a reliable model that predicts user fraud within an internal core system. This model provides a valuable understanding of the complexities involved in detecting fraud in this particular context.

In response to **Research Question #2**, the Random Forest algorithm emerges as the preferred choice for predicting fraud among internal core system users within the CBE internal core system user's dataset. After conducting a comparative analysis, Observations have shown that the Random Forest algorithm shows better performance as compared to other machine learning classifiers. It is highly effective in detecting fraudulent activities within the user data of the internal core system.

In addressing **Research Question #3**, which investigates how well the suggested model works in predicting fraud among internal core system users for CBE, the conducted experiments reveal that the Random Forest algorithm exhibits notable performance metrics. The chosen model demonstrates an accuracy rate of 71.63%, coupled with a precision of 96.90%, a recall of 72.96%, and an f1-score of 83.24%. The most effective model for predicting fraudulent activities within the dataset of internal core

system users for CBE is the Random Forest algorithm. Therefore, it is selected as the optimal predictive tool for this specific context based on these results.

4.6. Strengths of the Research

This study's primary strength was its exact identification of fraudulent activity among internal core system users. The findings indicate that the Random Forest algorithm outperformed other selected algorithms in predicting such activity. Furthermore, the study revealed several key predictors of internal core system user fraud of CBE, including User_Level, Age, Service_Year, Action_Performed, Profile, Gender, Reference, and Operating_Machine.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

This section provides an overview of the conducted study that aimed to predict internal core system user fraud. It consists of two primary parts that cover the conclusion, future work, and recommendations. The study's primary objective was to identify the machine learning algorithm that can predict CBE's internal core system user fraud with the highest accuracy, precision, recall, and f1-score among Logistic Regression, Random Forest, Support Vector Machine (SVM), or K-Nearest Neighbor (KNN). The chapter summarizes the experiment's results.

5.1. Conclusion

This research is crucial as it addresses the urgent demand for robust measures against internal user fraud, especially as financial institutions like CBE depend more on digital infrastructure. The expanding threat landscape within internal core systems poses significant risks to data integrity, financial assets, and user trust. The study employs machine learning algorithms to develop predictive models, enabling proactive identification of fraudulent activities among internal core systems users. The data was collected from Kirkos district core system users by simple random sampling technique, and a data pre-processing technique has been performed in this study to make the data clean and make it suitable for machine learning models. To test and train the model, the sample data is divided into 80% for training and 20% for testing. To ensure the features were optimized for machine learning algorithm compatibility, a careful selection process, efficient feature transformation, and feature engineering were used. Additionally, a challenge surfaced during the study: an inherent data imbalance issue. Specifically, a mere 4.4% of the entries reflected internal core system user fraud. To address this imbalance, the study employed the SMOTE method, ensuring a more representative and balanced dataset for the subsequent analysis and modeling processes. Four machine learning algorithms were chosen because of their diversity and applicability in this type of prediction. These algorithms are Logistic regression, Support Vector Machine (SVM), Random Forest, and K-nearest neighbor (KNN) algorithms. The Random Forest algorithm model has an accuracy of 71.63% and performed well across all metrics; based on the result found logistic regression, K-NN, and Support vector machines are in place.

5.2. Future Work and Recommendation

Future work in the field of internal core system user fraud prediction using machine learning algorithms could explore several opportunities to enhance the effectiveness and scope of the research:

- Enabling rapid identification and immediate action against potentially fraudulent activities. Implement a system for continuous monitoring of data streams to adapt the model in real time as user behavior patterns and fraud tactics evolve.
- Here in this study only one district of the CBE dataset was explored and analyzed, in the future, another district of the CBE dataset can be explored, and further research is needed to handle additional variables.
- Investigate and develop more advanced machine learning algorithms or ensemble methods to further improve the accuracy and efficiency of fraud prediction.
- Explore additional relevant features or alternative methods of feature engineering to capture nuanced patterns indicative of fraud. The model is also recommended for training and testing on a large number of data with complex configurations.

Finally, the research provides the following suggestions based on its findings.

- This study used different classification algorithms in which Random Forest performed better than other selected algorithms. So, this study recommends the Random Forest model to build the internal core system users' fraud prediction.
- The internal core system users' data has been analyzed using feature ranking algorithms. The algorithms have identified a basic feature that exhibits strong predictive power, which can be utilized during decision-making processes. This feature can help in combating fraud among internal core system users.
- Facilitate knowledge transfer within the CBE internal staffs to ensure that stakeholders understand the model's functionality and can contribute to its ongoing improvement.
- Explore opportunities for collaboration with other financial institutions facing similar challenges in internal user fraud. Sharing insights and experiences can contribute to a collective effort to combat fraud on a broader scale.

5.3. Contribution of the study

The contribution of this study lies in its focused application on predicting fraud among internal core system users through the implementation of a machine learning algorithm. By honing in on the unique context of internal user activities within core systems, the research aims to enhance the detection and prevention of fraudulent activities originating from within an organization. The utilization of a machine learning algorithm adds a layer of sophistication to the fraud prediction process, allowing for the identification of subtle patterns and anomalies in user behavior that might otherwise go unnoticed. The outcomes of this study are expected to provide valuable insights and tools for organizations to fortify their internal security measures, thereby mitigating the risks associated with fraudulent activities carried out by users within their core systems.

Bibliography

[1] Markowski And Ms. Mannan, “Fuzzy Risk Matrix”, 2008.

[2] Rae & Subramaniam, ""Quality Of Internal Control Procedures: Antecedents And Moderating Effect On Organizational Justice And Employee Fraud",," 2008.

[3] Beirstaker, Brody, and Pacini, “Reasons of Banking Fraud-A case of Indian Public Sector Banks”, *Int. J of Information System,* 2005.

[4] Samuel Ngigi, Samuel Nduati, Peter Kariuki , "“Application of Internal Control System in Fraud Prevention In Banking Sector.”," *IJST,* vol. VOLUME 9, no. ISSUE 03, MARCH 2020..

[5] Zager, L., Malis, S. S. and Novak, "‘The Role and Responsibility of Auditors in Prevention and Detection of Fraudulent Financial Reporting’, The Author(s), 39(November 2015),," *Procedia Economics and Finance.,* no. doi: 10.1016/s2212-5671(16)30291-x, p. pp. 693– 700., (November 2015).

[6] Abayomi, S. O. and Abayomi, S. O., ‘Personal Ethics and Fraudster Motivation: The Missing Link in Fraud Triangle and Fraud Diamond Theories’, *International Journal of Academic Research in Business and Social Sciences, 6(2),,* p. pp. 159–165. doi: 10.6007, (2016).

[7] “Occupational Fraud A Report To The Nation”, Association Of Certified Fraud Examiners (Acfe),, 2022.

[8] Adeyemo Kingsley, “Frauds In Nigerian Banks”," no. Department Of Accounting Covenant University,, May 2012..

[9] "“Report To The Nations On Occupational Fraud And Abuse”," *Association Of Certified Fraud Examiners (Acfe),,* no. Association Of Certified Fraud Examiners (Acfe), 2016.

[10] Lang, Matthias; Wambach, Achim,, " “The Fog Of Fraud -Mitigating Fraud By Strategic Ambiguity”," *Max Planck Institute For Research,,* May 2010..

[11] W. Steve Albrecht, Chad O. Albrecht, Conan C. Albrecht, , "“Fraud Examination”,," no. Brigham Young University,, 2011.

[12] Silverstone, H. & Davia, H, "Fraud 101: Techniques And Strategies For Detection.," vol. 2nd Edition. London: , no. John Wiley And Sons. , (2005).

[13] Albrecht, S., And Albrecht, C. , "Fraud Examination And Prevention.," no. Ohio: Thomson South Western., 2004.

[14] L. Fausett, " Fundamentals Of Neural Networks: Architectures, Algorithms And Applications,," no. Prentice-Hall, New Jersey, Usa, (1994)..

[15] Phua, C., Lee, V., Smith, K., & Gayler, R., " A Comprehensive Survey Of Data Mining-Based Fraud Detection Research.," vol. Arxiv Preprint Arxiv:1009.6119. , (2010)..

[16] C. Deng, X. Ji, C. Rainey, J. Zhang, and W. Lu,, " “Integrating Machine Learning with Human Knowledge,”," *iScience, ,* Vols. vol. 23, no. 11, , no. doi: 10.1016/j.isci.2020.101656., pp. p. 101656,, 2020.

[17] L. G. Kabari, D. N. Nanwin, And E. U. Nquoh, " “Telecommunications Subscription Fraud Detection Using Artificial Neural Networks,” Transactions On Machine Learning And Artificial Intelligence,," Vols. Vol. 3, No. 6, , p. P. 19., (2016),.

[18] A. Onibudo, "Bank Frauds Problems And Solutions,," *B.Sc. Research Project, ,* no. University Of Benin, Nigeria, 2007.

[19] D. R. Pankhurst, " “01. Early Ethiopian Banking History,” no. Early Ethiopian Banking History, [Online]. Available," [Online]. Available: https://www.linkethiopia.org/article/2-the-bank-of abyssinias-bank-notes/..

[20] Tariku, " A. Mining Insurance Data For Fraud Detection: The Case Of Africa Insurance Share Company.," *Aau, Faculty Of Informatics, Department Of Information Science.,* 2011.

[21] Al Marri, Matar And Alali, Ahmad,, " "Financial Fraud Detection Using Machine Learning Techniques"," 2020.

[22] A. Cutler, D. R. Cutler, And J. R. Stevens,, " “Ensemble Machine Learning,” Ensemble Mach. Learn.,," no. Doi: 10.1007/978-1-4419-9326-7., 2012.

[23] Khalifa Alsenaani, "“Fraud Detection In Financial Services Using Machine Learning”," 2022.

[24] M. R. Albougha, "“Comparing Data Mining Classification Algorithms In Detection Of Sim-Box Fraud,”," *St. Cloud State University The Repository At St. Cloud State.,* (2016),.

[25] S. Pan, T. Morris, And U. Adhikari , "“Developing A Hybrid Intrusion Detection System Using Data Mining For Power Systems,”," *Ieee Trans. Smart Grid, ,* Vols. Vol. 6, No 1, (2015), .

[26] M. ,. AGWU, "“REPUTATIONAL RISK IMPACT OF INTERNAL FRAUDS ON BANK CUSTOMERS IN NIGERIA”,," *International Journal of Development and Management Review (INJODEMAR) ,* Vols. Vol. 9, No 1,, June 2014..

[27] G. Williams, "“Data Mining With Rattle And R”:," *The Art Of Excavating Data For Knowledge Discovery. Springer Science & Business Media ,* 2011.

[28] Oraka, A.O., Ph.D., Egbunike, F.C. , "“Corporate Fraud And Performance Of Micro Finance Banks In Nigeria”," Department Of Accountancy, Nnamdi Azikiwe University,, Vols. Vol. 5, No. 4,, August 2016,.

[29] Leah Njeri Kabue, ""“The Effect Of Internal Controls On Fraud Detection And Prevention Among Commercial Bank In Kenya”," *School Of Business, University Of Nairobi,,* 2015.

[30] Aeran, "Comprehensive Overview Of Insider Threats And Their Controls.," *Royal Holloway,* 2006.

[31] G. L. Tang, "Trusted Computing: Addressing Insider Threats To The Banking And Financial Sector," (2005).

[32] Burke, B.E. & Christiansen, C.A.,, " Insider Risk Management: A Framework Approach To Internal Security. “Rsa, The Security Division Of Emc," *Rsa, The Security Division Of Emc,* 2009.

[33] R. Richardson, " Computer Crime And Security Survey,”," *Computer Security Institute,,* 2007.

[34] Willison, R., And Warkentin, M., “An Expanded View Of Employee Computer Abuse,” in *Beyond Deterrence*, 2013, p. Pp.1–20.

[35] ""“What Is Banking?,”," 2021. [Online]. Available: https://www.thebalance.com/what-isbanking-3305812..

[36] Qimin Cao , Yinrong Qiao,, " "Machine Learning To Detect Anomalies In Web Log Analysis,"," in *In 3rd Ieee International Conference On Computer And Communications,*, Shanghai, China ,, 2017.

[37] K. Kanwal, S. Ahmad, And A. S. Malik , "Fraud Detection Using Data Analytics And Machine Learning Techniques"., 2012.

[38] Zhang Et Al., "Fraud Detection In The Banking Sector Using Machine Learning", 2020.

[39] "N. Bank and S.-S. Africa, “Governor ’ s Note,”," 2010.

[40] H. Jiawei & M. Kamber, Data Mining: Concepts And Techniques., San Francisco: San Francisco, Ca, Itd: Morgan Kaufmann., 2001.

[41] ""“No Title,” no. COMMERCIAL BANK OF ETHIOPIA (CBE) ISSUED AWARDS, [Online]. Available:," 2021. [Online]. Available: https://uptimeinstitute.com/uptime-instituteawards/client/commercial-bank-of-ethiopia-cbe/592..

[42] H. Tesfaye, ” Constructing A Predictive Model For Subscription Fraud Detection Using Data Mining Technique For The Case Of Ethio Telecom”,, *Unpublished Master Thesis Department Of Information Science Aau, Ethiopia (2013).,* 2013.

[43] I. H. Sarker, " “Machine Learning: Algorithms, Real-World Applications and Research Directions,”," *SN Comput. Sci.,,* Vols. vol. 2, no. 3,, no. doi: 10.1007/s42979-021- 00592-x., pp. pp. 1–21,, 2021,.

[44] S. Pirzada, “Machine Learning and Logistic Regression Umme Salma,” *Mach. Learn. Algorithms Logist. Regres.,,* May, 2020..

[45] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, “A internal user Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for internal user Prediction and Factor Identification in Telecom Sector,” *IEEE Access, ,* Vols. vol. 7,, no. doi: 10.1109/ACCESS.2019.2914999., pp. pp. 60134–60149,, 2019.

[46] Z. Jin, J. Shang, Q. Zhu, C. Ling, W. Xie, and B. Qiang, , "Forecasting Turkish Lira (TRY)/US Dollar (USD) Interest Exchange Rates Using Machine Learning Methodologies," *Journal of Soft Computing and Artificial Intelligence, ,* Vols. vol. 12343 , no. 2, , pp. pp. 120 - 131,, 2020.

[47] Saima Saleem, Muhammad Sheeraz, Muhammad Hanif and Umar Farooq,, ""Web Server Attack Detection using Machine Learning,"," *Pakistan Institute of Engineering and Applied Sciences , Islamabad, ,* 2021.

[48] Gustavo Betarte, Rodrigo Martınez and Alvaro Pardo,, ""Web Application Attacks Detection Using Machine Learning Techniques,"," in *in 17th IEEE International Conference on Machine Learning and Applications,*, Uruguay,, 2018..

[49] R.Ravinder Reddy ,B.Kavya & Y Ramadevi,, ""A Survey on SVM Classifiers for Intrusion Detection,"," *International Journal of Computer Applications (0975 – 8887), ,* Vols. vol. 98, no. 19,, pp. pp. 38-44, , 2014..

[50] M. Sun, "“Support Vector Machine Models For Classification,”," *Encycl. Bus. Anal. Optim., ,* no. Doi: 10.4018/978-1-4666-5202-6.Ch215., pp. Pp. 2395–2409,, 2014,.

[51] K. J. Kim, "“Financial Time Series Forecasting Using Support Vector Machines,”," *Neurocomputing, ,* Vols. Vol. 55, No. 1–2, , no. Doi: 10.1016/S0925- 2312(03)00372-2, pp. Pp. 307–319, , 2003,.

[52] Breiman, "“Random Forests,”," *Machine Learning, ,* Vols. Vol. 45, No. 1, , p. Pp. 5–32., (2001).

[53] K. Mishra, S. V. Ramteke, P. Sen, And A. K. Verma, , "“Random Forest Tree-Based Approach For Blast Design In Surface Mine,”," *Geotechnical And Geological Engineering, ,* Vols. Vol. 36, No. 3,, p. Pp. 1647–1664., (2017).

[54] Z. Zhang, " “Introduction To Machine Learning: K-Nearest Neighbors,”," *Ann. Transl. Med.,,* Vols. Vol. 4, , no. Doi: 10.21037/Atm.2016.03.37., pp. P. 218,, Jun. 2016,.

[55] J. T. Point, " "K-Nearest Neighbor(KNN) Algorithm for Machine Learning," Java T Point,," 06 08 2021. [Online]. Available: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning. [Accessed 06 08 2022].

[56] A. Ali and S. O. S. E. T. A.-D. A. N. M. E. T. H. S. A. Abd Razak, "Financial Fraud Detection Based on Machine Learning:," *A Systematic Literature Review. ,* no. Appl. Sci. 12, 9637. https://doi.org/10.3390/ app1, 2022.

[57] Ali Moradi Vartouni , Saeed Sedighian Kashi and Mohammad Teshnehlab ,, " "An Anomaly Detection Method to Detect Web Attacks Using Stacked Auto-Encoder,"," *in 6th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS), Tehran, Iran, ,* 2018..

[58] T. Haddish, ""Constructing Predictive Model for Subscription Fraud Detection Using Data Mining Techniques,"," *Diss,* no. Diss. AAU, 2013., 2013.

[59] Jonathan K, Kassim T., and Wilhemina A., ""“A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions”.," (2023).

[60] Kha S. L., Lam H. and Yee-Wai S.,, " “A Review of Machine Learning Algorithms for Fraud Detection in Credit Card Transaction”,," *IJCSNS International Journal of Computer Science and Network Security,,* Vols. Vol.21 No.9, , September 2021..

[61] Bushra B., and Najwa A.,, ""“Insider Threat Detection Using Machine Learning Approach”,," *Appl. Sci. 2023, 13(1), 259;,* no. https://doi.org/10.3390/app13010259., 2023.

[62] J. Creswell, "Research Design:," in *Research Design: Qualitative, Quantitative And Mixed Method Approaches (3rd Ed.).* , Los Angeles:, Sage Publications. , 2009.

[63] El Arass, Mohammed & Khadija, Ouazzani Touhami & Souissi, Nissrine., " Data Life Cycle: Towards A Reference Architecture.," *International Journal Of Advanced Trends In Computer Science And Engineering. ,* Vols. Vol. 9, , no. 10.30534/Ijatcse/2020/215, pp. Pp. 5645 - 5653. , 2020.

[64] V. U. Library, ""Research Methods Guide: Research Design & Method," Carol M. Newman Library, Virginia Tech, 21 09 2018.," 21 09 2018. [Online]. Available: Https://Guides.Lib.Vt.Edu/Researchmethods/Design-Method.. [Accessed 21 10 2018].

[65] R. Sebastian, "Model Evaluation, Model Selection, And Algorithm Selection In Machine Learning.," *University Of Wisconsin–Madison Department Of Statistics.,* 2018.

[66] A. R. Hevner, S. T. March, J. Park, and S. Ram,, " “Design science in information systems research,”," *MIS Q. Manag. Inf. Syst.,* Vols. vol. 28, no. 1,, no. doi: 10.2307/25148625., pp. pp. 75–105,, 2004.

[67] A. Hevner and S. Chatterjee, “Design Science Research in Information Systems,” *in Management Information Systems Quarterly - MISQ, ,* Vols. vol. 28, 2010, , p. pp. 9–22..

[68] Y. Roh, G. Heo, and S. E. Whang,, "A Survey on Data Collection for Machine Learning: A Big Data-AI Integration Perspective," *IEEE Trans. Knowl. Data Eng., ,* Vols. vol. 33, no. 4,, no. doi: 10.1109/TKDE.2019.2946162., pp. pp. 1328–1347,, 2021.

[69] S. B. Kotsiantis and D. Kanellopoulos, , "Data preprocessing for supervised leaning,", *Int. J. ...,,* Vols. vol. 1, no. 2, , pp. pp. 1–7,, 2006..

[70] D. Z. Abidin, S. Nurmaini, R. Firsandava Malik, Erwin, E. Rasywir, and Y. Pratama,, ""RSSI Data Preparation for Machine Learning,"," in *Proc. - 2nd Int. Conf. Informatics, Multimedia, Cyber, Inf. Syst. ICIMCIS 2020,* , none, 2020.

[71] M. Quintero and A. LeBoulluec, , ""Missing Data Imputation for Ordinal Data,"," *Int. J. Comput. Appl., ,* Vols. vol. 181, no. 5, , no. Int. J. Comput. Appl., , pp. pp. 10–16, , 2018. .

[72] H. Kang, ""The prevention and handling of the missing data,"," *Korean J. Anesthesiol.,,* Vols. vol. 64, no. 5,, no. doi: 10.4097/kjae.2013.64.5.402., pp. pp. 402–406,, 2013,.

[73] Patrick S., and Christa B., " "Correlation Coefficients: Appropriate Use and Interpretation"," Volume 126, Number. 2, (May 2018)..

[74] E. Rosenzweig, "Successful User Experience: Strategies And Roadmaps," Success. User Exp. Strateg. Roadmaps, , no. Doi: 10.1016/C2013-0-19353-1., p. Pp. 1–344, 2015.

[75] N. J. Nilsson, "INTRODUCTION TO MACHINE LEARNING AN EARLY DRAFT OF A PROPOSED TEXTBOOK Department of Computer Science,", Vols. vol. 56, no. 2, , Mach. Learn., , 2005, pp. pp. 387–99, .

[76] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," Oper. Syst. Des. Implementation, OSDI 2016," in Proc. 12th USENIX Symp., May 2016.

Appendix I

Results of the Algorithms

1. Logistic Regression (LR) Result

	precision	recall	f1-score	support
0	0.04	0.51	0.08	53
1	0.97	0.58	0.73	1498
accuracy			0.58	1551
macro avg	0.51	0.55	0.40	1551
weighted avg	0.94	0.58	0.71	1551

2. Random Forest (RF) Result

	precision	recall	f1-score	support
0	0.04	0.34	0.08	53
1	0.97	0.73	0.83	1498
accuracy			0.72	1551
macro avg	0.51	0.53	0.45	1551
weighted avg	0.94	0.72	0.81	1551

3. K-Nearest Neighbor (KNN) Result

	precision	recall	f1-score	support
0	0.04	0.68	0.08	53
1	0.98	0.47	0.63	1498
accuracy			0.48	1551
macro avg	0.51	0.57	0.36	1551
weighted avg	0.94	0.48	0.61	1551

4. Support Vector Machine (SVM) Result

	precision	recall	f1-score	support
0	0.04	0.53	0.07	53
1	0.97	0.55	0.70	1498
accuracy			0.55	1551
macro avg	0.51	0.54	0.39	1551
weighted avg	0.94	0.55	0.68	1551

Appendix II
Random forest Model Python Code

```
forest = RandomForestClassifier(n_estimators=100, criterion="gini", max_depth=5)
forest.fit(X_balance, y_balance)
forest_pred = forest.predict(X_test)
print("Val Accuracy: ", accuracy_score(forest_pred, y_test))
print("Train Accuracy: ", accuracy_score(forest.predict(X_balance), y_balance))

aa = forest.predict(X_test)
print(aa)
# unscaling the ypred values
aa_lis = []
for i in aa:
   if i>0.5:
      aa_lis.append(1)
   else:
      aa_lis.append(0)
print(aa_lis)
cmm = confusion_matrix(y_test, aa_lis)
#print('Accuracy random forest : {:.2f} %'.format(forest_pred.score(X_test,y_test)*100))
modello = 'Random Forest HeatMap'
cmap = ['green', 'blue', 'red', 'black','brown', 'yellow']
plt.figure(figsize= (10,8))
plt.title(modello)
sns.heatmap(cmm, annot=True,fmt='.7g',cmap=cmap ,cbar= True, robust=True ,linewidths=2.0)

data = {'orignal_fraud':y_test, 'predicted_fraud':forest_pred}
df_check = pd.DataFrame(data)
df_check.tail(20)


print(confusion_matrix(y_test,forest_pred))
print(classification_report(y_test,forest_pred))
plt.savefig('cff.RF.png', bbox_inches='tight', pad_inches=0.0)


print(conf_mat)
print(confusion_matrix(y_test,forest_pred))
print(classification_report(y_test,forest_pred))
plt.savefig('cff.RF.png', bbox_inches='tight', pad_inches=0.0)

print ("Accuracy:", metrics.accuracy_score(y_test, ypred_lis))
print ("Precision:", metrics.precision_score(y_test, ypred_lis, average='weighted'))
rint ("Recall:", metrics.recall_score_score(y_test, ypred_lis, average='weighted'))
print ("F1 Score:", metrics.f1_score_score(y_test, ypred_lis, average='weighted'))
```