Democratic and People's Republic of Algeria

Ministry of Higher Education and Scientific Research

University of 8th May 1945 – Guelma -

Faculty of Mathematics, Computer Science and Material Sciences

Computer Science department

Master's Thesis

Speciality: Computer Science

Option: Information and Communication Science and Technology

SIMBox Fraud Detection in Telecommunication Operators: A Case Study on "Djezzy"

Presented by:

Ahmed Rami Bouguettoucha

Jury Members:

Dr. Karima Benhamza	*Chairwoman*
Dr. Aicha Aggoune	*Supervisor*
Dr. Ali Seridi	*Examiner*
Dr. Ghania Berkat	*Incubator Representative*

Socio-Economic Sector Representative

July 1st, 2024

Acknowledgement

Above all, we thank **ALLAH**, the Most Merciful, the One, the Powerful. Master of the heavens and the earth for having guided, protected, helped and enabled us to carry out this work.

It is with great pleasure that we reserve these lines as a sign of gratitude and recognition to those who contributed directly or indirectly to the development of this work in particular: To my supervisor **Dr.Aggoune Aicha**, please accept the expression of my most sincere thanks for your precious help and support that you have given me, your advice, your favorable comments and your efforts.

I would like to express my most sincere thanks to the people **Mohamed Yacine Bechiri**, **Radhouane Chaiere** and **Abderrahmane Zebboudj** who gave me their help and who contributed to the development and success of this dissertation.

I also thank my **Computer Science teachers** at the **University of Guelma** each in his own name.

Dedications

We thank **Almighty Allah** for giving us the will and courage to carry out this work.

To my dear mother **Fatiha Cherifi** No dedication can express my respect, my eternal love and my consideration for the sacrifices you have made for my education and my well-being. I thank you for all the support and love you have given me since my childhood and I hope that your blessing always accompanies me. May this modest work be the fulfillment of your long-expressed wishes, the fruit of your countless sacrifices. May God grant you health, happiness and long life.

To my dearest father **Farid**, you have been an exemplary father, friend, and advisor throughout my life. Your prayers have greatly supported me during this long journey. May God, Almighty, keep you and grant you health, happiness, and long life so that you remain the light illuminating my path.

To my dear brothers **Abd El Mounaim**, **Chihab**, and my dear sister **Hadjer**, words are hardly enough to express the attachment, love, affection, and respect that I have for you. As a testimony of my fraternal affection, deep tenderness, and gratitude, I wish you a life full of happiness and success.

Special thanks to my friend **Mohamed Yacine Bechiri**, who has always encouraged and helped me in my work, to whom I wish more success. And to my dear friends: **Anes**, **Abd El Basset**, **Abd El Kader**, **Abd El Wadoud**, and **Akrem**, in memory of our bursts of laughter and good times, in memory of everything we experienced together, I hope with all my heart that our friendship will last forever.

Ahmed Rami Bouguettoucha

Abstract

The SIMBox is a telecommunications device that reroutes international voice calls through the internet to a local SIM card within the device, making them appear as local calls and bypassing standard network gateways. This Master's thesis presents a SIMBox Fraud Detector, an AI-powered web application for SIMBox fraud detection in mobile telecommunications networks. A thorough experiments with machine learning techniques using the dataset CDR (Call Detail Records) from Djezzy mobile operator show that the XGBoost model with Tomek+RUS undersampling exhibits promising advantages in both detection and balance performance. XGBoost not only achieved a precision and accuracy of 96% for fraudulent detection, but it also provided the best tradeoff between false positive rate and true positive rate.

Keywords: SIM Box Fraud, SIM Box Fraud Detection, Data Balancing, Call Detail Record, Machine Learning

Résumé

La SIMBox est un appareil de télécommunications qui redirige les appels vocaux internationaux via Internet vers une carte SIM locale située dans l'appareil, les faisant apparaître comme des appels locaux et contournant les passerelles réseau standard. Ce mémoire présente un SIMBox Fraud Detector, une application Web basée sur l'IA pour la détection des fraudes SIMBox dans les réseaux de télécommunications mobiles. Des expériences approfondies avec des techniques d'apprentissage automatique utilisant l'ensemble de données CDR (Call Detail Records) de l'opérateur mobile Djezzy montrent que le modèle XGBoost avec sous-échantillonnage Tomek+RUS présente des avantages prometteurs en termes de performances de détection et d'équilibrage. XGBoost a non seulement atteint une précision et d'exactitude de 96% pour la détection frauduleuse, mais il a également fourni le meilleur compromis entre le taux de faux positifs et le taux de vrais positifs.

Keywords: Fraude liée à la boîte SIM, Détection de fraude liée à la boîte SIM, Équilibrage des données, Enregistrement des détails des appels, Apprentissage Automatique

الملخص

صندوق بطاقات سيم هو جهاز اتصالات يقوم بإعادة توجيه المكالمات الصوتية الدولية عبر الإنترنت إلى بطاقة سيم محلية داخل الجهاز، مما يجعلها تظهر كمكالمات محلية وتجاوز بوابات الشبكة القياسية. تقدم مذكرة الماستر هذه جهاز كشف احتيال صندوق بطاقات السيم ، وهو تطبيق ويب مدعوم بالذكاء الاصطناعي للكشف عن الاحتيال صندوق بطاقات السيم في شبكات الاتصالات المتنقلة. أظهرت تجارب شاملة مع تقنيات التعلم الآلي باستخدام مجموعة البيانات CDR (سجلات تفاصيل المكالمات) من مشغل الهاتف المحمول جيزي أن نموذج XGBoost مع RUS Tomek يُظهر مزايا واعدة في كل من أداء الكشف والتوازن. لم يحقق XGBoost دقة تصل إلى 96% للكشف عن الاحتيال فحسب، بل قدم أيضًا أفضل توافق بين المعدل الإيجابي الكاذب والمعدل الإيجابي الحقيقي.

كلمات مفتاحية: احتيال صندوق بطاقات السيم ، كشف احتيال صندوق بطاقات السيم ، موازنة البيانات، تسجيل تفاصيل المكالمات، التعلم الآلي

Contents

List of Figures

List of Tables

General Introduction

Mobile telecommunications have become an integral part of modern society, facilitating communication and connectivity on a global scale. However, with the increasing reliance on mobile networks comes the challenge of safeguarding these systems against fraudulent activities, which can have significant financial implications for telecommunications companies. One such form of fraud is SIM box fraud, a sophisticated scheme that involves the illegal routing of international calls through SIM boxes to bypass legitimate telecommunication networks.
In this Master's thesis, we make several key contributions to the field of fraud detection in mobile telecommunications:

- We provide a comprehensive overview of the structure and operations of mobile telephony networks, emphasizing the challenges posed by SIM box fraud.
- We review and analyze various existing fraud detection techniques, highlighting their strengths and limitations in the context of detecting SIM box fraud.
- We implement AI-based techniques for SIM box fraud detection, detailing the data preprocessing, model training, and evaluation processes.
- We present experimental results demonstrating the efficacy of our AI-based detection methods, providing insights into their potential for practical application in the telecommunications industry.
- We emphasize the importance of continued research and collaboration to advance fraud detection technologies.

The remainder of this manuscript is organized as follows:

Chapter 1: **Mobile Telecommunications**, provides an overview of mobile telecommunications, including the structure of mobile telephony networks and the architecture of GSM (Global System for Mobile Communications). Additionally, it introduces the concept of SIM box fraud, highlighting its definition, the challenges it poses to telecommunications companies, and its impact on the industry. The chapter also presents a case study of Djezzy, a mobile operator, and discusses its organizational structure, missions, and the digital transformation initiatives undertaken to address emerging challenges.

Chapter 2: **Techniques of SIM Box Fraud Detection**, delves into the techniques used for SIM box fraud detection. It outlines the challenges associated with detecting SIM box fraud, such as the ongoing evolution of fraudulent tactics and the need for fast and accurate detection methods. The chapter explores various fraud detection techniques, including machine learning algorithms and deep learning approaches. It also discusses data balancing techniques utilized to address class imbalances in fraud datasets. Moreover, recent related work in the field of SIM box fraud detection is examined to provide context for the proposed methodology.

Chapter 3: **Implementation And Experimental Results**, focuses on the implementation and experimental results of SIM box fraud detection using artificial intelligence (AI) techniques. It details the experimental setup, including the software tools and libraries used for data preprocessing, feature engineering, and model training. The chapter presents the process of preparing the call detail record (CDR) data for analysis, including data preprocessing steps such as data conversion, feature selection, and outlier removal. It then describes the methodology for model training and selection, highlighting the performance metrics used for evaluating model performance.

The concluding chapters of this Master's thesis offer insights into the application of AI for SIM box fraud detection and discuss the implications of the findings for the telecommunications industry. By leveraging advanced AI techniques, this research aims to contribute to the development of robust fraud detection systems that can effectively mitigate the risks associated with SIM box fraud.

Chapter 1

Mobile Telecommunications

1.1 Introduction

Telecommunications, often abbreviated as telecom, is the electrical transmission of speech, data, and video signals over vast distances. This phrase refers to a broad range of communication infrastructures and technologies, including telegraph networks, optical fibers, microwave links, satellites, radio and TV transmission, conventional wired telephones, and the internet [64].

Many service providers cater to the telecommunications industry, offering everything from current internet and wide area network (WAN) services to old phone services, as well as metropolitan and international connectivity. Although these providers were formerly frequently state-run, privatization has gained traction. Despite the move towards privatization, the telecommunications industry remains subject to regulation by national and international bodies to ensure fair practices, protect consumers, and maintain standards.

This chapter provides an overview of mobile telecommunications, covering the structure and functionality of mobile networks, GSM architecture, and differences between traditional cellular networks and VoIP. It examines call routing, international call handling, and financial aspects, with a case study of Djezzy illustrating practical applications. Fraud in telecommunications, especially SIM boxing, is discussed. The chapter concludes by summarizing key points and their significance in mobile telecommunications.

1.2 Mobile Telephony Networks

Mobile telecommunication services means wireless communication services carried on between mobile stations. Two stations, each with a transmitter and a receiver, or a combination device known as a transceiver, make up a standard telecommunications circuit. Electrical lines (copper cables), optical fibres, electromagnetic fields, and light are some of the media via which signals can be sent. Wireless communications, on the other hand, refers to the transfer and receipt of data using electromagnetic fields without the requirement for a physical conductor [64].

1.2.1 Calls in cellular networks

Call routing, a feature of second-generation (2G) GSM networks, involves creating a direct circuit connection between callers. Even though newer wireless technologies provide a plethora of high-speed data services, they continue to depend on the circuit-switched framework of 2G to manage phone calls [29].

The Mobile Equipment, identified by its distinct International Mobile Equipment Identity (IMEI), serves as the mobile user's gateway to the cellular network. A Subscriber Identity Module (SIM) card is required in order to make use of the network services provided by a provider. The SIM card's unique International Mobile Subscriber Identity (IMSI) is used by the network to identify it, and it also contains an encryption key that the provider has provided to secure communications. [29]

GSM (Global System for Mobile communications) is a standard outlining the protocols for second-generation (2G) cellular digital networks. These 2G networks were introduced as an advancement over the first-generation (1G) analog networks, leading the way for further developments in mobile technology, including 2.5G, 2.75G, 3G, and 4G LTE [4].

The primary distinction between 1G and 2G networks is the transition to digital technology with 2G, particularly beyond the base station, along with digital encryption for enhanced security. Initially, 2G networks were designed to support voice communications only, operating on a circuit-switched basis. The introduction of 2.5G, or GPRS (General Packet Radio Service), marked the beginning of data support in the cellular network technology, providing theoretical data transfer speeds of up to 50 Kbps, although practical speeds are around 40 Kbps [4].

1.2.2 GSM Architecture

The GSM network is comprised of two main components [4]: the BSS (Base Station Subsystem) and the NSS (Network Switching Subsystem). To fully grasp the structure and functionality of the GSM network as depicted in Figure 1.1, it is essential to first understand what these abbreviations stand for. Let's examine each of these subsystems in detail.

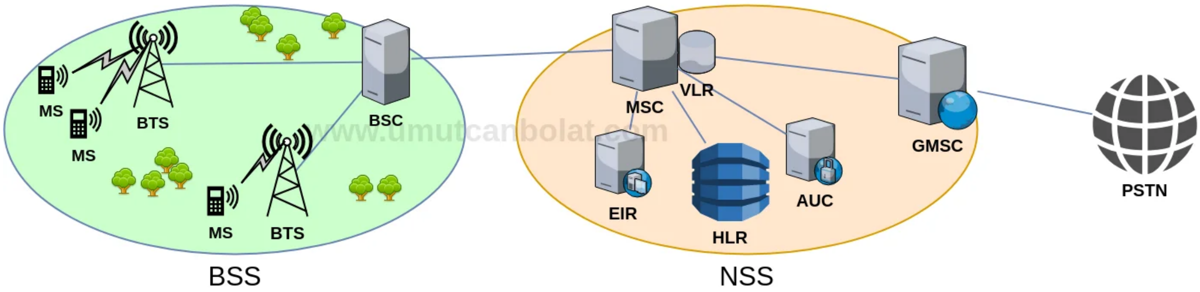


FIGURE 1.1: GSM Architecture [4]

SIM card: Which first appeared alongside 2G GSM technology, are an abbreviation for "Subscriber Identity Module." They are designed to store critical information such as the IMSI (International Mobile Subscriber Identity), a unique code that identifies the subscriber. In addition to the IMSI, SIM cards contain security authentication and ciphering information, along with a list of services available to the subscriber. Now, let's delve into the specifics of these acronyms and their significance [9].

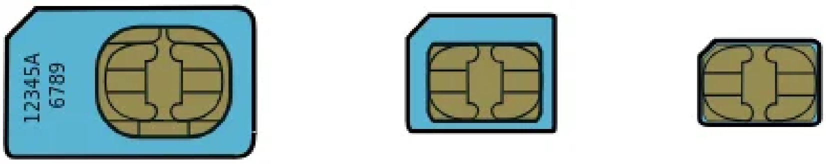


FIGURE 1.2: Mini, Micro and Nano SIM [9]

Mobile Station (MS) MS, or Mobile Station, refers to the user equipment in the context of GSM networks. This includes devices like cellphones, mobile computers, or any other device equipped with a SIM card and the necessary software to communicate with the GSM network. In the evolution of mobile telecommunications, when moving to 3G systems, the terminology shifts from MS to UE (User Equipment) to describe these devices [4].

Base Transreceiver Station (BTS): The BTS, or Base Transceiver Station, is critical equipment in the cellular network infrastructure, facilitating the transmission and reception of radio signals between the Mobile Station (MS) and the network. BTS units lay the groundwork for the cellular network's structure by creating cells, within which mobile devices can communicate with the broader network. These BTS units are interconnected with BSCs (Base Station Controllers), which are responsible for controlling and managing the BTS operations.

Base Station Controller (BSC): The BSC, or Base Station Controller, serves as the controlling intelligence for multiple BTSs (Base Transceiver Stations). It plays a pivotal role in managing the radio resources of a cellular network, including the allocation of radio frequencies, power management, and signal measurements. Furthermore, the BSC is responsible for orchestrating handovers, ensuring seamless transitions for mobile devices moving from one cell to another within its jurisdiction. This capability is crucial for maintaining continuous communication as users travel across different areas covered by the same BSC. [4]

Mobile Switching Center (MSC): The MSC, or Mobile Switching Center, is the central component of the Network Switching Subsystem (NSS) in a cellular network. Its primary functions include routing voice calls and SMS messages, establishing end-to-end circuit-switched connections between subscribers. Beyond these core tasks, the MSC is instrumental in managing key mobile services, such as subscriber registration, authentication, and location updates, ensuring that users are accurately identified and can seamlessly access network services as they move within or across network areas.

Gateway MSC (GMSC): The GMSC, or Gateway Mobile Switching Center, represents a specialized form of MSC (Mobile Switching Center) tasked with facilitating call routing to and from the mobile network and external networks, such as the Public Switched Telephone Network (PSTN). When a call is made to a mobile subscriber from outside the mobile network, or when a mobile subscriber initiates a call to a recipient outside the mobile network, the GMSC serves as the pivotal point through which these calls are routed. This enables seamless communication between the mobile network subscribers and the wider telecommunication network beyond the mobile system. [4]

Home Location Register (HLR): The HLR, or Home Location Register, stands as the most critical database within a carrier's network infrastructure. It is a comprehensive repository that stores extensive information about subscribers, including the GSM services they utilize and their current locations. Essentially, the HLR maintains detailed records for each SIM card issued by the carrier, ensuring that the network can efficiently manage and support the service requirements and connectivity of its subscribers. [4]

Visitor Location Register (VLR): The VLR, or Visitor Location Register, functions as a database similar to the HLR (Home Location Register), but it is designed to store temporary information about subscribers that might be required by the MSC (Mobile Switching Center). The VLR is always integrated with an MSC. Whenever a Mobile Station (MS) enters a new area covered by an MSC, the associated VLR will request the subscriber's information from the HLR. This setup ensures that when the MS initiates a call, the MSC will consult the VLR for the necessary information before reaching out to the HLR. By doing so, the VLR effectively reduces the workload and number of queries directed to the HLR, optimizing the overall efficiency of the network's operation. [4]

Equipment Identity Register (EIR): The EIR, or Equipment Identity Register, is a crucial database in a mobile network that maintains a list of Mobile Stations (MS) based on their eligibility to access the network or their prohibition from it. This system plays a vital role in monitoring and controlling the use of the network, especially in cases of lost or stolen devices. Mobile Stations are uniquely identified by their IMEI (International Mobile Equipment Identity) numbers, which allows for precise tracking and management of devices in relation to network access. The EIR ensures that only authorized devices can connect to the network, enhancing security and protecting against fraudulent use. [4]

Authentication Center (AuC): The AuC, or Authentication Center, is a critical security function within the GSM network tasked with authenticating mobile subscribers attempting to connect to the network. This process involves identifying and verifying the validity of the subscriber's SIM card. By ensuring that the SIM card is legitimate, the AuC plays a key role in safeguarding the network against unauthorized access, thereby maintaining the integrity and security of the communication services provided to legitimate users. [4]

1.2.3 Call Scenario

In the following, we will explore a typical scenario illustrating how a call is processed within a GSM network. This walkthrough aims to demystify the complex orchestration of network components that collaborate to enable seamless communication between two parties. Understanding this process sheds light on the intricacies of mobile telecommunications and the critical role of each network element. Let's examine the sequence of steps from the moment a call is initiated to its conclusion.

1. **Initiation:** The calling party initiates a call by dialing the recipient's number on their mobile device (MS).

2. **Signal to the BTS:** The call request is first sent to the nearest Base Transceiver Station (BTS) from the caller's mobile device.

3. **Connection to the BSC:** The BTS forwards the call request to the Base Station Controller (BSC) responsible for the BTS. The BSC manages the radio resources and decides on the best path for the call.

4. **Authentication by the AuC:** Before proceeding, the network authenticates the caller's identity through the Authentication Center (AuC), which verifies the validity of the SIM card.

5. **MSC Handling:** Once authenticated, the call is routed to the Mobile Switching Center (MSC), the core element of the Network Switching Subsystem (NSS). The MSC is responsible for setting up the call, managing the connection, and routing the call to the recipient.

6. **Query the HLR:** The MSC queries the Home Location Register (HLR) to retrieve the service profile of the recipient and to find out the current location of the recipient's mobile device.

7. **Interaction with the VLR:** The HLR responds with information indicating that the recipient is registered in a different area, managed by another MSC. The recipient's current MSC consults the Visitor Location Register (VLR) associated with it to get the precise location of the recipient and to ensure the recipient's device is allowed to receive calls.

8. **Routing the Call:** With the recipient's location confirmed, the MSC routes the call through the network infrastructure (potentially involving the GMSC if the call is to cross into another network) towards the recipient's current MSC and then to the BSC and BTS serving the recipient.

9. **Call Establishment:** The BTS nearest to the recipient sends a signal to the recipient's mobile device, causing it to ring. If the recipient answers, a voice channel

is established between the caller and recipient through the network's infrastructure, allowing them to communicate.

10. **Call Termination:** Once the call is finished, either party can end the call. The network then disassembles the dedicated voice channel and releases the resources used for the call.

1.2.4 Calls in VoIP networks

Voice over Internet Protocol (VoIP) facilitates the transmission of voice communications over both wired (such ADSL, or optical fiber) and wireless (including satellite, Wi-Fi, or LTE) internet networks. The foundation of a VoIP network lies in its servers, which authenticate, manage, and direct client communications, alongside optional gateways for linking with traditional telephone systems like the Public Switched Telephone Network (PSTN) and cellular networks. [29]

Clients utilizing VoIP technology range from IP-enabled hard-phones and softphones to conventional analog phones equipped with Analog Telephone Adapters (ATAs). Within corporate environments, these devices are overseen by an IP Private Branch Exchange (IP PBX), responsible for assigning IP phone numbers to individual extensions and managing internal call connectivity. [29]

Additionally, VoIP functionality extends to mobile data services via Over-The-Top (OTT) applications such as Skype, Discord, and WhatsApp. These applications leverage VoIP protocols to offer cost-effective calling options, challenging traditional mobile service providers due to their affordability. The economic advantage of VoIP calls over cellular calls stems from their use of pre-existing internet services and infrastructures. Prior to transmission, voice data undergoes compression and is packaged into IP packets through codecs, which balance sound quality against bandwidth consumption. However, the quality of VoIP calls can suffer from shared bandwidth, network latency, and issues like packet loss, and delay, leading to disruptions in the audio stream. [29]

1.3 Call Routing in Cellular Networks

1.3.1 Engaged Parties

The involved parties and call routing schemes can differ significantly based on whether the call is on- or off-network, domestic (within the nation), or international. In contrast to off-network (off-net) calls, on-net calls occur when both the caller and the called parties are subscribers to the same mobile operator.[29]

End-Users: Make and accept calls, often utilizing devices equipped with multiple SIM card slots. This allows them to insert two or three SIM cards from various service providers, a practice prevalent in emerging markets. It enables end-users to continuously benefit from the most competitive offers among different network operators.[29]

Mobile Operators: Facilitate call routing by relaying traffic from the radio access network through the base station to the core network. Within the core network, the Mobile Switching Center (MSC) determines a path to the intended recipient. For calls that are off-net, the call traffic is forwarded to the Gateway MSC, which then connects with the mobile operator of the called party. This interconnection may occur directly or via intermediary carriers.[29]

Intermediate Carriers: Whether they are publicly owned entities like Tata Communications in India or private companies such as Belgacom ICS and Telia Carrier, provide pathways to termination or transit countries. They obtain these routes through partnerships and then offer them to others for resale. Their primary role is in facilitating international call routing, particularly when there is no direct connection between the originating and destination operators. As a result, international call traffic typically moves from one carrier to another, forming a series of hops between the originating and destination mobile operators.[29]

The agreements between carriers and operators outline the terms and conditions for their interconnection, including traffic measurement, Points of Interconnection, and quality of service standards. Carriers have a variety of transport link technologies at their disposal to handle traffic, such as satellite links, submarine cables, fiber rings, and others, which influence both the cost and quality of the route. Moreover, carriers are increasingly utilizing VoIP technology to send voice traffic as data packets. A hop in the international termination route is deemed legal if the carrier is licensed in its country to employ a regulated transport link technology. However, VoIP links are difficult to regulate and monitor, posing certain challenges and risks; they may be considered illegal depending on local regulations. There are three categories of international termination routes: white, grey, and black. A route is classified as white when it is completely legal with no unauthorized hops. Grey routes involve at least one illegal hop, where the traffic is sent by the originating operator through a licensed carrier but ends at the destination via an unlicensed carrier, a common scenario for calls from the USA to India. Black routes involve unauthorized interconnections at both the source and destination. [29]

The telecommunications sector is characterized by its dynamism. Mobile operators often maintain interconnections with numerous carriers, potentially reaching into the hundreds, for each destination country, necessitating a choice among them for the termination path. Moreover, the quality and cost of these routes can fluctuate on a weekly basis even with the same carrier. To adapt to these changes, Gateway MSCs employ Least Cost Routing algorithms that automatically identify the route that offers the best balance of quality and cost. This approach ensures optimal utilization of the network infrastructure and enhances the revenue generation for operators.

Regulators: In certain countries, mobile operators' operations and alliances are regulated by entities, which can be either public organizations such as ministries or private bodies. Generally, governments view telecommunications as a crucial public service and aim to guarantee that these services are delivered in alignment with the country's understanding of the public good. [29]

1.3.2 International Call

Figure 1.3 illustrates the process of an international call from Phone X to Phone Y, who is a subscriber of a foreign operator, Operator B. The call initiated by Operator A is routed through two intermediary carriers. Carrier 1 is selected via a least cost routing algorithm, while Carrier 2, which has a direct link, forwards the call to the core network of Operator B. Operator B then establishes the connection to Phone Y. This scenario exemplifies a white call routing, characterized by conventional, well-regulated carrier-to-carrier connections that are defined by formal agreements. Although the example mentions only two intermediary steps between Operator A and Operator B, the actual number of hops in practice is typically undisclosed, as the routing process lacks transparency. Generally, a carrier is only aware of the hop immediately preceding and following it in the call's path, as well as the phone numbers of both the calling and receiving parties. Furthermore, the calling number might sometimes be omitted or incorrect.

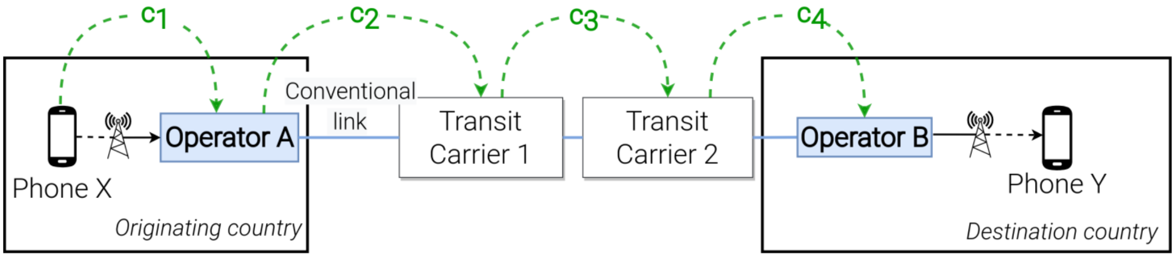


FIGURE 1.3: International Call Route [29]

1.3.3 Money Flow

In every type of call routing scenario, including domestic (both on-net and off-net) and international, it is the caller who is responsible for covering the call termination fees. However, the cost associated with international calls is typically higher than that of domestic calls. This higher cost can be attributed to the call's journey across various intermediate operators before it reaches the intended recipient. Each operator involved in the transit process receives a portion of the call's revenue, known as the settlement rate, while the terminating operator on whose network the call concludes collects the termination fee. Illustrated by a green dashed line in **Figure[1.3]**, an example of the flow of funds shows Operator A charging the end customer a billing rate "c1", which includes both its own retention and the amount "c2" it pays to carrier 1 for call routing. Similarly, transit carriers invoice for their services including the costs to facilitate the call's progression towards its final destination. Finally, the termination fees, denoted as "c4", are levied by the destination operator, Operator B.

1.4 Presentation of the mobile operator Djezzy

1.4.1 Organization of the company

Optimum Telecom Algeria, commonly known as Djezzy, is a prominent telecommunications provider in Algeria that was established in July 2001. As a leading entity in the mobile telephony sector, it boasted 14.59 million subscribers as of 2021. The company offers an extensive array of services, including prepaid and post-paid plans, data services, as well as various value-added services and SUT. In January 2022, the National Investment Fund (FNI) acquired full ownership of Djezzy's capital. Presently, Djezzy's 4G network extends coverage to over 37% of the Algerian population.

Djezzy is currently undergoing a transformation to position itself as the leading digital operator in Algeria, aiming to facilitate customers' access to the digital world. The company is under the leadership of MAHIEDDINE ALLOUCHE, who serves as the Acting Managing Director.

Djezzy, a global communication and technology firm, operates with a vision founded on entrepreneurial principles, prioritizing customer satisfaction, innovation, partnership, and integrity. The company was granted its 2G license on July 30, 2001, and later received its 3G license on December 2, 2013, with an extension of the license occurring on September 4, 2016.

FIGURE 1.4: Logo of Djezzy

Djezzy's corporate framework comprises various crucial divisions and roles that collaborate to maintain the company's operations seamlessly.

Executive Management: The executive management team is tasked with guiding the company's strategic direction, making major decisions, and overseeing all aspects of operations (see Figure 1.5).

Sales Department: This division is dedicated to sales, marketing, and the creation of new products. Its responsibilities include crafting marketing strategies, promoting the company's services, overseeing customer relations, and introducing new products and deals.

Technical Department: The technical department oversees the administration and functioning of Djezzy's telecommunications network. Its duties encompass maintaining and enhancing the network, broadening service coverage, and ensuring the quality of the services provided.

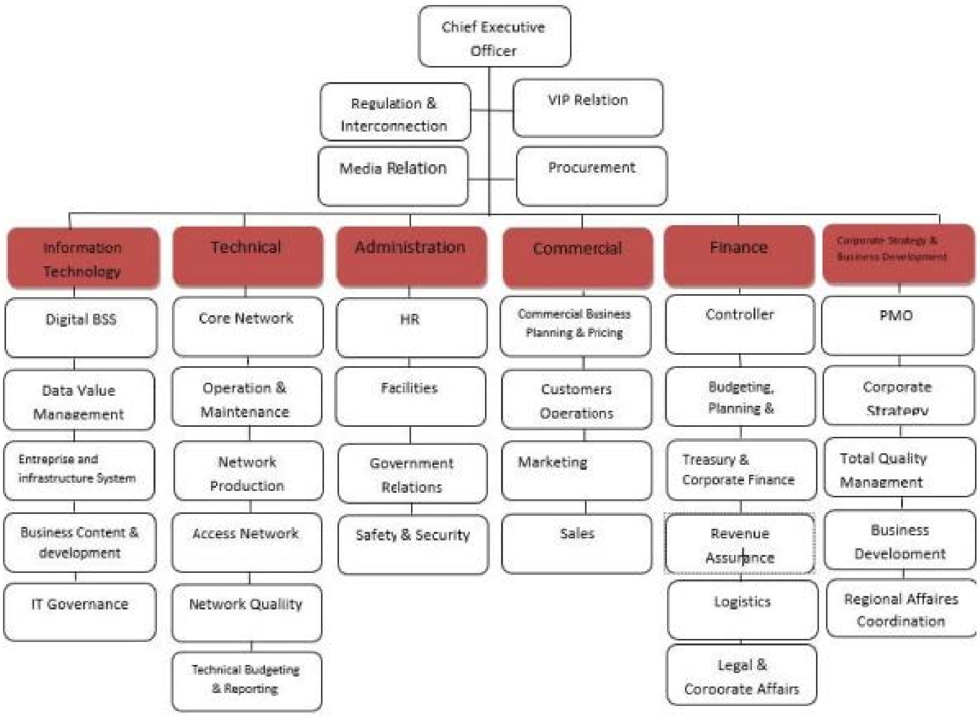


FIGURE 1.5: Organization of Djezzy

Human Resources Department: The Human Resources department is engaged in hiring, talent management, training, compensation, and implementing HR policies within the company. Its goal is to cultivate a nurturing work atmosphere and promote the professional growth of its employees.

Financial Department: The finance department handles the management of the company's financial assets, encompassing tasks such as accounting, budgeting, risk management, and the preparation of financial statements.

Legal and Compliance Department: The legal and compliance department addresses matters related to legal affairs and adherence to regulations, including managing contracts, ensuring regulatory compliance, protecting intellectual property rights, and overseeing litigation.

Data science Department: Our internship was conducted in the Department of Data Science, a unit under the EIM (Enterprise and Infrastructure Management) department. This department plays a critical role in the company's operations by undertaking key initiatives such as:

- Ensuring the accuracy of company data through regular analyses and verifications to maintain the integrity of data crucial for the company's operational and strategic decisions.
- Structuring and cataloging the company's diverse business areas by developing data catalogs and glossaries, which facilitates a better understanding and organization of data tailored to the specific needs of each business sector.
- Complying with national and international data protection regulations, like the DPA (Data Protection Authority) and GDPR (General Data Protection Regulation), to guarantee that data management and processing align with legal standards.
- Enhancing communication across different departments by implementing data science tools for collaboration, thus streamlining the exchange of information and coordination of data-related tasks.

The impact of the Department of Data Science is observed in several key areas, including:

- Production of accurate and reliable reports based on high-quality data, significantly improving the company's decision-making capabilities.
- Facilitation of smoother communication between various departments, thereby enhancing collaboration and the flow of information.
- Compliance with data management guidelines and recommendations from both national and international regulatory authorities, ensuring the company fulfills its legal and regulatory responsibilities while minimizing data protection risks.

1.4.2 Missions of Djezzy

Djezzy is determined to achieve its vision through the following strategic commitments:

- Providing top-quality products at competitive prices to guarantee customer satisfaction.
- Investing in cutting-edge infrastructure to offer reliable and efficient connectivity throughout Algeria.
- Creating a conducive work environment that encourages the professional and personal growth of its employees.
- Actively enhancing the welfare of the Algerian community by offering services tailored to their communication needs and by participating in social initiatives.
- Ensuring the appreciation of shareholders through the optimization of operations and strict cost management.
- Implementing an environmental policy with the goal of reducing the ecological footprint of its operations.
- Constantly refining internal processes to adhere to the highest quality standards as outlined in its quality policy.

By implementing these strategies, Djezzy intends to establish itself as the unparalleled leader in Algeria's telecommunications industry, providing premium products and services, generating value for its stakeholders, and making a constructive contribution to the nation's advancement and prosperity.

1.4.3 Digital Transformation

Digital transformation, also known as digitization, denotes the comprehensive integration of digital technologies across an organization's entire operations. Djezzy initiated a digital transformation program in 2015, stemming from an agreement between

FNIVimpelCom, aiming to position itself as a leading digital pioneer in Algeria. This program encompasses several objectives:

- Optimizing the use of resources, both human and material, through the consolidation of information systems and computer networks.
- Cutting expenses by delegating certain technical operations to external providers.
- Establishing a presence in the digital realm by introducing innovative services.
- Leveraging the potential of emerging technologies, such as big data, to capture new opportunities.
- Ensuring that employees work in conditions conducive to heightened productivity.

1.4.4 The Challenges of Djezzy

Djezzy, like any other telecommunications company, faces different problems and challenges. Here are some common problems Djezzy faces:

Intense Competition: The telecommunications industry in Algeria is characterized by intense competition due to the existence of multiple operators. Djezzy encounters strong rivalry in attracting and keeping customers, potentially impacting pricing strategies and profit margins.

Network Coverage and Quality of Service: Maintaining widespread network coverage and superior service quality throughout the nation poses significant technical and logistical hurdles. To offer dependable connectivity to its clientele, Djezzy is required to persistently fund the growth and modernization of its infrastructure.

Technological Evolution: Swift progress in telecommunications technologies necessitates considerable investment to remain up-to-date. Djezzy needs to keep pace with emerging innovations like 5G and devise strategies for integrating these advanced technologies, all while controlling related expenses.

Data Management: As a telecommunications provider, Djezzy handles a large volume of confidential customer data, making it crucial for the company to implement robust security protocols to safeguard this information and adhere to existing data protection laws. Furthermore, Djezzy needs to verify the consistency of data collected during customer registration and its match with the information on scanned identity documents. This practice is essential for maintaining the accuracy and integrity of the data collected, and for preventing any fraudulent or unauthorized use of customer information.

Changing Needs of Customers: In the fast-evolving telecommunications sector, customer needs and expectations are constantly shifting. It is essential for Djezzy to accurately grasp these changing demands and adjust its products and services accordingly to stay ahead in the competitive landscape.

Regulation and Legal Framework: Telecommunications firms face rigorous regulatory and legal mandates concerning licensing, consumer rights, network safety, and more. Djezzy is required to adhere to these regulatory demands and adjust to any modifications in the legal environment.

Reputation: Factors like service quality, customer satisfaction, and network coverage significantly impact consumers' view of Djezzy. Upholding a strong brand image and positive reputation is crucial for attracting new clients and keeping current ones.These matters highlight potential challenges Djezzy might encounter as a telecommunications provider. Addressing these challenges effectively demands strategic foresight, continuous focus on consumer requirements, suitable investments in infrastructure, and flexible responses to technological and regulatory shifts.

1.5 Fraud problem

1.5.1 Definition

"Fraud" refers to any activity where the goal is to obtain an unfair advantage through deception. According to Black's Law Dictionary, it becomes a criminal violation when there is deliberate misrepresentation or concealment of material facts that would otherwise force a party to behave in a way that would be detrimental to them. Fraud is basically the use of dishonesty to take other people's property or financial assets without their consent. [18]

The notion of the Fraud Triangle encapsulates the motivations for fraud committed by individuals. According to this framework, the ability to justify wrongdoing as necessary or acceptable, a perceived opportunity to commit the act with little risk of detection, and a pressing financial need that the individual feels cannot be met through legal channels are the main drivers of fraudulent behavior. Therefore, the Fraud Triangle offers a brief overview of the situational and psychological elements that contribute to fraud **Figure[1.6]**.[18]

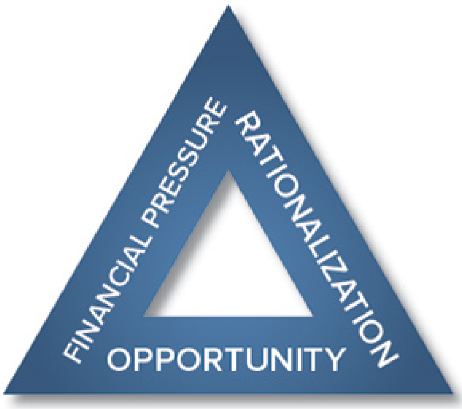


FIGURE 1.6: Fraud triangle [18]

Financial Pressure: This happens when someone has a pressing need for funds or resources that they don't think they can obtain legally. This demand, which is frequently hidden, may result from internal problems like debt, addiction, or the desire for a more luxurious way of life.

Opportunity: This is a reference to the person's belief that they can commit fraud secretly. This perception could be the result of poor oversight, a lack of internal controls within an organization, or the person's trusted authority position giving them access to resources or data.

Rationalization: This is the method by which people convince themselves that their dishonest behavior is acceptable by mentally justifying it. Some rationalizations for their behavior could be that they are just "borrowing" the money, that they should be compensated more for their efforts, or that no one is harmed directly.

1.5.2 Fraud in Telecom

Telecommunications companies are increasingly beleaguered by fraud, leading to significant financial repercussions and the potential loss of customers for those unable to effectively counteract these activities. The development of adaptive and automated systems presents a viable solution to mitigate such fraud. The implementation of cutting-edge technologies, such as artificial intelligence (AI) and machine learning (ML), can enhance the detection and prevention of fraudulent activities by analyzing patterns and predicting potential threats with greater accuracy. Moreover, engaging in industry-wide collaborations and sharing intelligence about fraud trends can further bolster defenses against such activities. The financial toll of telecommunication fraud on companies is substantial each year, yet quantifying the exact losses remains challenging, As stated by [57], fraudulent activities result in telecommunication companies experiencing a loss of approximately 7% of their revenue. This difficulty arises from the reluctance of some companies to disclose information to safeguard their reputation and the reality that not all fraudulent activities are detected. In this article [43] published in Communications Fraud Control Association (CFCA) a summary of top 10 fraud types in **Table[1.1]**.

Top 10 Fraud Types
$5.04 B – International Revenue Share Fraud (IRSF)
$3.28 B – Arbitrage
$2.71 B – Interconnect Bypass (e.g., SIM Box)
$2.27 B – Domestic Premium Rate Service (In Country)
$2.00 B – Traffic Pumping (includes: Domestic Revenue Share, TFTP)
$1.76 B – Commissions Fraud
$1.76 B – Device / Hardware Reselling
$1.49 B – Theft / Stolen Goods
$1.17 B – Friendly Fraud
$.98B – Wholesale SIP Trunking Fraud

TABLE 1.1: Top 10 Types of Telecommunication Frauds [43]

In this work, we will specifically focus on Interconnect Bypass fraud, commonly known as SIM Box fraud, due to its significant impact on telecommunications revenue and its complex challenges in detection and mitigation.

1.6 SIM Boxing

Before delving into the specifics of SIM Boxing, it's crucial to know the components of a SIM Box system.

SIM Box Voice over IP (VoIP) gateway configurations use a device called a SIM box, sometimes referred to as a SIM bank **Figure[1.7]**. It has a significant amount of SIM cards that are kept apart from a VoIP gateway but connected to it. A SIM box can function with numerous GSM gateways in different places by holding SIM cards from different mobile operators in the area. In essence, a SIM box facilitates the installation and management of many SIM cards from different providers, enabling the use of GSM gateways located in separate locations. It is possible to integrate many SIM boxes into a system so that an infinite number of SIM cards can be used.[30]

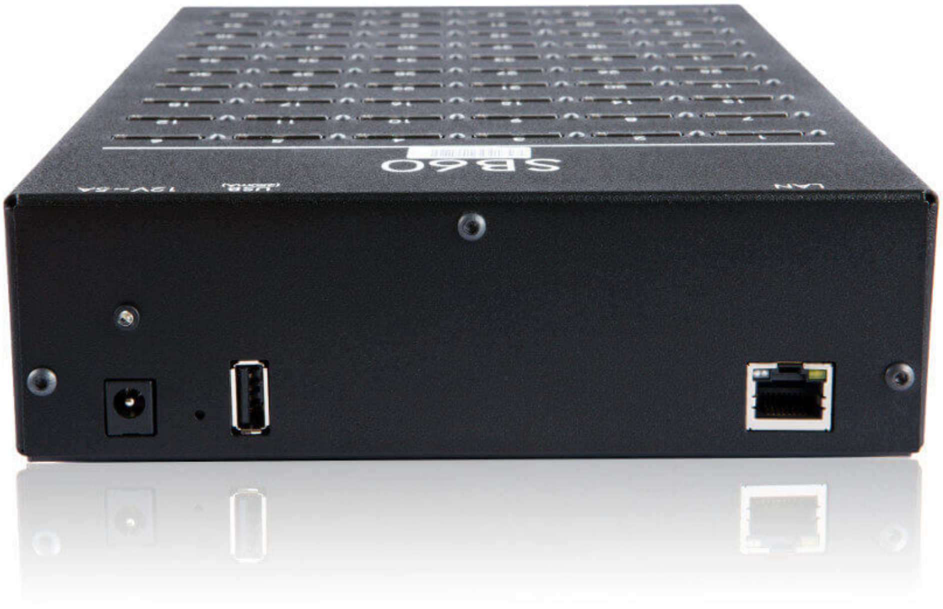


FIGURE 1.7: Sim Box device [7]

GSM/VOIP Gateway The GSM/VOIP Gateway is the component that actually performs the conversion between GSM calls and VoIP calls. GSM/VOIP gateway interfaces with the SIM Box to access the GSM network via the SIM cards. The gateway then converts the GSM signal into VoIP packets for transmission over the internet, or vice versa, for calls coming from the internet to be sent out on the GSM network **Figure[1.8]**.

1.6.1 Sim Box Fraud

SIM Boxes exploited for fraudulent purposes to evade standard telecom network connections through advanced voice technologies like VoIP (Voice over Internet Protocol). They manipulate international voice calls to appear as local calls by routing them

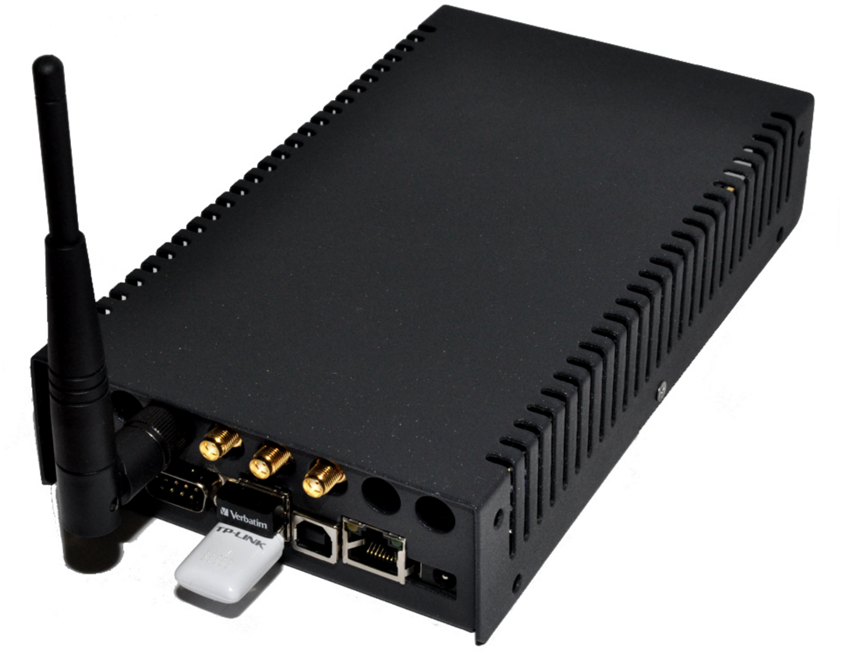


FIGURE 1.8: GSM/VOIP Gateway device [61]

through internet to a local SIM card within the device, bypassing the usual network gateways. This practice leads to financial losses for telecom providers by exploiting the cost discrepancy between international and local call rates, costing the industry billions annually. It not only results in significant revenue loss for network operators due to the avoidance of interconnection fees but also degrades the quality of service for users, manifesting in poor call quality and increased dropped calls. Fraudsters utilize these devices with prepaid national SIM cards to reroute incoming international calls as local, thereby circumventing legal billing rates and causing substantial financial harm to telecom companies.[37]

We will next examine the operational scenarios of both legitimate call routing and illicit call routing facilitated by SIM box fraud. This comparative analysis aims to shed light on the stark differences in the routing processes, illustrating how legitimate calls traverse through sanctioned network pathways, incurring standard fees, versus how SIM box fraud manipulates this system to reroute international calls as local ones, evading rightful charges and undermining the integrity of telecom operations.

1.6.2 Legitimate Route

Consider a scenario where Subscriber "a" from Country A wants to call Subscriber "b" in Country B, as illustrated by the legitimate call routing process in **Figure[1.9]**.

1. Subscriber "a" places a call to Subscriber "b" via their mobile network provider, paying the relevant charges for the call to the service provider.
2. The call from Subscriber "a" is sent to Country A's international gateway. This gateway then relays the call to an intermediary operator, taking on the payment obligations for the call transmission.
3. This intermediary operator then forwards the call to Country B's international gateway, paying a fee to the receiving international operator for call handling.
4. Country B's international gateway then completes the connection, routing the call to Subscriber "b" through its local network.

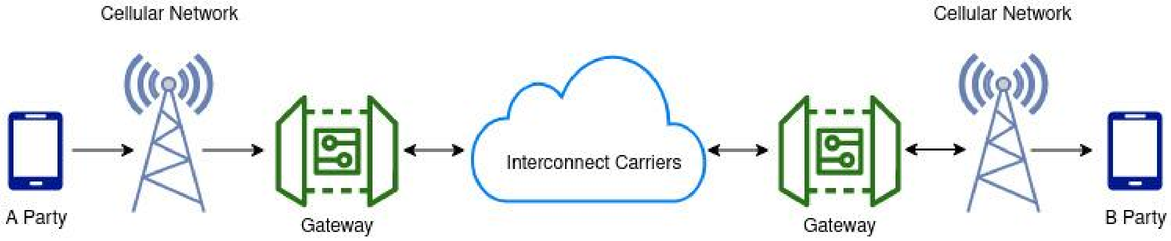


FIGURE 1.9: Legitimate International Route[27]

1.6.3 SIM Box Fraud Route

Consider a scenario involving Subscriber "a" from Country A, Subscriber "b" from Country B, and a fraudster operating a SIM-Box in Country B, as depicted by the process of SIM-Box Fraud in **Figure[1.10]**.

1. Subscriber "a" makes a call to Subscriber "b" through their domestic mobile network and pays for the call.

2. This call from Subscriber "a" gets routed to Country A's international gateway.

3. Country A's international gateway then sends the call to an intermediary operator, which takes on the financial responsibility for the call's transfer.

4. The intermediary operator forwards the call via Voice over Internet Protocol (VoIP) to a SIM-Box in Country B, paying a fee to the operator of the SIM-Box.

5. The SIM-Box then initiates a new call to Subscriber "b" using a local SIM card within Country B's network. This approach makes the call appear domestic, thereby avoiding interconnection charges and incurring only the cost of a local call.

6. As a result, Subscriber "b" receives what seems to be a local call in Country B, though it's actually an international call from Subscriber "a" in Country A.

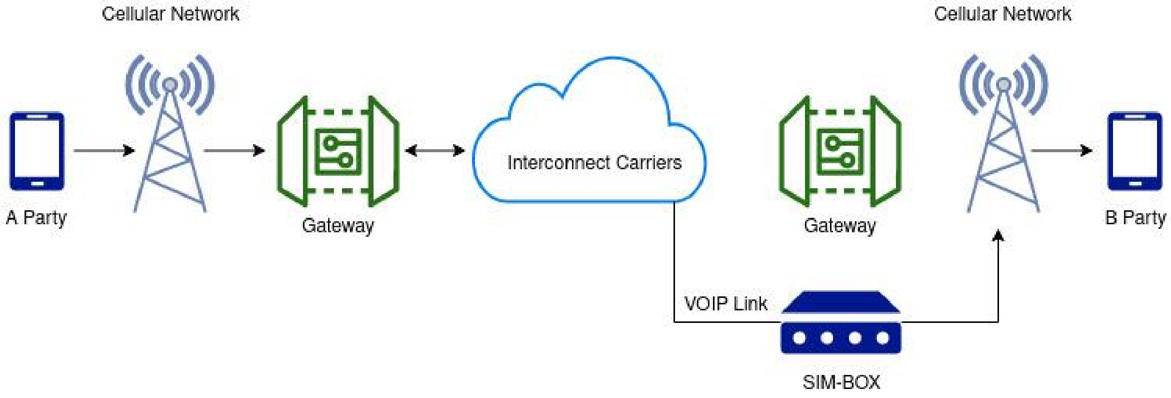


FIGURE 1.10: Sim Box Fraud Route[27]

1.7 Conclusion

In conclusion, this chapter has explored mobile telecommunications, detailing mobile network structure, GSM architecture, and call handling differences between traditional cellular networks and VoIP. It examined call routing, international call handling, financial flows, and included a case study of Djezzy to illustrate practical applications. The chapter also addressed SIM box fraud and its distinctions from legitimate routing, highlighting the importance of understanding these aspects in mobile telecommunications. The next chapter will focus on techniques for detecting SIM box fraud, including an introduction to the topic, challenges in detection, and various detection techniques. It will also cover machine learning algorithms, data balancing techniques, and recent related work, providing a comprehensive overview of current SIM box fraud detection methods.

Chapter 2

Techniques of SIM Box Fraud Detection

2.1 Introduction

A flexible fraud detection system is necessary to quickly and efficiently handle various types of fraud. The key strategy in uncovering fraud is to "follow the money," a concept illustrated by Willie Sutton’s famous reason for robbing banks: targeting the source of significant money [5].

While it is ideal to detect fraudulent calls as soon as they occur, in practice, detecting some types of fraud may require monitoring a number of calls. However, steps to combat fraud can be taken more quickly the earlier it is discovered. The impact of voice termination fraud, commonly known as subscriber identity module (SIM-box) fraud or bypass fraud, on mobile networks is particularly serious. So, detecting fraudulent SIMboxes is a very important task. In this chapter, we address the complexities and methodologies related to SIM box fraud detection. We begin by exploring the challenges in detecting SIM box fraud. We then examine various machine learning and deep learning algorithms and techniques used for detection. Additionally, we discuss data balancing techniques to handle imbalanced datasets, such as oversampling and undersampling. The chapter also reviews related works to provide context and insights from previous studies. Finally, we synthesize the findings, discuss practical challenges, and suggest directions for future research. This structure aims to provide a comprehensive understanding of SIM box fraud detection, equipping researchers and practitioners to effectively tackle this issue.

2.2 Challenges of SIMBox fraud detection

The principal challenges of fraud detection are [29]:

- The rapid spread of the use of SIM boxes with sophisticated techniques,
- The difficulty of fast detecting the bypass,

- The class imbalance is due to the small number of fraudulent activities compared to legitimate activities.

2.2.1 The Ongoing Development of Fraud

In the realm of security challenges, the tactics of fraudsters evolve over time to circumvent detection mechanisms, particularly in the context of SIMBox fraud. The capabilities of SIMBox devices are continuously enhanced to slip past detection methods that primarily scrutinize cellular network data, such as Call Detail Records (CDRs), to separate fraudulent activities from legitimate user actions. Fraudsters adapt by crafting sophisticated techniques that closely resemble genuine human communication patterns [29].

Despite the high accuracy of current detection approaches documented in the literature, SIMBox fraud persists. This is because these detection methods are largely based on analyzing historical data from operators, which represent the assumed ground truth within CDRs. Such approaches tend to merely automate existing detection processes without advancing them, thus failing to identify emerging fraudulent tactics. Moreover, the literature often lacks detailed information on the behavior of fraudsters within the datasets analyzed. This absence significantly affects the interpretation and application of detection methods, leaving it ambiguous whether their success is attributable to the identification of simplistic, predictable fraud likely to evolve or the effectiveness of the detection methodologies in adapting to changes in fraudulent activities [29].

2.2.2 The Requirement For Fast Detection

In 2023, given the minimal investment needed to carry out SIMBox fraud, involving a one-time purchase of SIMBox devices estimated at $550 per gateway and the ongoing acquisition of inexpensive SIM cards, fraudsters can quickly see profits from this illicit activity, potentially earning $100 per day for each SIMBox gateway. The current timeframe for gaining the necessary insights for detection and proceeding with SIM blocking ranges from a day to a week, which grants fraudsters sufficient time to inflict financial losses. This considerable profit margin encourages the continuation of fraud activities, as fraudsters readily replace any blocked SIM cards. Consequently, effective detection of SIMBox fraud demands rapid identification methods that can accurately pinpoint fraudulent activities [29].

2.2.3 Class Imbalance

The class imbalance issue in the telecommunications sector significantly affects the performance of fraud detection systems. This phenomenon arises when the number of fraudulent activities (the minority class) is much smaller compared to legitimate activities (the majority class), leading to a bias in predictive models towards the majority class. Such an imbalance complicates accurate modeling and increases the likelihood of misclassifying fraudulent activities. Moreover, it raises the incidence of false positives, where legitimate activities are erroneously flagged as fraudulent. This not only undermines the accuracy of fraud detection but also negatively impacts customer satisfaction and trust. Addressing class imbalance is crucial for improving the effectiveness of fraud detection systems, requiring advanced analytical techniques and algorithms adept at managing uneven class distributions [29].

2.3 SIM Box Fraud Detection techniques

The likelihood of successfully committing fraud is significantly tied to the ability to secure SIM cards from network operators or their distributors worldwide. Consequently, the impact of SIM box fraud differs from one country to another. In jurisdictions where the law mandates SIM card registration and classifies SIM box devices as contraband, the prevalence of SIM box fraud is noticeably lower than in places where SIM cards are readily available for little or no cost and where regulations do not enforce subscriber registration. This variation in the impact of fraud across the global telecommunications sector could explain the scant public research on the matter. Although completely eradicating fraud is unattainable, it is imperative to implement effective fraud management measures to curtail these illegal activities to a manageable extent.

The strategy behind SIM Box Detection is to develop methods that progressively identify and obstruct traffic channeled through SIM Boxes, by severing connections with identified SIM Box numbers and potentially enforcing severe penalties. Such measures aim to shield network operators from the risk of a surge in fraudulent activities. Furthermore, the SIM Box Detection Service allows for the determination of whether SIMs belonging to a network operator are being misused to clandestinely redirect international voice traffic to the same network's destinations. The installation of minimal hardware at the customer's site enables the automatic blocking of such fraud, along with providing details on the IMSI and International Mobile Equipment Identity (IMEI) of the implicated SIM Box numbers, thus enhancing the process of tracking down the perpetrators. [37]

As shown in **Figure [2.1]**, there are two main operational modes for approaches to detect SimBox fraud: active and passive. Active strategies represent the first line of defense against fraud for mobile operators because they require ongoing human intervention and the allocation of substantial resources. Conversely, passive strategies are designed to detect fraudulent activity in a network on their own and with the least amount of human involvement. Based on the type of data analyzed, passive techniques are further divided into three subcategories: call detail record (CDR) analysis-based methods, audio data analysis methods, and signaling data analysis strategies. [29]

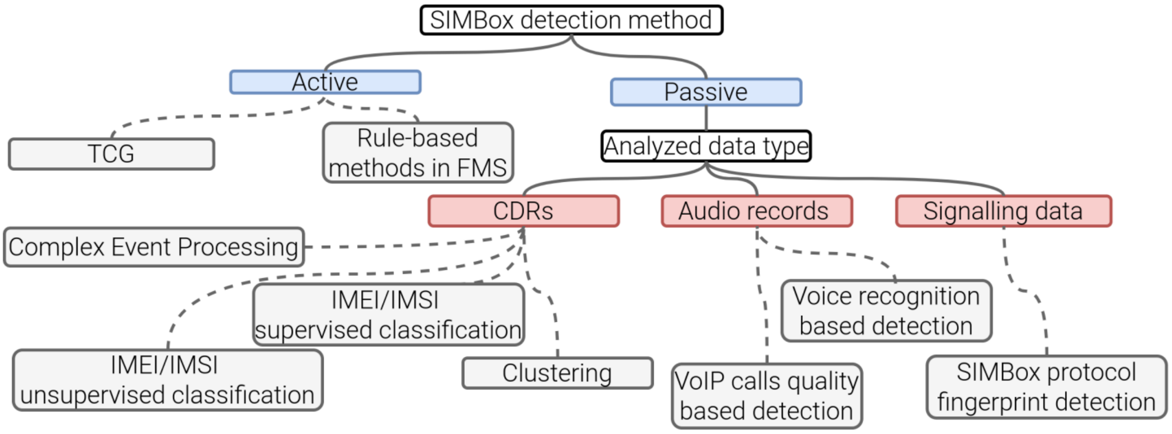

SIMBox detection method
Active
Passive
TCG
Rule-based methods in FMS
Analyzed data type
CDRs
Audio records
Signalling data
Complex Event Processing
IMEI/IMSI supervised classification
IMEI/IMSI unsupervised classification
Clustering
Voice recognition based detection
VoIP calls quality based detection
SIMBox protocol fingerprint detection

FIGURE 2.1: Categorization of existing SIMBox fraud detection methods [29]

Building on the categorization described in **Figure [2.1]**, our work focused on the Call Detail Record (CDR) analysis-based method, specifically employing supervised classification. This approach allows us to leverage detailed call records to systematically identify and classify instances of SimBox fraud within the network, utilizing predefined labels to enhance the accuracy and efficiency of fraud detection.
In the next section, we present a theoretical background of machine learning and deep learning techniques as well as techniques for handling imbalanced data.

2.4 Machine learning algorithms

The concept of machine learning has been defined in various ways across the literature, reflecting its broad application and fundamental importance in the field of computer science. One of the earliest definitions was provided by Arthur Samuel, who described machine learning as "a field of study that gives computers the ability to learn without being explicitly programmed" [51]. This definition emphasizes the autonomous learning capability of machines, distinguishing machine learning from traditional programming paradigms.

Further refining the concept, Tom Mitchell provided a more formalized definition, stating, "A computer program is said to learn from experience "E" with respect to some class of tasks "T" and performance measure "P", if its performance at tasks in "T", as measured by "P", improves with experience "E" [39]. This definition introduces the essential components of the learning process: experience, tasks, and performance metrics, highlighting the iterative improvement of machine performance through learning.

Ethem Alpaydin contributed to the conceptualization of machine learning by defining it as the process of "Programming computers to optimize a performance criterion using example data or past experience" [1]. Like the previous definitions, Alpaydin's emphasizes the goal-oriented nature of machine learning, focusing on optimization based on empirical data.

These definitions collectively underscore the central idea of machine learning as the discipline concerned with equipping computers to autonomously perform tasks by generalizing from data, rather than following explicitly programmed instructions. This perspective has fundamentally shifted the approach to designing computer algorithms, enabling systems to adapt to new scenarios and perform complex tasks by learning from examples and experiences.

2.4.1 Supervised Learning

Supervised learning focuses on establishing a connection between a set of input variables, X, and an output variable, Y, to enable the prediction of outcomes for new, unseen data. This method is the cornerstone of machine learning and holds significant relevance in the handling of multimedia data [14].

2.4.1.1 Decision Trees

A decision tree is a classification model structured as a series of divisions within the space of instances, forming a hierarchical arrangement known as a tree. The structure begins with a primary node, known as the root, which has no incoming connections. Subsequent nodes each receive exactly one incoming edge. Nodes that lead to further divisions are termed internal or test nodes, while terminal nodes, which do not branch further, are called leaves. Test nodes segment the instance space into smaller parts based on specific criteria related to the input attributes. For simple scenarios, each test examines a solitary attribute, dividing the space based on that attribute's value. With numerical attributes, the division is based on whether values fall within a certain range. Leaves represent classifications, often with a probability distribution indicating the likelihood of each possible class. Classification is performed by progressing from the root to a leaf, following the path determined by test outcomes. Figure 3 illustrates a basic decision tree where nodes denote the tested attribute and branches represent attribute values. For numerical attributes, decision trees can be visualized as a series of orthogonal hyperplanes [35]. Simpler decision trees are favored for their ease of interpretation and understandabilit [41].

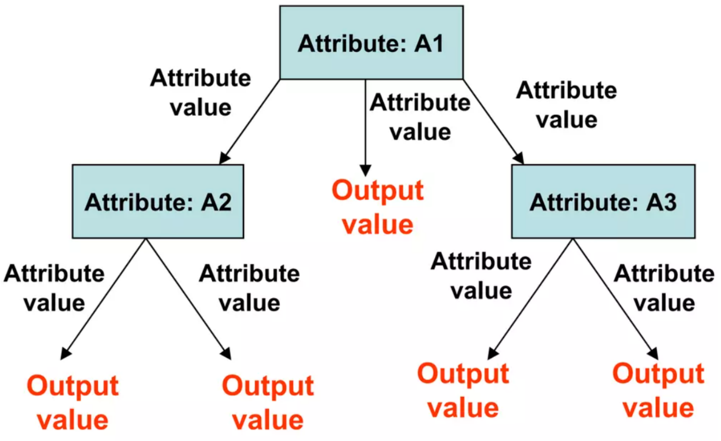


FIGURE 2.2: Decision Tree [25]

2.4.1.2 SVM

The Support Vector Machine is a supervised learning model extensively used in machine learning for both classification and regression challenges, with notable effectiveness in binary classification tasks. Its main goal is to find the best hyperplane or decision boundary that divides data points of different classes. This is achieved by optimizing the margin, which is the space between the hyperplane and the nearest data points of each class, thus ensuring a clear separation between categories. For situations where data cannot be linearly separated, SVMs utilize a mathematical method to project the data into a higher-dimensional space, thereby improving the model's capacity to recognize complex patterns and delineate boundaries. This feature renders SVMs particularly useful for working with complex datasets that cannot be easily categorized through linear methods [11].

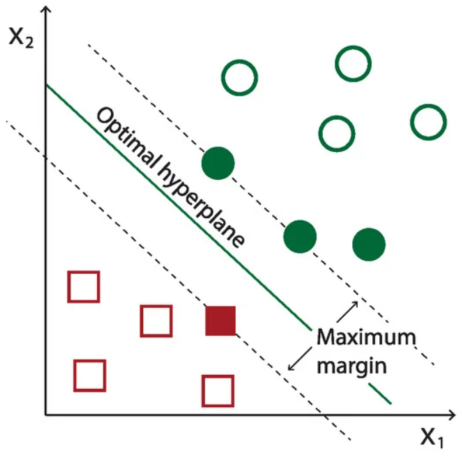


FIGURE 2.3: SVM Geometric Definition [58]

2.4.1.3 Artificial Neural Networks

The Artificial Neural Network is a computational model inspired by the structure of the human brain's neural network. It iteratively processes information, adjusting biases and synaptic weights to learn from the data it receives. ANNs are capable of enhancing their performance through various learning algorithms, including error-correction and competitive learning methods. The learning process within ANNs can be either supervised, where the network is trained using specific input and output patterns, or unsupervised, where the network only requires input patterns to autonomously develop its representations [27]. According to [22], there are specific guidelines for determining the optimal number of hidden layers and neurons in an ANN:

- The number of hidden neurons should ideally be between the size of the input layer and the size of the output layer.
- A recommended calculation for the number of hidden neurons is two-thirds (2/3) the size of the input layer plus the size of the output layer.
- It is advised that the number of hidden neurons does not exceed double the size of the input layer.

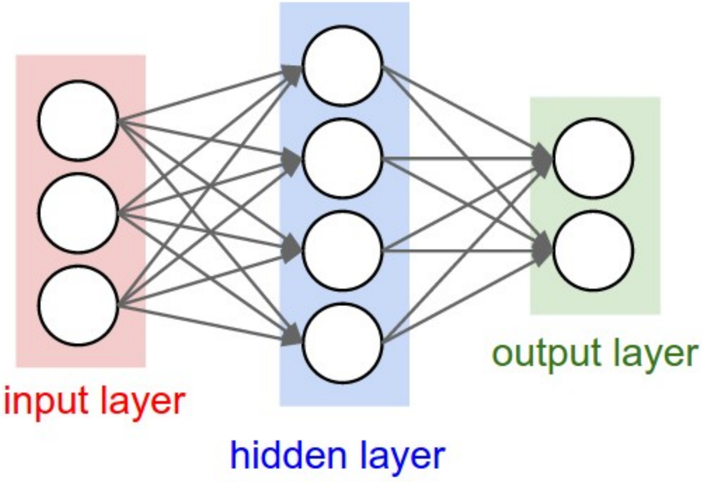


FIGURE 2.4: Simple Neural Network Architecture [32]

2.4.1.4 Random Forests

The Random Forest algorithm, created by Leo Breiman [8], is celebrated for its straightforwardness and versatility in dealing with classification and regression tasks. It functions by aggregating the predictions from numerous decision trees to produce a unified outcome. This method enhances the bagging technique by integrating both bagging and feature randomness, leading to a collection of independent decision trees. Feature randomness, specifically, entails the selection of a random assortment of features, which helps decrease the correlation between individual trees. This aspect sets random forests apart from conventional decision trees, which evaluate all possible feature splits. The strategic choice of feature subsets plays a key role in the high efficiency and effectiveness attributed to random forests [52].

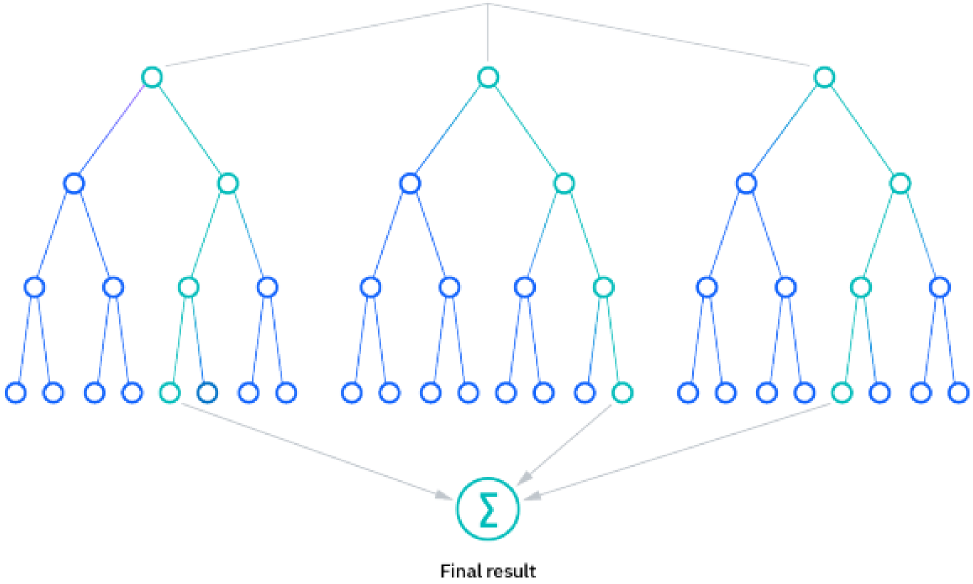


FIGURE 2.5: Random Forest [52]

2.4.1.5 XGBoost

Standing for eXtreme Gradient Boosting, has gained widespread acclaim in the machine learning community for its rapid processing, ability to scale, and proficiency in managing structured or tabular data. This algorithm, devised by Tianqi Chen, functions within a gradient boosting framework that methodically introduces decision trees to amend inaccuracies from preceding trees. Noteworthy attributes comprise its capability to manage missing data, apply regularization methods to curb overfitting, and employ an advanced optimization algorithm for efficient model training [12].

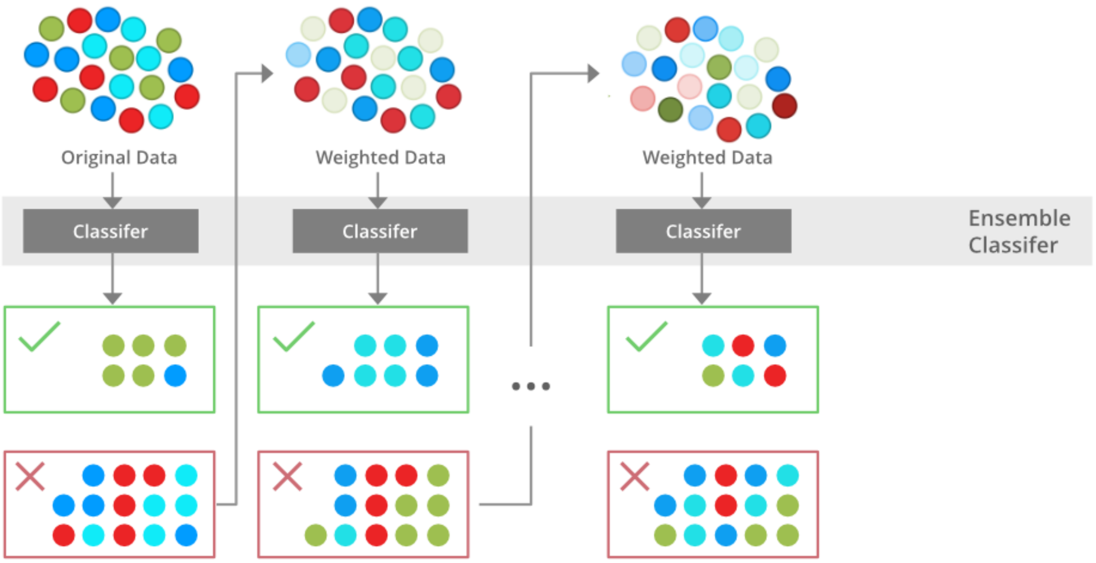


FIGURE 2.6: Boosting [13]

2.4.2 Unsupervised Learning

Unsupervised learning, or unsupervised machine learning, employs algorithms to sift through and categorize datasets that lack labels. These algorithms autonomously unearth underlying patterns or groupings within the data, eliminating the necessity for manual guidance. This technique's prowess in identifying resemblances and variances among data points renders it exceptionally suitable for tasks such as exploratory data analysis, devising cross-selling tactics, segmenting customers, and recognizing images. [59]

2.4.2.1 K-means Clustering

K-means clustering is an uncomplicated algorithm within unsupervised learning aimed at addressing clustering challenges. It operates on a straightforward principle of distributing a provided dataset into a certain quantity of clusters, denoted by "k," which is determined in advance. These clusters are initially positioned as points in space, and each data point in the dataset is assigned to the closest cluster based on distance. The cluster centers are then recalculated, and the assignments are updated, with this cycle repeating until a satisfactory outcome is achieved. The application of K-means clustering spans various fields, including search engines, market segmentation, statistics, and astronomy. [55]

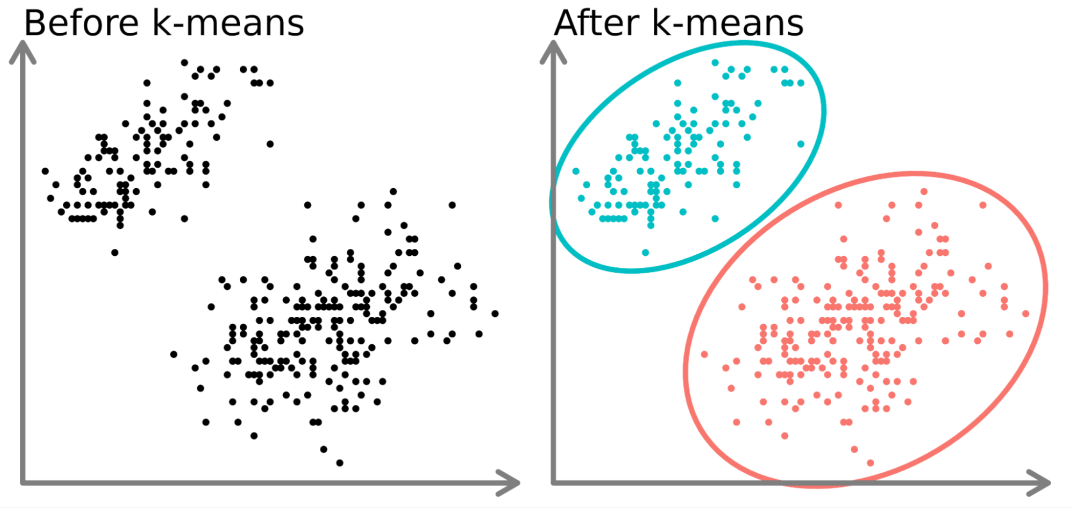


FIGURE 2.7: Kmeans Clustering [38]

2.4.2.2 Principle Component Analysis

Principal component analysis (PCA) is a technique for reducing the dimensionality of extensive datasets. It achieves this by converting a broad set of variables into a more compact set that retains most of the original dataset's information. While reducing the dataset's variables does lead to a slight decrease in accuracy, the essence of dimensionality reduction lies in finding a balance between accuracy and simplicity. Smaller datasets are not only simpler to examine and visualize, but they also facilitate quicker and more efficient analysis by machine learning algorithms, as these algorithms have fewer irrelevant variables to consider. [31]

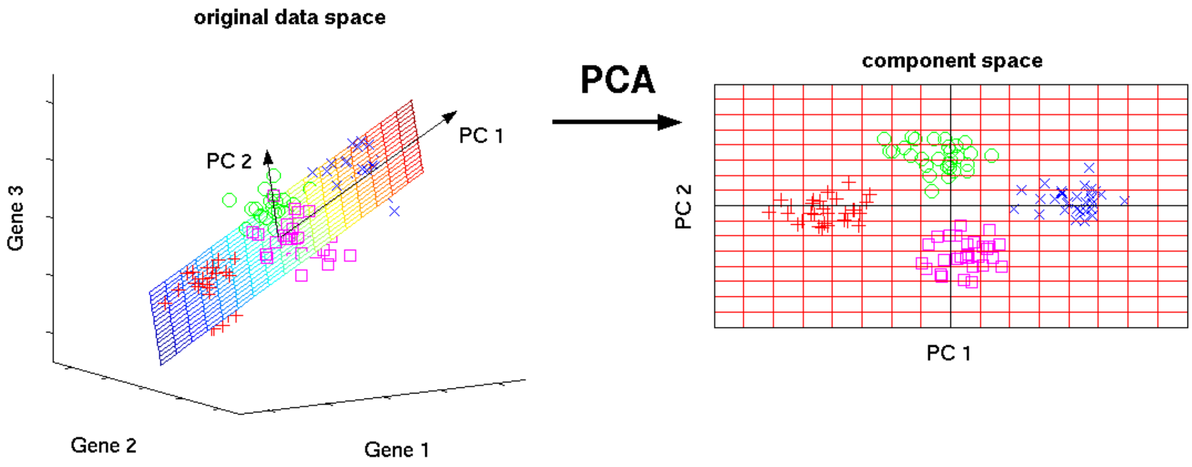


FIGURE 2.8: PCA [20]

2.5 Deep learning

Deep learning is a branch of machine learning, which itself falls under the broader fields of artificial intelligence and statistics. It uses neural networks to address complex problems by learning data representations. These neural networks are designed to mimic the brain's structure, featuring interconnected nodes (neurons) arranged in layers, allowing them to identify patterns and features within the data autonomously. Deep learning algorithms can learn from vast amounts of both labeled and unlabeled data, enabling them to perform tasks like image and speech recognition, natural language processing, and autonomous driving [42].

2.5.1 DNN

Deep Neural Networks are proficient in tasks such as speech recognition, image classification, and object tracking due to their ability to approximate universal functions. They acquire intricate patterns by utilizing numerous layers of interconnected neurons, effectively mimicking highly nonlinear functions. Deep neural networks consist of an input layer, numerous hidden layers, and an output layer, adjusting parameters through backpropagation and gradient descent to reduce forecasting errors. Their profundity enables the automatic extraction of features from unprocessed data, removing the requirement for manual feature construction and rendering them well-suited for tasks involving substantial amounts of unlabeled data. Nevertheless, they encounter obstacles like interpretability, resilience against adversarial attacks, and guarantees of generalization, especially in areas such as medical image analysis [50].

The difference between DNNs and ANNs lies primarily in their depth. DNNs have many more hidden layers than traditional ANNs, enabling them to model more complex, highly nonlinear functions and automatically extract features from raw data. This depth allows DNNs to handle large, unlabeled datasets more effectively than ANNs.

2.5.2 CNN

Convolutional Neural Networks are a type of deep learning model designed for processing data with a grid pattern, such as images. Inspired by the organization of the animal visual cortex, CNNs automatically learn spatial hierarchies of features, progressing from low- to high-level patterns. They consist of three types of layers: convolution, pooling, and fully connected layers. Convolution and pooling layers extract features, while fully connected layers map these features for classification. CNNs efficiently process images by applying a small grid of parameters (kernel) at each image position, allowing features to be detected anywhere in the image. Through training, which optimizes parameters like kernels, CNNs minimize differences between outputs and ground truth labels using backpropagation and gradient descent [63].

2.5.3 RNN

Recurrent Neural Networks (RNNs) represent a broader category within artificial neural networks, as they feature connections that aren't strictly feed-forward. Unlike traditional neural networks, RNNs have connections between units that create directed cycles, which establish an inherent internal memory. This design makes RNNs

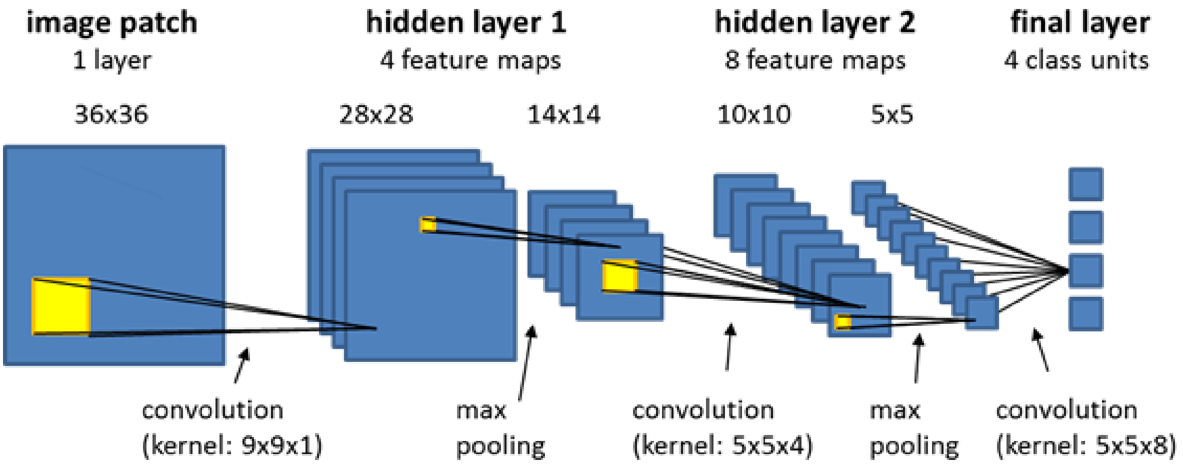


FIGURE 2.9: Convolutional neural networks [34]

particularly suitable for tasks involving time-evolving signals, as their internal memory naturally incorporates temporal information. Notably, RNNs have been proven effective in approximating complex dynamical systems, showcasing their utility in various applications [10].

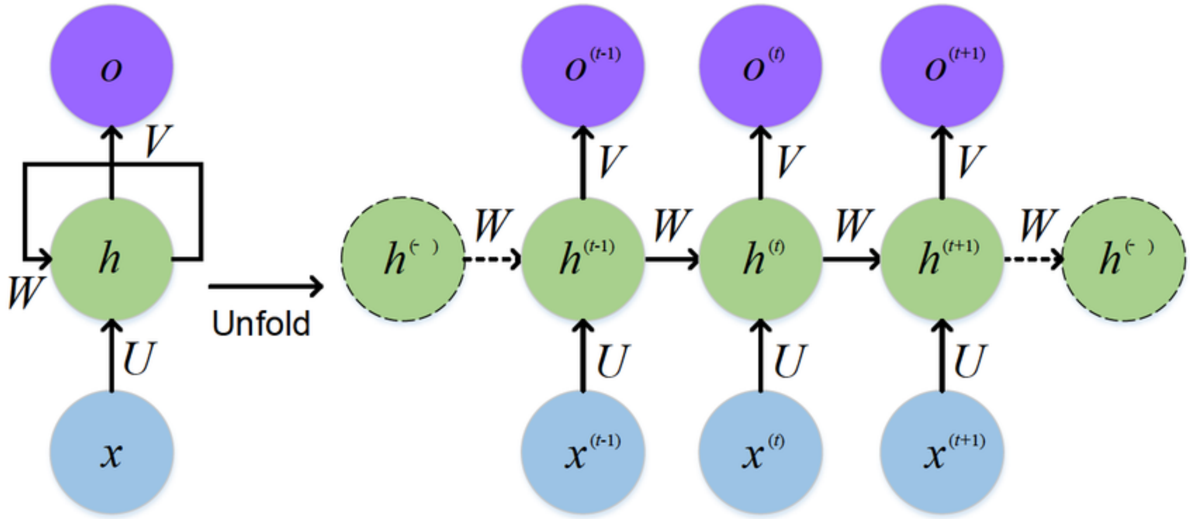


FIGURE 2.10: Recurrent neural networks [16]

2.6 Data balancing techniques

Imbalanced data describes datasets characterized by a disproportionate distribution of observations across target classes, wherein one class label significantly outnumbers the other. This disparity often manifests in scenarios where the majority class eclipses the minority class by a large margin, leading to challenges in predictive modeling and analysis.[60]

We will examine the data-level approach as a strategy to achieve balanced class distribution within the dataset, specifically through the exploration of different techniques of over-sampling, under-sampling and over-sampling followed by under-sampling. These methods are pivotal in adjusting the proportions of the minority and majority classes to address the challenges posed by class imbalance.

2.6.1 Over-Sampling

Oversampling, alternatively known as upsampling, constitutes a sampling strategy aimed at equalizing the dataset by duplicating samples from the minority class. This technique generates a superset that preserves the integrity of the original dataset, albeit it may precipitate overfitting issues and escalate computational complexity, particularly with substantial data volumes. Oversampling is bifurcated into two distinct categories: Random oversampling and Informative oversampling, as delineated by [49]. Random oversampling involves the indiscriminate replication of minority class samples to achieve dataset equilibrium. Conversely, Informative oversampling entails the creation of synthetic minority class data points based on specific criteria to facilitate balance. Various oversampling methodologies have been devised to address class distribution disparities in imbalanced datasets, a selection of which are examined within this section [56].

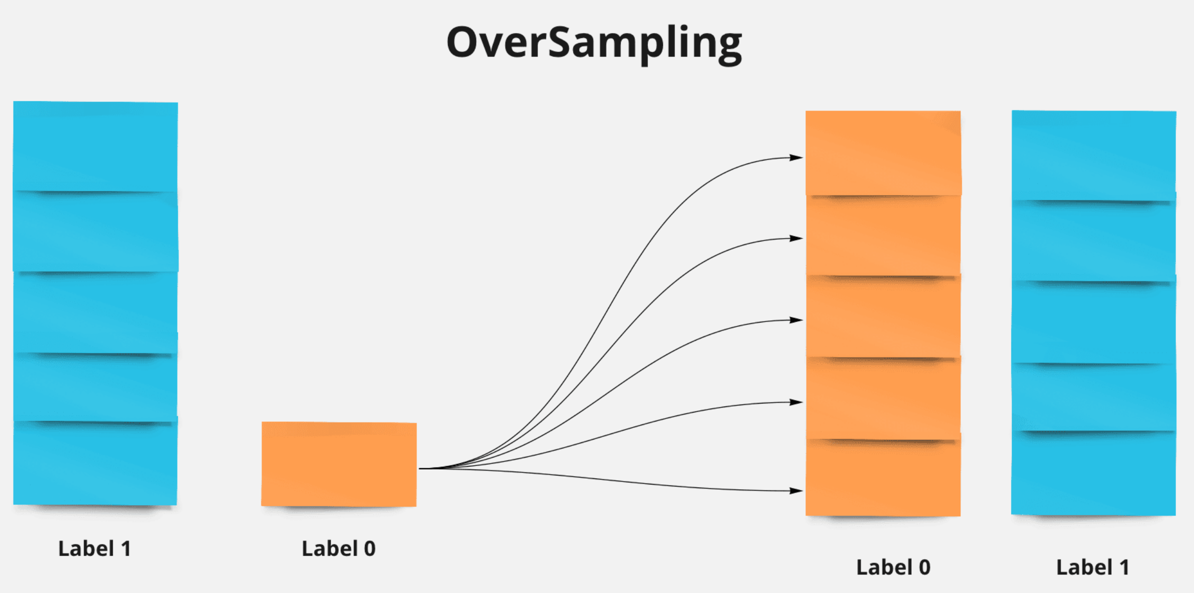


FIGURE 2.11: Over-Sampling [40]

2.6.1.1 SMOTE

Synthetic Minority Oversampling Technique is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together [62].

2.6.1.2 AdaSyn

Adaptive Synthetic Sampling, an algorithm for generating synthetic samples, focuses on augmenting the minority class in a dataset through the creation of synthetic instances. This method operates by evaluating the density distribution of each sample within the minority class, subsequently generating synthetic instances in alignment with this density. The adaptive nature of ADASYN prioritizes the generation of additional synthetic samples for those minority class samples deemed more challenging for machine learning models to classify. This strategy is aimed at enhancing the overall classification accuracy of such models by addressing imbalances in the dataset more effectively [21].

2.6.1.3 ROS

Random Over-Samlping, as introduced by [33], stands as a prevalent and efficacious technique for mitigating the issue of class imbalance. This method entails the random selection and replication of samples from the minority class, which are subsequently incorporated into the training dataset to facilitate the training of machine learning models. By generating identical new data points from the original dataset, random oversampling enhances the representation of the minority class but also elevates the risk of model overfitting due to the repetition of the same data points [56].

2.6.2 Under-Sampling

Undersampling, a method articulated by [17], involves the reduction of majority class instances to achieve a balanced dataset, subsequently utilizing this newly formed subset for the training of learning models. This technique systematically eliminates data from the majority class until an equilibrium is reached within the dataset. Undersampling is categorized into two primary strategies: Random undersampling and Informative undersampling, as identified by [49]. Random undersampling entails the arbitrary selection and removal of instances from the majority class to restore balance. Conversely, Informative undersampling involves the elimination of instances based on specific, pre-established criteria, ensuring that the removal process is guided by informative parameters. Detailed discussions of various undersampling approaches will be provided in the following sections. [56]

2.6.2.1 RUS

Random Under-Sampling is recognized as a prevalent and effective technique for rectifying class imbalances within datasets [17]. This method, known for its speed compared to alternative strategies, involves the selective removal of samples from the

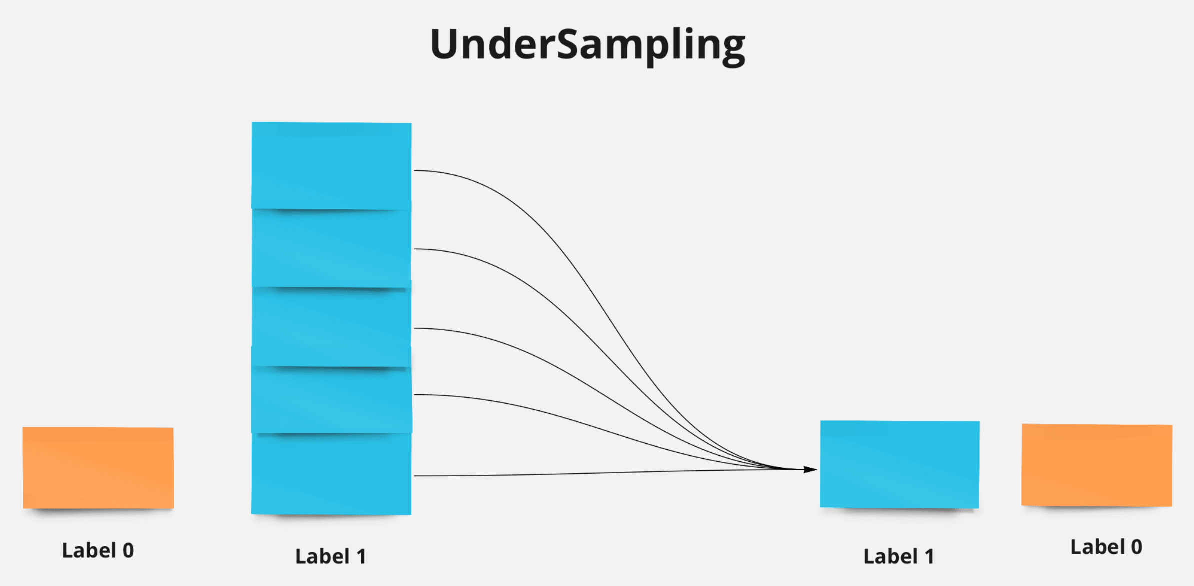


FIGURE 2.12: Under-Sampling [40]

majority class to achieve a balanced class distribution. However, a significant limitation of Random Undersampling (RUS) is the potential loss of critical information, as it may eliminate important instances from the dataset, which could, in turn, impact the model's accuracy [56].

2.6.2.2 Tomeklink

Tomek Links, an undersampling approach introduced by Ivan Tomek in 1976 [24], focuses on removing samples from the majority class that are identified as Tomek Links. Simply put, a Tomek Link enhances the Nearest Neighbor Rule (NNR) by representing the closest pair of instances from opposing classes. This technique is particularly useful for reducing overlap within the dataset after the application of synthetic sampling methods. By eliminating these overlaps, Tomek Links can significantly enhance the performance of classification models. [56]

2.7 Recent related work

In this section, we will explore various literature works focused on SIMBox fraud detection employing supervised classification algorithms. These studies offer insights into how data-driven approaches, particularly machine learning models, are utilized to analyze patterns within Call Detail Records (CDR) to accurately identify fraudulent activities. Additionally, we will delve into how studies have addressed the challenge of imbalanced data, a common issue in fraud detection where the number of fraudulent instances is significantly lower than that of legitimate ones. This exploration will include a review of methodologies and techniques implemented to mitigate the impact of data imbalance on model performance, ensuring more accurate and reliable detection of fraudulent activities.

2.7.1 SIMBox Fraud Detection

For telecom companies, implementing effective measures to detect and counteract SIM-Box fraud is essential. As fraudulent schemes become more complex, it's vital for operators to deploy strategic defenses to protect their networks and financial interests. Recognized strategies for uncovering SIM-Box fraud encompass a range of methodologies, such as Random Forests (RF), Artificial Neural Networks (ANN), Support Vector Machines (SVM), among other sophisticated approaches.

In 2023, Djomadji and colleagues [15] discovered that RandomForest, SVM, and XGBoost algorithms are capable of identifying bypass SIM-Box fraud. The experimental findings revealed that the Random Forest algorithm outperformed the others in terms of accuracy. Specifically, it was found to be the most fitting for the classification model employed in detecting SIM-Box fraud, achieving an accuracy rate of 92%. Subsequently, this model was successfully applied to pinpoint fraudulent numbers within a mobile operator's network.

The research outlined in [27] developed an Artificial Neural Network (ANN) with a structure consisting of a 12-unit input layer, two hidden layers each containing 8 units, and a single-unit output layer. This model was trained using the MSC CDR dataset from Mobitel (Pvt.) Ltd, which included 12 specific features. The researchers trained three separate models on distinct datasets: one oversampled and two undersampled. Through an extensive evaluation process that measured accuracy, sensitivity, and AUC (Area Under the Curve), it was found that the model trained on the oversampled dataset performed remarkably well. This top-performing model achieved an accuracy of 99.59%, a sensitivity rate of 0.99988, and an AUC of 0.998, demonstrating its effectiveness in detecting SIM-Box fraud through classification models.

Kashir and Bashir's 2019 study [28] aimed to pinpoint the most effective classification methods for SIM-Box fraud detection. This research assessed various versions of neural networks and support vector machines (SVM) to thoroughly evaluate their performance. The findings showed that out of five tested artificial neural network (ANN) models, the one utilizing the Bayesian Regularization algorithm stood out by achieving the highest accuracy rate of 99.87%. Among the SVM variants tested, models incorporating Polynomial, Radial Basis, and Sigmoid Kernels emerged as top performers, each registering an impressive accuracy of 99.24%. However, it's important to highlight that artificial neural networks, on the whole, outperformed support vector machines in terms of accuracy.

In ALGHAWI's 2019 study [30], a rigorous experiment was carried out, testing 240 artificial neural network (ANN) models with various architectures, learning rates, and momentum values, alongside 40 support vector machine (SVM) models that utilized different kernels. The research led to the identification of an optimal SVM model that stood out for its efficiency, notably through a significant reduction in runtime when compared to the wide array of tested ANN configurations. Crucially, both the ANN and SVM methodologies were found to be highly effective, each achieving an accuracy rate of 98.5%, highlighting their capability in reliably identifying SIM-Box fraud.

2.7.2 Imbalance Data

In the study [19], researchers evaluated the performance of various classification algorithms for Credit Card Fraud Detection data modeling, with the XGBoost Classifier standing out for its superior performance even before implementing data balancing techniques. Despite achieving an initial accuracy of over 99%, which suggested data imbalance, the evaluation shifted towards the F1 score for a more accurate comparison, revealing the XGBoost Classifier's F1 score of 0.856, precision of 0.913, recall of 0.805, and accuracy of 0.99. To enhance classifier performance further, the study explored three data balancing techniques, identifying Random Over Sampling as the most effective, significantly outperforming Random Under-Sampling and SMOTE. With Random Over Sampling, the XGBoost Classifier achieved remarkable scores across all metrics: accuracy (0.998), precision (0.997), recall (1.0), and F1 score (0.998), marking it as the optimal approach for achieving unbiased and high-quality classification results.

Amit et al. [56] present a comparative analysis of class imbalance handling methods in the context of credit card fraud detection using machine learning. It highlights the challenge of imbalanced datasets where fraudulent transactions are much fewer than legitimate ones. By conducting extensive experiments with state-of-the-art classification models like AdaBoost, XGBoost, and Random Forest, the study evaluates various techniques on performance metrics such as Precision, Recall, K-fold Cross-validation, AUC-ROC curve, and execution time. The key finding is that employing a combination of Oversampling and Undersampling methods notably improves the detection capabilities of ensemble classification models, making a significant contribution to the development of more effective fraud detection systems.

Bart et al. [3] proposed robROSE, an innovative approach to address the challenge of fraud detection within highly imbalanced datasets, where fraud instances are rare. Traditional oversampling methods, while attempting to balance the dataset, often falter due to the presence of outliers, leading to distorted models. RobROSE integrates robust statistical techniques to effectively identify and exclude outliers, coupled with oversampling to better represent the minority fraud class. This methodology enhances fraud detection accuracy by ensuring that synthetic samples accurately mimic genuine fraud cases without being influenced by anomalies. Validated on both simulated and real datasets, robROSE demonstrates improved fraud detection capabilities and offers its source code to the public, promoting further research and application in the field.

2.7.3 Discussion

Simbox fraud detection frequently fails to address the significant problem of data imbalance, which can greatly affect the efficiency of detection systems. Even though recent studies have made progress in developing algorithms, they often fail to tackle this imbalance, which could result in inefficiencies when identifying fraud. Our review [6] points out this discrepancy. Below is a resume table **Table[2.1]** summarizing key approaches and findings from recent works on Sim Box Fraud detection. This table provides an overview of the methodologies, tools, and outcomes associated with each study.

Following the presentation of a table summarizing recent research on SIM box fraud detection, it is clear that only the study by [15] specifically tackles the problem of data imbalance. They utilize strong algorithms like Random Forest and XGBoost, known for their effectiveness in dealing with imbalanced data distributions. This

TABLE 2.1: Summary of Recent Works on Simbox Fraud Detection [6]

Works	**Techniques (Accuracy)**				**Features**
	SVM	**ANN**	**XGB**	**RF**	
[27]	-	**99.59%**	-	-	Mobile Number, Number of Unique MO SMS, Number of Unique Location Area Codes for MT SMS, Number of Unique MT SMS, Number of Unique Location Area Codes for MT SMS, Number of unique MO calls, Number of unique LAC's of MO calls, Number of unique IMEI's of MO calls; Total MO call minutes, Number of unique MT calls, Number of unique LAC's of MT calls, Number of unique IMEI's of MT calls, Total MT call minutes.
[28]	99.24%	**99.87%**	-	-	-
[30]	**98.5%**	98.5%	-	-	Call sub, Total calls, Total numbers called, Total minutes, Total night calls, Total numbers called at night, Total minutes at night, Total incoming, Called numbers to total calls ratio, Average minutes.
[15]	69%	-	81%	**92%**	-

method improves both fraud detection accuracy and reduces bias towards the majority class in the model. These approaches are essential in creating trustworthy and efficient fraud detection systems, particularly in settings where fraudulent behaviors are infrequent compared to valid behaviors.

Our research aims to fill this gap by developing methodologies that further tackle data imbalance. By integrating advanced machine learning algorithms and novel data preprocessing techniques, we seek to create a detection system that not only improves upon existing accuracy metrics but also ensures fairness and robustness across varying data distributions. This approach will significantly contribute to the reliability of fraud detection systems, particularly in scenarios where fraudulent activities are vastly outnumbered by legitimate activities.

2.8 Conclusion

In conclusion, this chapter delved into the techniques of SIM box fraud detection, starting with an introduction to the topic and the challenges faced in detection, such as the ongoing evolution of fraud, the need for rapid detection, and class imbalance. It then reviewed various detection techniques, including machine learning algorithms like supervised (e.g., decision trees, SVM, neural networks, random forests, XGBoost) and unsupervised learning (e.g., K-means clustering, PCA), as well as deep learning methods (DNN, CNN, RNN). Data balancing techniques, both over-sampling (e.g., SMOTE, AdaSyn, ROS) and under-sampling (e.g., RUS, Tomeklink), were also explored. Finally, recent related work in SIM box fraud detection and handling imbalanced data was discussed, providing a comprehensive overview of the current state of detection methods.

In the next chapter, we will discuss the implementation and experimental results of our SIM box fraud detection techniques. This will include the tools and libraries used, details about the dataset, preprocessing steps, and the training and evaluation processes, offering a practical insight into the application of these techniques.

Chapter 3

Implementation And Experimental Results

3.1 Introduction

In this chapter, we investigate how artificial intelligence methods can be used to identify SIM box fraud in a dataset given by Djezzy an Algerian telecommunications company. Analyzing SIM box fraud is essential due to the major financial risks and regulatory challenges it presents for telecommunications companies on a global scale. The process outlined here consists of several crucial stages: preprocessing data, feature engineering, and training predictive models. Every stage is designed to tackle the specific obstacles posed by the intricacies of SIM box fraud. We begin by outlining the dataset, which comprises transactional telecommunication data that has been anonymized in order to adhere to privacy regulations. The following parts outline the extensive preprocessing steps implemented to clean and organize the data, the intentional development of features to showcase fraudulent actions, and the precise choice and fine-tuning of machine learning models for accurate fraud prediction. This section acts as a building block for the thesis, connecting theoretical methods from previous sections to practical results.

3.2 Experimental setup

To achieve optimal performance, the entire study is conducted using Python programming language version 3.11.7 installed on a Windows 10 system with 20 GB of RAM and equipped with NVIDIA MX110 GPU. We present in this section a brief description of the different software tools used in this study.

3.2.1 Python

Python is a programming language with dynamic semantics that is interpreted, object-oriented, and high-level [47]. Due to its advanced built-in data structures, along with dynamic typing and dynamic binding, it is highly appealing for Rapid Application Development, as well as for serving as a scripting language or glue to link existing components. Python's syntax is straightforward and easy to grasp, making it conducive to readability and subsequently lowering the expenses associated with program upkeep. Python has the ability to use modules and packages, which promotes the organization of programs and the reuse of code. The Python interpreter and the wide-ranging standard library can be obtained for free in source or binary form for all main platforms and can be shared without any cost .

3.2.2 Jupyter Notebook

Jupyter Notebook, previously named IPython Notebook, is a web-based tool that allows users to create and distribute computational documents interactively. Initially called IPython, the project underwent a name change to Jupyter in 2014. It is a completely open-source product, and all features can be used by users for no cost. It provides assistance for over 40 languages such as Python, R, and Scala [26].

3.2.3 Anaconda

Anaconda is a distribution of the Python and R programming languages designed for data science, with a focus on simplifying package management and deployment, that is open-source. Anaconda manages package versions through its package management system, conda, which checks the current environment beforehand to prevent interference with other frameworks and packages during installation. The Anaconda distribution includes more than 250 packages that are pre-installed. There are more than 7500 extra open-source packages available for installation from PyPI in addition to the conda package and virtual environment manager. Additionally, there is a GUI called Anaconda Navigator which serves as a graphical substitute for the command line interface. Anaconda Navigator comes bundled with Anaconda distribution, enabling users to open applications and handle conda packages, environments, and channels without the need for command-line instructions. The navigator is able to find packages, install them in a specific setting, execute the packages, and keep them up to date [2].

3.2.4 Pandas

Pandas, a Python package that is open source, is predominantly utilized for tasks related to data science, data analysis, and machine learning. It is constructed on top of a different package called Numpy, which offers assistance for arrays with multiple dimensions. Being one of the most widely used data manipulation libraries, Pandas integrates effectively with various other data analysis tools within the Python environment, and is commonly integrated in all Python distributions, ranging from default operating system installations to commercial distributions such as ActiveState's ActivePython [46].

3.2.5 Matplotlib

Matplotlib is a library for data visualization and graphical plotting in Python, compatible with multiple platforms, including histograms, scatter plots, and bar charts, among others, along with its numerical extension NumPy. Therefore, it provides a practical open source option to MATLAB. Programmers are able to utilize matplotlib's Application Programming Interfaces (APIs) to insert plots into GUI applications as well [44].

3.2.6 Scikit-Learn

Scikit-learn (sklearn) is a popular open-source Python library for machine learning and data modeling. It offers a wide range of algorithms for classification, regression, clustering, and more. Released in 2010, it has become a cornerstone of the Python machine learning ecosystem. Its simple APIs, compatibility with other Python libraries like NumPy and Pandas, and support for various tasks like classification, regression, clustering, and dimensionality reduction make it highly versatile. Scikit-learn is known for its standardized model interface, ease of use, and integration with tools like matplotlib and NumPy. Written primarily in Python with some performance-critical parts in Cython, it leverages NumPy for efficient array operations [53].

3.2.7 Pytorch

PyTorch is an open-source deep learning framework praised for its flexibility and user-friendliness. Built for Python, it's favored by machine learning developers and data scientists for tasks like image recognition and language processing. Notable for its GPU support and reverse-mode auto-differentiation, PyTorch allows for dynamic computation graphs, facilitating rapid experimentation and prototyping [48].

3.2.8 Pytorch Lightning

PyTorch Lightning is a lightweight wrapper for PyTorch that streamlines the training process for deep learning models. It simplifies the training loop, reducing boilerplate code and enabling researchers and practitioners to concentrate on model architecture and experiment setups [23].

3.2.9 Streamlit

Streamlit is a free, open-source framework tailored for machine learning engineers to quickly create and share attractive web applications for data science and machine learning projects. Built in Python, it simplifies the app development process, enabling users to effortlessly display data and gather model parameters with minimal code, catering to those who prioritize ease of use over extensive web development knowledge [54].

3.3 Call Detail Record (CDR) Data

3.3.1 Definition

Call detail records collect data about calls on telephone networks such as caller and recipient's names and numbers, call date and time, call duration, and various usage and diagnostic details like features used and call termination reasons. CDRs are regularly gathered for the purpose of creating reports on usage, capacity, performance, and diagnostics. Having this information makes it simpler to identify deviations from typical calling habits, like calls made outside of working hours, international calls, noticeable changes from previous reports, and calls to destinations that are not usual for the company [36].

Expanding on this basic knowledge, this research utilizes a particular set of CDRs from Djezzy, a top telecom provider in Algeria. This dataset is crucial as it contains a detailed collection of telecommunications interactions that are vital for identifying fraudulent activities like SIM Box fraud. The abundant data in Djezzy's complex telecommunications system provides a special chance to uncover and recognize fraudulent patterns. Djezzy's CDR data is crucial for various reasons. To begin with, it represents a rich aggregation of real-world telecommunications interactions, which includes time-stamped call details, call durations, the locations where calls were made and received, and other relevant metadata. Detailed records are crucial for developing strong machine learning models that can detect abnormalities indicative of fraudulent behavior. Additionally, Djezzy, a key player in the telecom industry in Algeria, faces a variety of complex efforts in telecom fraud, which highlights the importance of their CDR data in this research. Analyzing this dataset provides valuable information to help Djezzy improve their fraud detection abilities and also adds to the overall knowledge of telecom fraud patterns in the area.

3.3.2 Data Overview

The dataset provided by Djezzy, critical for this thesis, consists of Call Detail Records (CDRs) specifically for the "Gross-Ads," which are SIM cards sold within the last two months (1st March to 30th April 2024). Furthermore, it is important to note that personal information contained within these datasets has been encrypted. This measure ensures the protection of individual privacy and complies with data protection regulations, while still allowing for a comprehensive analysis of usage patterns necessary for detecting SIM box fraud. Additionally, we focused on the calls made during the last two days of each subscriber's most recent call activity. This targeted approach is crucial because fraudulent users often operate the SIM card like a regular user for several days, then switch to using it in a SIM box. By analyzing the last two days of call activity, we aim to capture this shift in behavior and effectively identify fraudulent usage patterns.

The dataset provided includes two distinct text files in TSV (Tab-Separated Values) format, one containing Fraudulent data and the other containing Non-Fraudulent data.**Table[3.1]** depicts the description of the dataset used. Each file organizes the data into a structured format, where the features relevant to our analysis are laid out. **Table[3.2]** lists these features used along with a brief description of each to provide clarity on the data being used. Each feature these files plays a crucial role in identifying patterns that may indicate fraudulent activity. The use of encrypted identifiers for personal information, Phone Number and Caller/Called Number ensures that privacy is maintained while still allowing a comprehensive analysis of communication patterns.

Data **Proprites**	**Fraudulent Data**	**Non-Fraudulent Data**
Total Number of Objects	1,048,575	22,922,715
Dimension	11	10
Data type	Categorical, integer	Categorical, integer
Data Size	110MB	2.5GB

TABLE 3.1: Description of Dataset used

Feature Name	**Description**
Phone Number	Phone number generating the CDR.
Caller/Called Number	Phone number of the caller or the called person.
Call Type	Outgoing/Incoming call.
Time Stamp	Date and Time the call was generated.
Duration	Call duration.
Destination	On-net/Off-net/International call.
Cell ID	Unique identifier assigned to a cell site.
Baring	Only for fraud data, Date and Time of blocking the phone number.

TABLE 3.2: Original Features and Their Descriptions in the CDR Dataset

3.4 Working Plan

Figure [3.1] provides a comprehensive summary of our working plan for the machine learning-based SimBox fraud detection. It illustrates the following key stages:

1. Preprocessing: Fraud and Non-Fraud Data undergo conversion to CSV format, removal of irrelevant features and duplicates, data type conversion, imputation, recent call extraction, feature engineering, and outlier removal. These streams are merged, standardized, and split into datasets.

2. Data Splitting: The processed data is divided into Test Set and Train Set.
3. Data Balancing: Applied to the Train Set, using oversampling techniques (SMOTE, SMOTETomek, AdaSyn) and undersampling (Tomek + RUS) to address class imbalance.
4. Model Training and Selection: 25 models including Linear SVM, RBF SVM, ANN, Random Forest, and XGBoost are trained using the balanced Train Data. ROC curve analysis is used to select the best-performing model.
5. Model Evaluation: The selected model is evaluated using the Test Data, employing metrics such as Accuracy, Precision, Recall, and F1-Score.

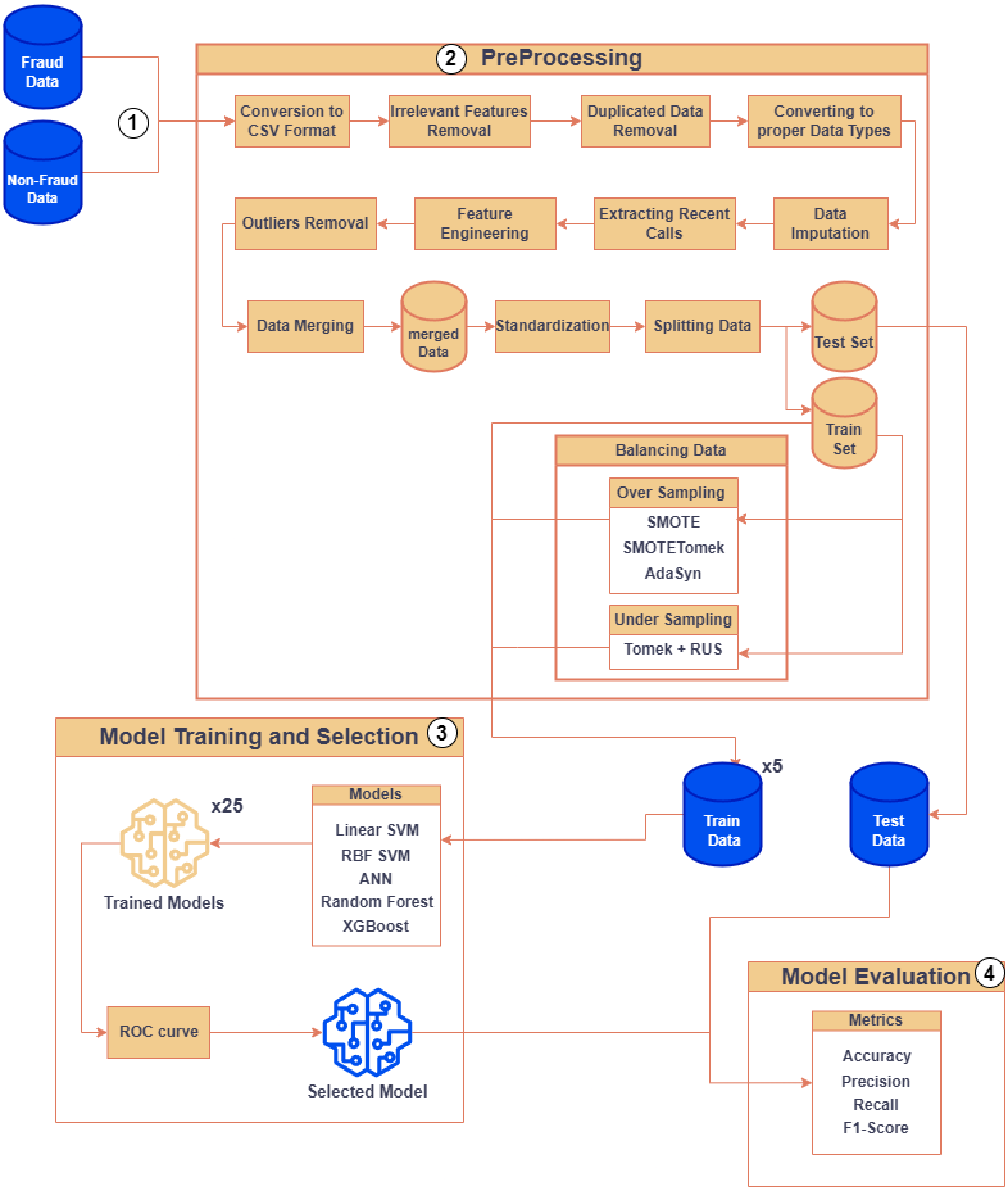


FIGURE 3.1: Project Working Plan

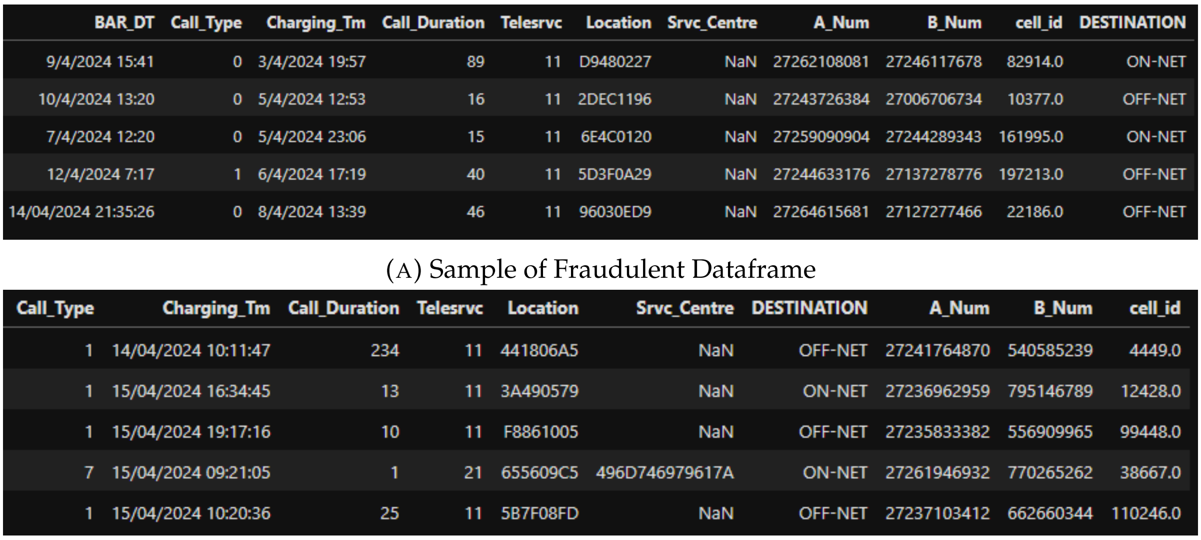
3.5 Data Preprocessing

The preprocessing of the dataset forms a crucial part of the implementation phase, as it ensures that the data used for training the machine learning models is clean and relevant.

We maintained separate datasets for fraud and non-fraud transactions and applied initial preprocessing steps to each dataset individually. This approach was essential because the characteristics and distribution of outliers differ significantly between fraud and non-fraud transactions. By removing outliers in their respective datasets, We ensured that the unique outlier patterns specific to each category were accurately addressed. Following the removal of outliers, We merged the datasets and proceeded with the remaining preprocessing steps on the combined dataset. This method ensured that the preprocessing was tailored appropriately to the distinct nature of each dataset, ultimately improving the quality and reliability of the subsequent analysis.

3.5.1 Conversion to CSV Format

The initial stage of the preprocessing involved converting the raw data, originally provided in text files in TSV format, into Comma-Separated Values (CSV) format. This conversion was crucial as CSV files are easier to handle and are widely supported by data analysis tools and libraries. The conversion process ensured that each piece of data was accurately segmented into distinct columns, facilitating easier manipulation and analysis in subsequent steps. The CSV files were then represented as DataFrames in Python for proper data manipulation and analysis. **Figure [3.2]** shows samples of fraudulent and non-fraudulent data.



BAR_DT	Call_Type	Charging_Tm	Call_Duration	Telesrvc	Location	Srvc_Centre	A_Num	B_Num	cell_id	DESTINATION
9/4/2024 15:41	0	3/4/2024 19:57	89	11	D9480227	NaN	27262108081	27246117678	82914.0	ON-NET
10/4/2024 13:20	0	5/4/2024 12:53	16	11	2DEC1196	NaN	27243726384	27006706734	10377.0	OFF-NET
7/4/2024 12:20	0	5/4/2024 23:06	15	11	6E4C0120	NaN	27259090904	27244289343	161995.0	ON-NET
12/4/2024 7:17	1	6/4/2024 17:19	40	11	5D3F0A29	NaN	27244633176	27137278776	197213.0	OFF-NET
14/04/2024 21:35:26	0	8/4/2024 13:39	46	11	96030ED9	NaN	27264615681	27127277466	22186.0	OFF-NET

(A) Sample of Fraudulent Dataframe

Call_Type	Charging_Tm	Call_Duration	Telesrvc	Location	Srvc_Centre	DESTINATION	A_Num	B_Num	cell_id
1	14/04/2024 10:11:47	234	11	441806A5	NaN	OFF-NET	27241764870	540585239	4449.0
1	15/04/2024 16:34:45	13	11	3A490579	NaN	ON-NET	27236962959	795146789	12428.0
1	15/04/2024 19:17:16	10	11	F8861005	NaN	OFF-NET	27235833382	556909965	99448.0
7	15/04/2024 09:21:05	1	21	655609C5	496D746979617A	ON-NET	27261946932	770265262	38667.0
1	15/04/2024 10:20:36	25	11	5B7F08FD	NaN	OFF-NET	27237103412	662660344	110246.0

(B) Sample of Non-Fraudulent Dataframe

FIGURE 3.2: Dataframes Samples

3.5.2 Removing Irrelevant Features

Once the data was converted into CSV format, the next step involved scrutinizing the dataset to identify and remove unnecessary features. These features present information that was not relevant to the analysis of SIM box fraud like "Location" and "Service Center".

3.5.3 Removing Duplicated Data

Having duplicate information can result in partial or incorrect outcomes while building and evaluating models. As a result, a comprehensive review of the dataset was carried out in order to detect and remove any duplicate records. This procedure ensured that each data entry was distinct, which is crucial for preserving the accuracy and dependability of the predictive models created in the subsequent phases of this research.

3.5.4 Converting Data Types

In our dataset, certain time and date features are currently stored as object data types ("Time Stamp" and "Baring"). To ensure consistency and facilitate easier analysis, we convert these features to the appropriate datetime data type. This conversion allows us to perform chronological operations and extract meaningful insights more efficiently.

3.5.5 Data Imputation

We addressed the issue of missing values, notably in the "Cell ID" column, which exhibited a 2.5% rate of missing data. To handle this, we first sorted the dataset by timestamp, arranging entries from the oldest date and time to the newest. Then, to maintain temporal consistency, missing "Cell ID" values were imputed with the last known cell visited by each user. This approach ensured that each user's trajectory remained coherent with their most recent location data. In cases where no previous information was available to infer the "Cell ID," we assigned a default value of 'unknown'. By employing this method and filling missing values with the preceding entry, we ensured the dataset's robustness, preserving data integrity throughout subsequent analyses.

3.5.6 Extracting Recent Call Activity

After ordering the data by timestamp, we focused on capturing the most recent call activities for each subscriber. Specifically, we retained the call records from the last two days of each subscriber's call history. This selection was based on the insight that fraudulent users typically operate the SIM card as a regular user for several days before employing it in a SIM box. By concentrating on the final two days of calls, we aim to detect this transition and identify patterns indicative of fraudulent behavior.

3.5.7 Feature Engineering

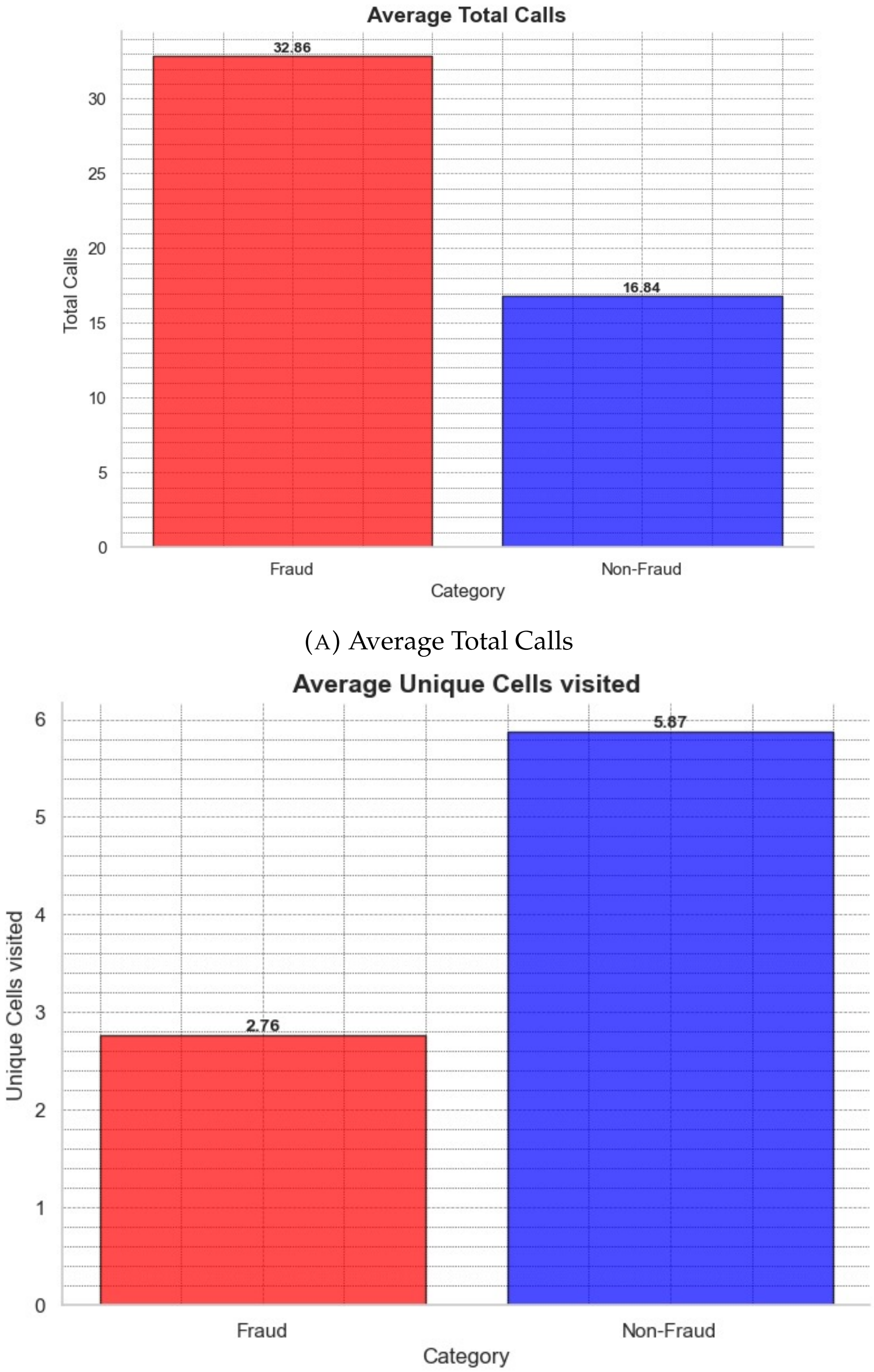
We focused on enhancing the predictive capacity of our models by deriving 24 new features from the existing dataset. These features were meticulously extracted from the original feature set, without directly utilizing the initial features for model training. This process involved a detailed analysis to uncover significant patterns and characteristics inherent in the data. By generating these additional features, we aimed to enrich the information available for model learning, thereby optimizing their performance and efficacy in addressing our research objectives.

- **Total Calls**: Total Calls: This feature involves tallying the overall number of calls (both outgoing and incoming). As illustrated in **Figure[3.3a]**, fraudulent numbers exhibit, on average, twice the total call volume compared to non-fraudulent phone numbers.

- **Unique Cells Visited**: This feature records the total number of unique cells visited. As depicted in **Figure[3.3b]**, fraudulent numbers demonstrate limited mobility, likely due to the stationary nature of SimBox devices. In contrast, non-fraudulent numbers tend to visit, on average, twice as many cells as fraudulent numbers.

- **Outgoing Calls by Total Calls Ratio**: This feature represents the ratio of outgoing calls to total calls, a significant metric as SimBox phone numbers often exhibit a higher frequency of outgoing calls compared to received calls. Referencing **Figure[3.4a]**, this disparity is visually evident.

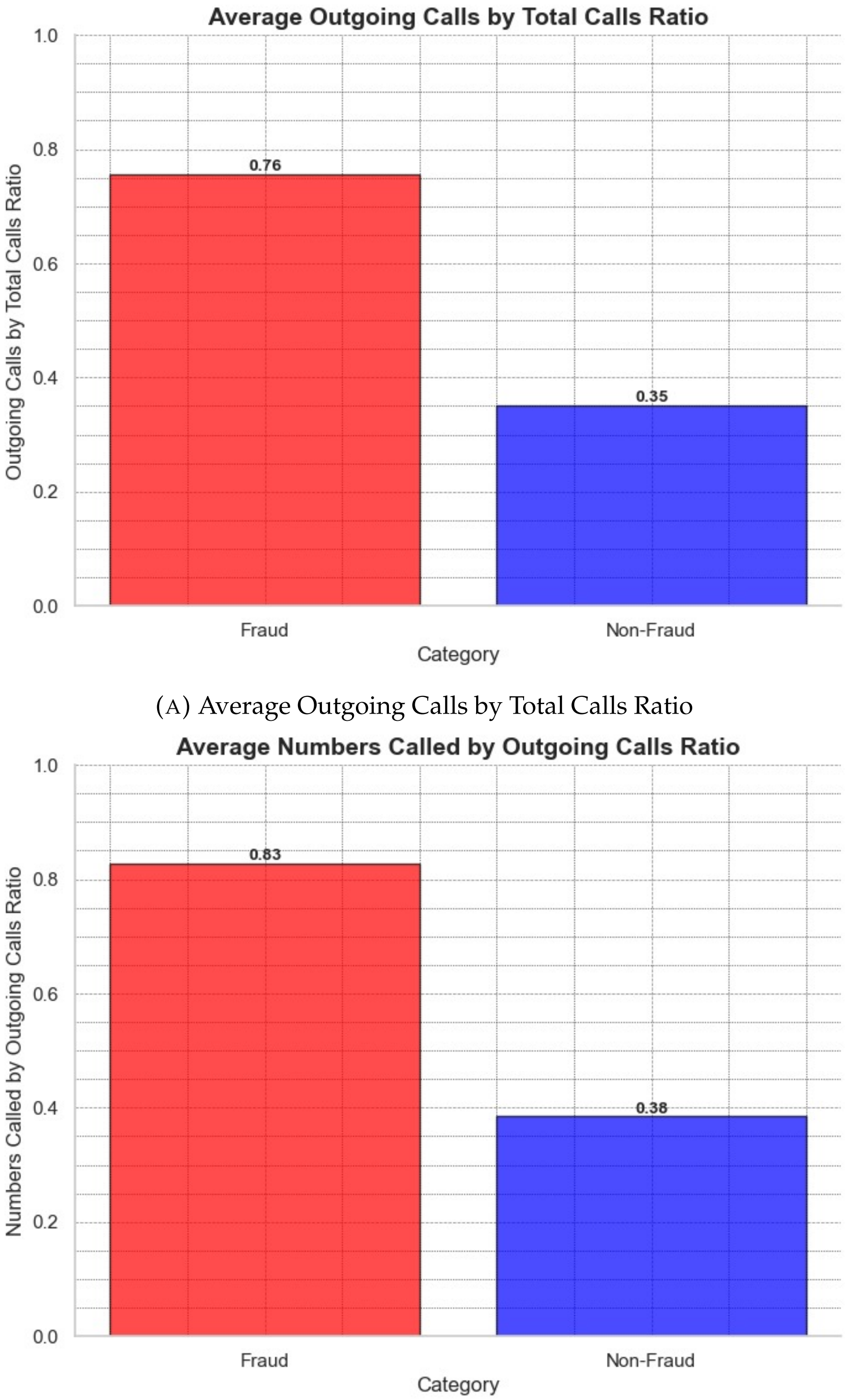
- **Unique Numbers Called by Outgoing Calls Ratio**: This feature calculates the ratio of unique numbers called to outgoing calls, serving as an indicator of calling behavior. Typically, individuals have a limited circle of contacts, resulting in a lower ratio of unique numbers called. Conversely, SimBox operations often involve frequent calls to a wide range of numbers, resulting in a higher ratio. As illustrated in **Figure[3.4b]**, the average proportion of calls made to new phone numbers by SimBox devices is 83%, whereas for non-fraudulent cases, it's only 36%.



(A) Average Total Calls

(B) Average Unique Cells Visited

FIGURE 3.3: Average Call and Cell Visitation Metrics



(A) Average Outgoing Calls by Total Calls Ratio

(B) Average Called Number by Outgoing Calls Ratio

FIGURE 3.4: Average Ratios Metrics

3.5.8 Removing Outliers

We addressed outliers within the dataset by removing data points beyond the 99th percentile. Outliers are extreme values that can skew the performance of machine learning models. We specifically removed outliers from features that are sums, such as "Total Calls" and "Total Night Calls." By removing these outliers, we ensured that our models were trained on more representative data, leading to improved performance and accuracy.

3.5.9 Data Merging

To distinguish between fraud and non-fraud data for our classification task, we created a new column called "Class" for both DataFrames to label these datasets. This column was assigned a value of 0 for non-fraudulent data and 1 for fraudulent data. After adding this column, we combined the two datasets into a single DataFrame. This merged dataset allows for a comprehensive analysis and model training, incorporating both fraud and non-fraud instances, and enabling the classification of data based on the "Class" column.

3.5.10 Standardization

Standardization is a vital step in preparing data for machine learning models. It involves adjusting the scale of features within a dataset to ensure uniformity and comparability. By transforming features to have a mean of 0 and a standard deviation of 1, standardization ensures that all features contribute equally during model training. For instance, in our dataset, features like "Total Duration" and "Total Calls" may have originally been measured in different units or scales. Standardizing these features ensures that the model treats them on an equal footing, thereby improving its performance and interpretability.

3.5.11 Splitting Data

In the process of splitting the data, 90% was allocated for training and 10% for testing. Within the test set, a balanced subset consisting of 75 positive cases and 75 negative cases was deliberately selected due to the very low number of positive cases. This approach ensures that the evaluation process includes an equal representation of both classes. By maintaining a balanced test set, the evaluation aimed to provide a more accurate assessment of the model's performance on both positive and negative instances, thereby enhancing the reliability and generalization capability of the models in real-world scenarios.

3.5.12 Balancing Data

The training data is extremely imbalanced, with approximately 99.88% negative instances and 0.12% positive instances, comprising 608,857 negative samples and 725 positive samples. This imbalance poses a substantial challenge for model training and performance evaluation, as it may lead to biased classifiers favoring the majority class. To address this issue and determine which balancing technique yields the best results, we experimented with various methods **Table[3.3]**.

For oversampling, we employed the Synthetic Minority Over-sampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), and a hybrid technique, SMOTE-Tomek. Using SMOTE, we balanced the dataset to have 608,857 positive and 608,857 negative samples. With SMOTETomek, we achieved 608,657 samples of each class. ADASYN resulted in 608,657 negative and 608,925 positive samples. Each resulting dataset was saved for further analysis.

For undersampling, we applied Tomek Links followed by Random Under-Sampling (RUS), resulting in a balanced dataset with 725 positive and 725 negative samples.

By training models on datasets balanced using each technique, we aimed to identify the most effective method for improving the classifier's ability to generalize effectively across both positive and negative classes.

Technique	Positive Instances	Negative Instances
Original Data	725	608,857
SMOTE	608,857	608,857
SMOTETomek	608,657	608,657
ADASYN	608,925	608,657
TomekLinks + RUS	725	725

TABLE 3.3: Class Distribution for Different Balancing Techniques

3.6 Model Training And Selection

In this part, we outline our approach to training machine learning models for detecting SIMBox fraud. We explore the effectiveness of various algorithms, including Support Vector Machines (SVM), Random Forests, XGBoost, and Artificial Neural Networks (ANN), across different datasets. Our primary aim is to strike a balance between minimizing false positives (incorrectly identifying non-fraudulent cases as fraudulent) and maximizing true positives (correctly identifying fraudulent cases). To achieve this balance, we rely on the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) to identify the most suitable model for SIMBox fraud detection, contributing to the robustness and security of telecommunication networks. Due to the high complexity of SVM, we encountered significant computational challenges when using the entire datasets balanced with oversampling techniques. Consequently, for the datasets manipulated with oversampling techniques, we restricted the training data to only 10% of the entire dataset, comprising 5% positive instances and 5% negative instances. This strategy was implemented to mitigate the computational burden associated with training SVM models, which proved to be excessively time-consuming. It's worth noting that the sampling process was both stratified and random, ensuring that the subset of data used for training remained representative of the overall dataset distribution while introducing variability necessary for model robustness. **Table[3.4]**.

Technique	**Positive Instances Used**	**Negative Instances Used**
Original Data	725	60,886
SMOTE	60,886	60,886
SMOTETomek	60,866	60,866
ADASYN	60,893	60,866
TomekLinks + RUS	725	725

TABLE 3.4: Used Distribution for Different Balancing Techniques

3.6.1 Artificial Neural Networks

According to [22] guidelines that are cited in Chapter 2, we constructed several neural networks with the same architecture **Table[3.5]**, each having two hidden layers with 16 neurons each. The output layer consists of a single neuron with a sigmoid activation function, suitable for our binary classification task of detecting SIMBox fraud. We utilized the Adam optimizer for its efficiency and ability to handle sparse gradients, with a learning rate set at 0.001 to ensure smooth convergence. To comprehensively evaluate performance, we trained separate neural networks on different datasets for 10 epochs: one on the original unbalanced dataset, and others on various balanced datasets. This approach allowed us to compare the performance of the same neural network structure, ensuring a robust evaluation. The ROC plot below illustrates the comparative performance of these models **Figure [3.5]**.

Input Layer	**Hidden Layer 1**	**Hidden Layer 2**	**Output Layer**	**Optimizer**	**Learning Rate**	**Epochs**
24	16	16	1	Adam	0.001	10

TABLE 3.5: ANN Model architecture

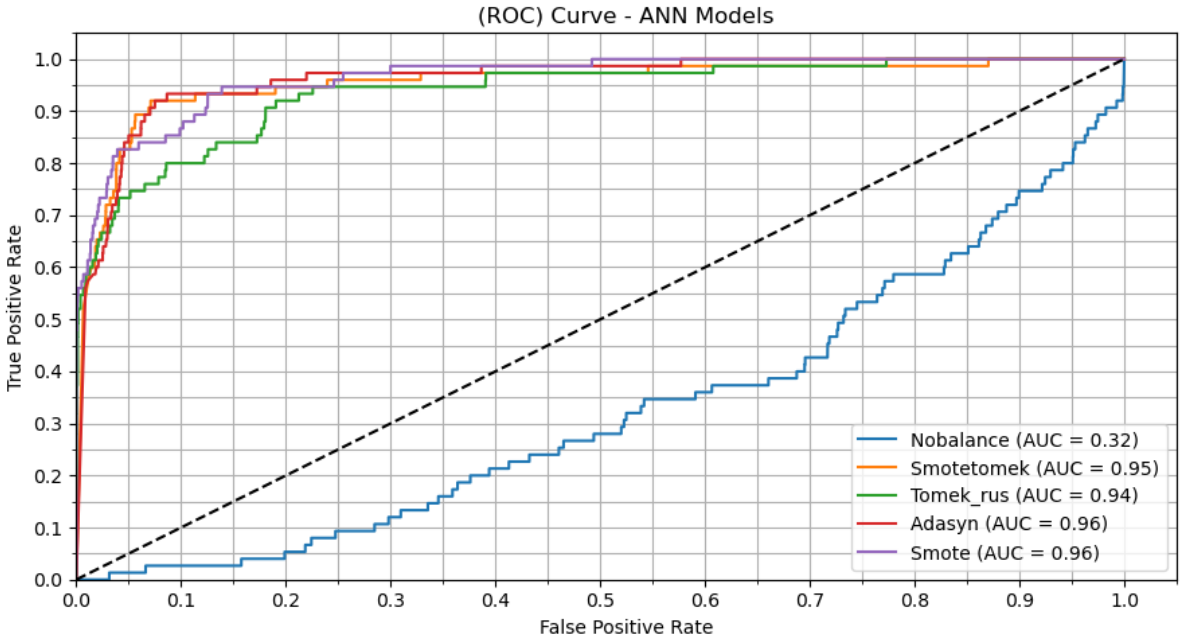
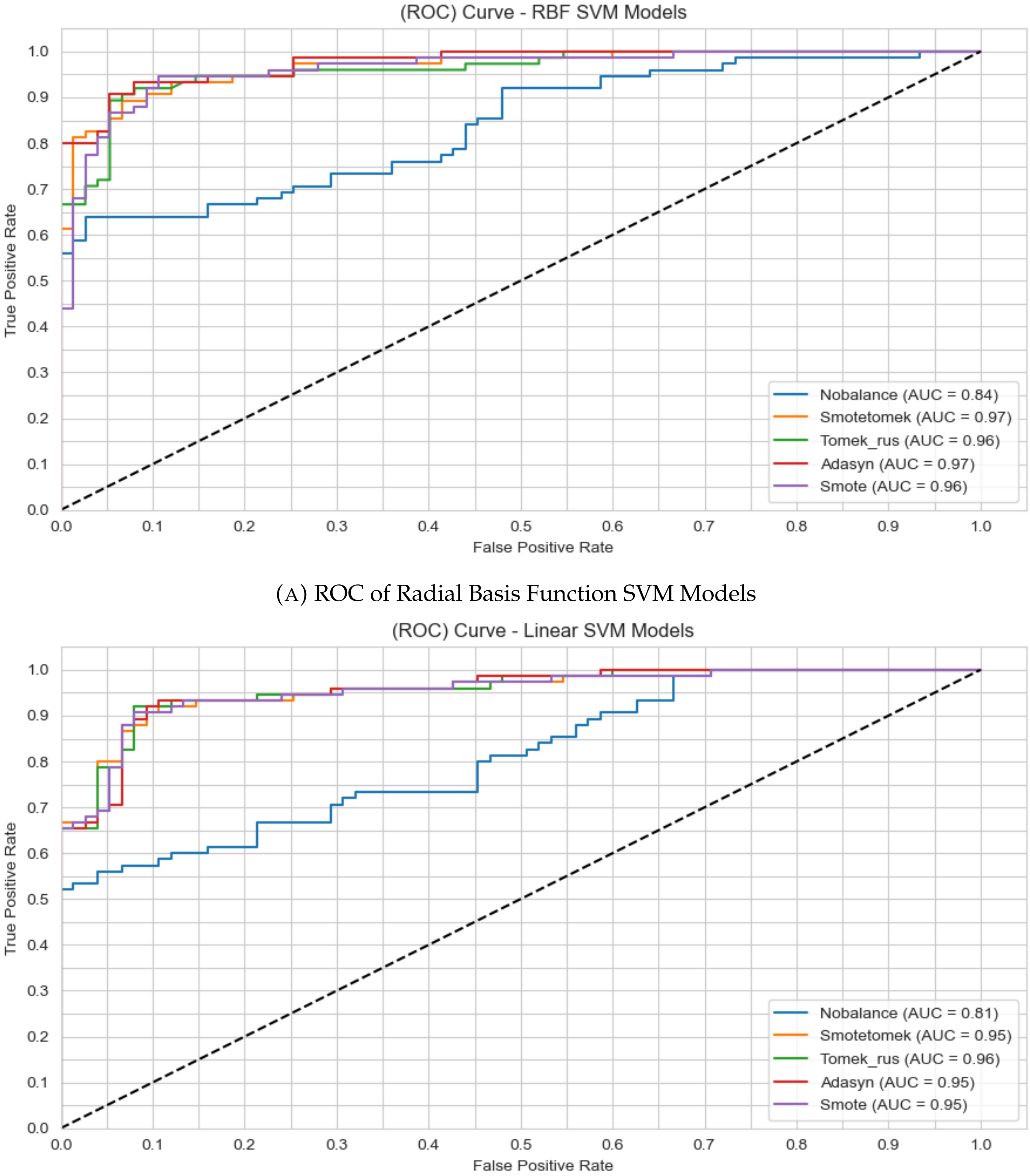


FIGURE 3.5: ROC of ANN Models

3.6.2 Support Vector Machine

We experimented with two different kernel functions: the Radial Basis Function (RBF) and the linear kernel. The RBF kernel is known for its ability to capture complex decision boundaries, while the linear kernel provides a simpler approach with linear decision boundaries. Each SVM model was trained on separate dataset **Figure [3.6]**.



(A) ROC of Radial Basis Function SVM Models

(B) ROC of Linear SVM Models

FIGURE 3.6: ROC of SVM Models

3.6.3 Random Forest

For our Random Forest models, we utilized an ensemble of 100 trees. Each model was trained on different versions of our dataset, including the original unbalanced dataset. The ROC plot demonstrates the effectiveness of these models across the different datasets **Figure [3.7]**.

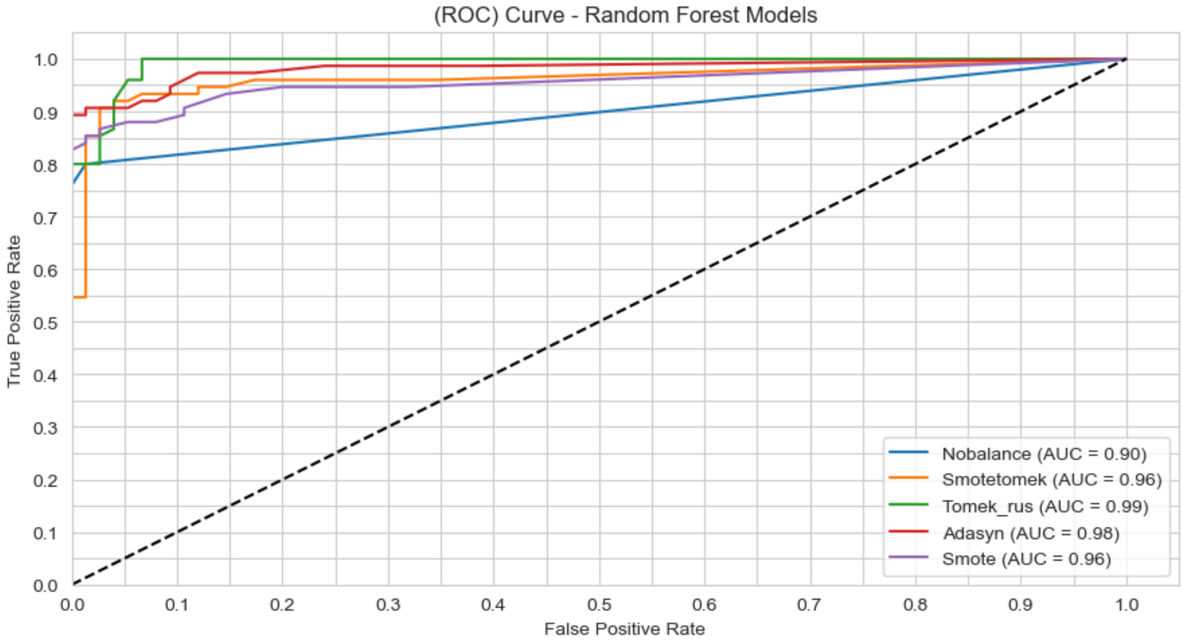


FIGURE 3.7: ROC of Random Forest Models

3.6.4 XGBoost

We trained XGBoost with maximum depth of 6 and a learning rate of 0.3, which are the default parameters in the XGBoost Python library. The resulting performance metrics are illustrated in **Figure [3.8]**.

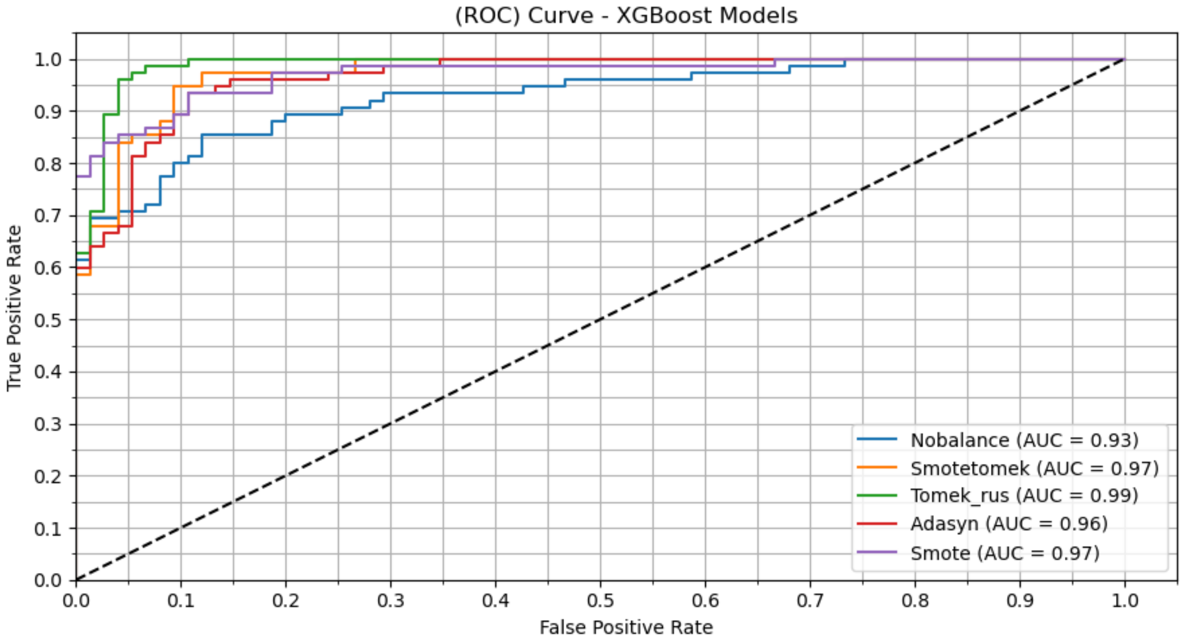


FIGURE 3.8: ROC of XGBoost Models

3.6.5 Model Selection

Our primary objective is to detect as many SIM cards used in SIMBox fraud as possible while minimizing false positives, where normal SIM cards used by legitimate users are mistakenly identified as fraudulent. Considering this trade-off, we found that XGBoost model trained on data balanced using Tomek Links followed by Random Undersampling offers the best performance for our problem. This model achieved an approximately 4% false positive rate and a high 96% true positive rate, making it the optimal choice for our SIMBox fraud detection task **Figure [3.9]**.

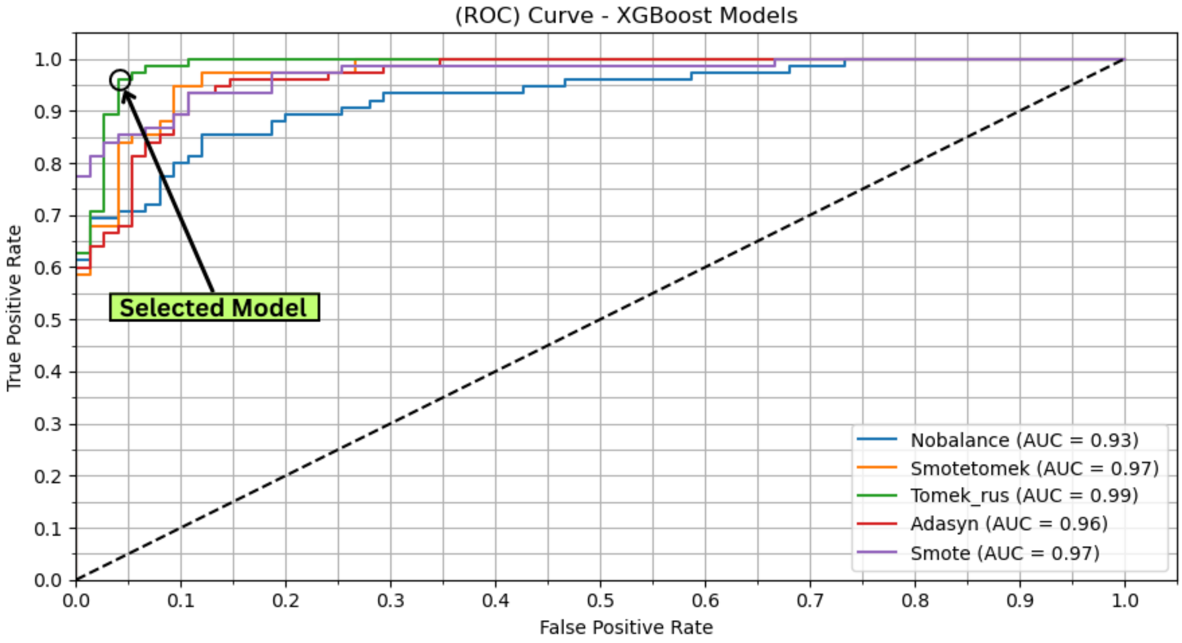


FIGURE 3.9: Selected Model

3.7 Model Evaluation

In this section, we will explore the metrics employed for evaluating the performance of our chosen model and subsequently assess its performance.

3.7.1 Performance Metrics

3.7.1.1 Accuracy

Accuracy is a measurement that calculates how often a machine learning model accurately forecasts the result. It is determined by the total number of correct predictions divided by the overall number of predictions made [45]. It is defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

3.7.1.2 Recall

Recall is a metric that quantifies how effectively a machine learning model identifies positive instances (true positives) from the total actual positive samples in the dataset. It is calculated by dividing the number of true positives by the sum of true positives and false negatives [45]. It is defined as:

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

3.7.1.3 Precision

Precision is a metric that quantifies the frequency with which a machine learning model accurately predicts the positive class. It is calculated by dividing the number of true positive predictions by the total number of instances predicted as positive (including both true positives and false positives) [45]. It is defined as:

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

3.7.1.4 F1-Score

The F1-Score is the harmonic mean of Precision and Recall, providing a single metric that balances both concerns. The F1-Score is defined as:

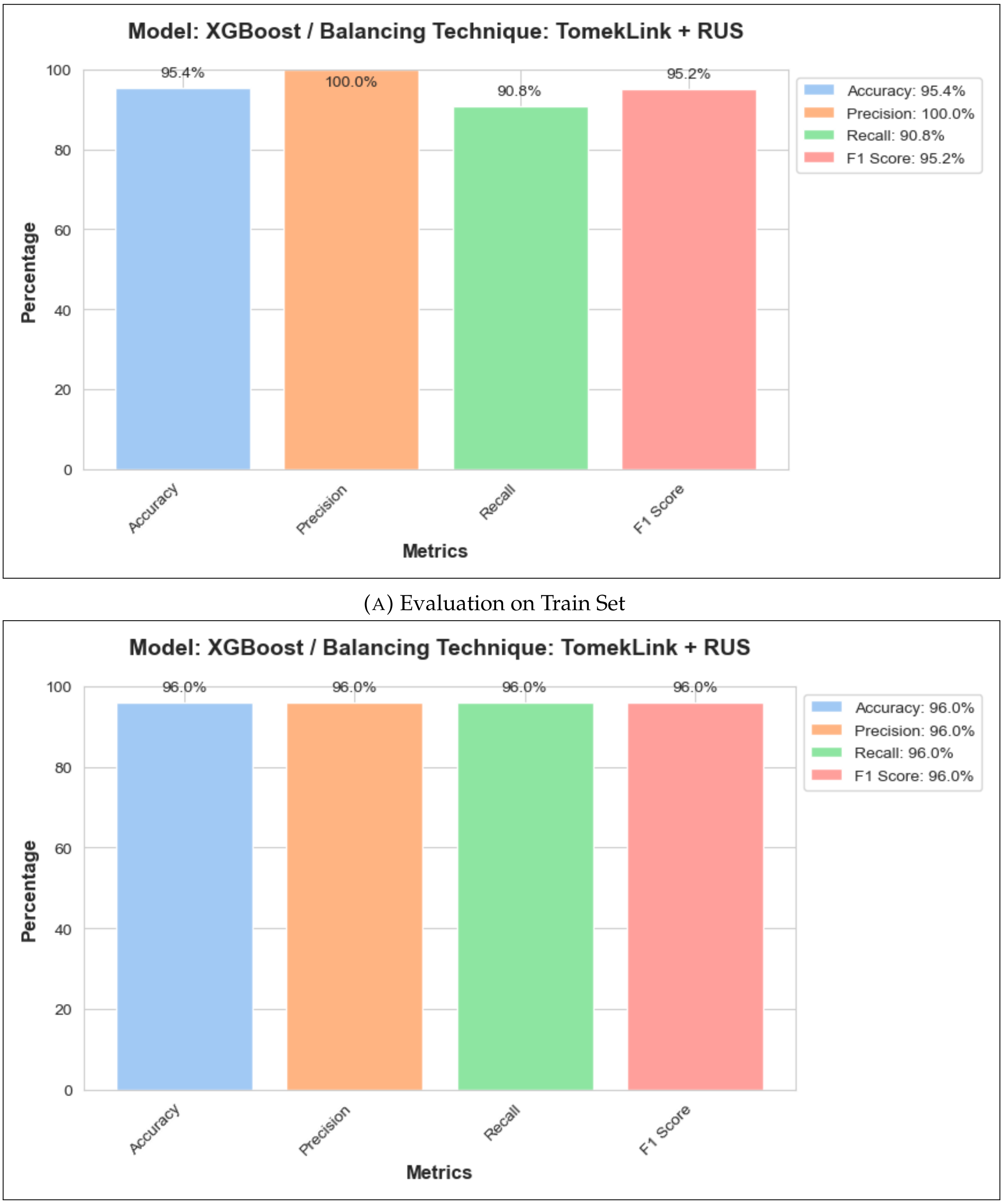
$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.7.2 Performance Evaluation

In this subsection, we will evaluate our selected XGBoost model using various evaluation metrics, including accuracy, precision, recall, and F1-score, on both the training and test datasets. **Table[3.6]** and **Figure [3.10]** summarize the results, confusion matrices **Figure [3.11]**.

Dataset	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Train Set	95.4	100.0	90.8	95.2
Test Set	96.0	96.0	96.0	96.0

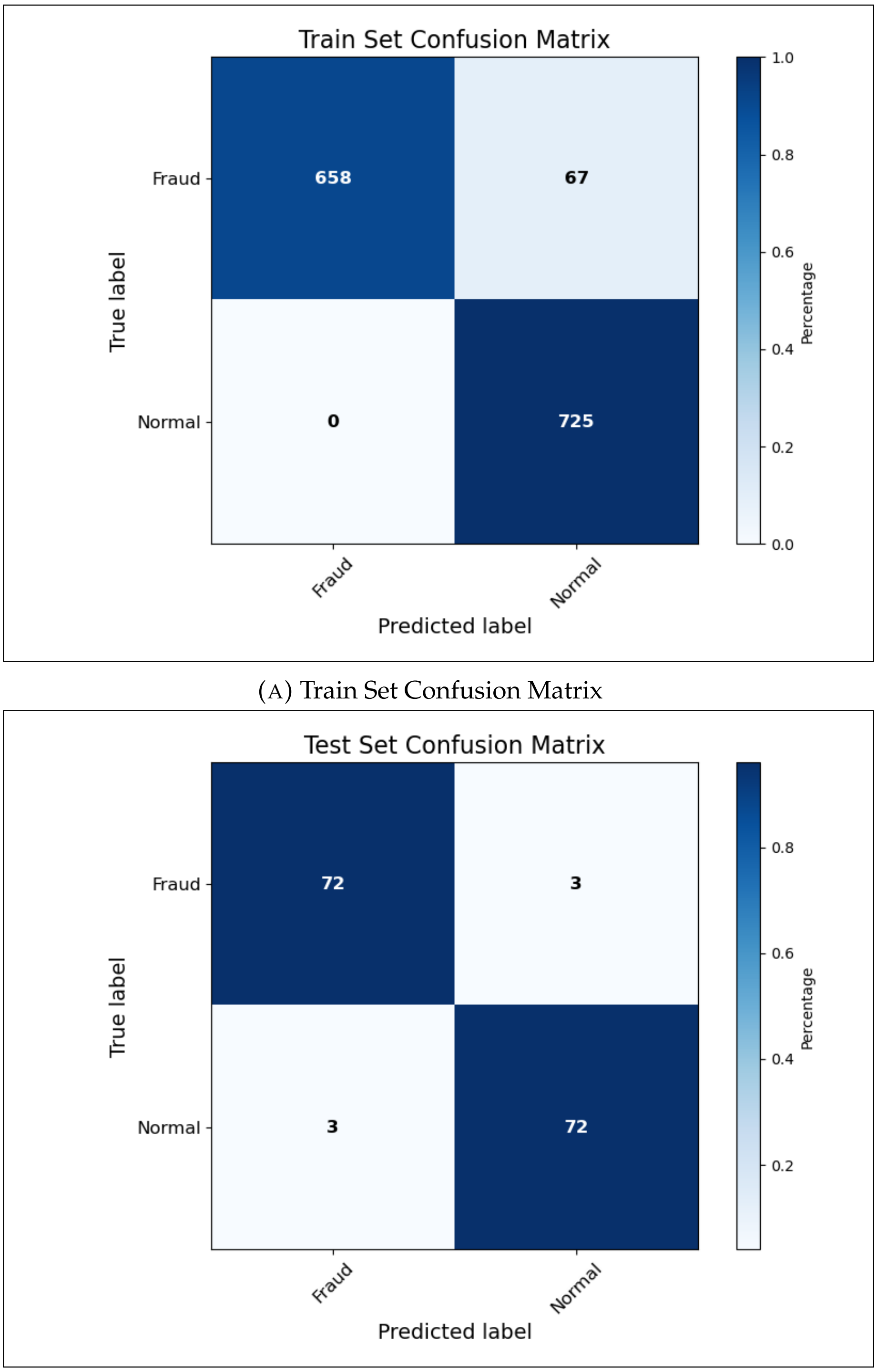
TABLE 3.6: Performance Metrics for Train and Test Sets



(A) Evaluation on Train Set

(B) Evaluation on Test Set

FIGURE 3.10: Evaluation on Train Set and Test Set



(A) Train Set Confusion Matrix

(B) Test Set Confusion Matrix

FIGURE 3.11: Confusion Matrix

3.7.3 Discussion

We explored several machine learning algorithms, including Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). Among these, XGBoost demonstrated the best performance in detecting fraudulent activities.

A significant challenge in fraud detection is the severe class imbalance in the dataset, with fraudulent activities being vastly outnumbered by legitimate ones. To address this issue, we implemented multiple data balancing techniques, including undersampling and oversampling. Our experiments showed that the combination of Tomek Links and Random Undersampling (RUS) yielded the most effective results when used with the XGBoost model.

The Tomek Links method helped in cleaning the dataset by removing borderline instances, which often contribute to noise and overlap between classes. Following this, Random Undersampling was applied to further balance the class distribution by reducing the number of majority class instances. This sequential application of Tomek Links and RUS significantly improved the performance metrics of the XGBoost model, leading to a higher detection rate of fraudulent activities and a reduction in false positives and false negatives.

XGBoost not only achieved approximately 96% in accuracy, recall, precision, and F1-score, but it also provided the best tradeoff between false positive rate and true positive rate. This balance is crucial in fraud detection, as it ensures that legitimate Sim cards are not incorrectly flagged as fraudulent while maintaining a high detection rate for actual fraudulent activities.

The superior performance of XGBoost can be attributed to its robust handling of imbalanced data and its ability to model complex interactions between features. XGBoost's gradient boosting framework efficiently combines the strengths of multiple weak learners, resulting in a powerful and accurate predictive model.

In comparison, other models like ANN, SVM, and RF did not perform as well, despite various attempts at data balancing. This suggests that XGBoost's algorithmic advantages, make it particularly suitable for fraud detection tasks where data imbalance is a critical issue.

Overall, our study highlights the importance of selecting appropriate data balancing techniques and machine learning algorithms in the context of fraud detection. The combination of Tomek Links and RUS, along with the XGBoost model, offers a promising approach for effectively identifying fraudulent activities in highly imbalanced datasets, providing an optimal balance between false positive and true positive rates.

3.8 AI web-Application for SIMBox Fraud Detection

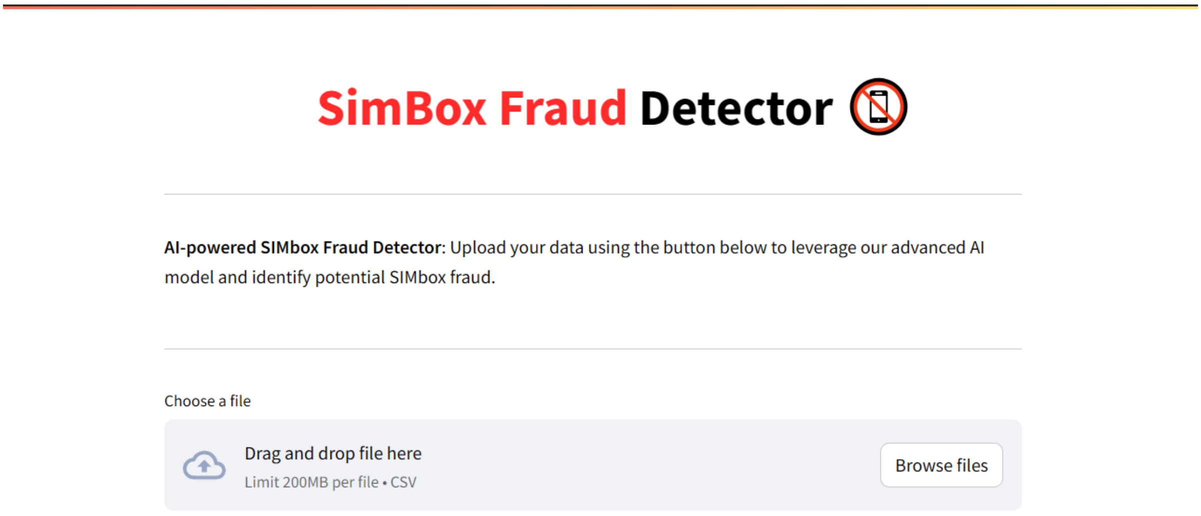
We have developed an AI-powered web application called SimBox Fraud Detector that harnesses the power of our AI model for SIMBox fraud detection. Through our platform, users can effortlessly upload datasets containing phone number records, initiating the fraud detection process.

Before detection, we apply rigorous preprocessing to the data. This involves cleaning, normalizing, and transforming the dataset to ensure optimal performance of our AI model. This preprocessing step includes handling missing values, removing duplicates, standardizing data types, and extracting relevant features to enhance the detection accuracy.

Upon upload, our advanced AI model swiftly analyzes the preprocessed dataset, flagging fraudulent numbers for further action. The results are presented in two distinct datasets for user convenience. The first dataset retains all original rows, with fraudulently identified numbers visually highlighted in red, enabling easy identification within the dataset. Moreover, our platform generates a second dataset exclusively featuring the identified fraudulent numbers. Users can conveniently download this dataset as a CSV file, facilitating additional investigation or remedial measures against the detected fraud.

With our intuitive web application and powerful AI-driven detection capabilities, we provide a seamless solution for combating SIMBox fraud, empowering users to safeguard their networks effectively.

Upon accessing the web application, users are presented with a clean and minimalistic interface, featuring a prominent button for uploading the dataset, as depicted in **Figure [3.12]**.

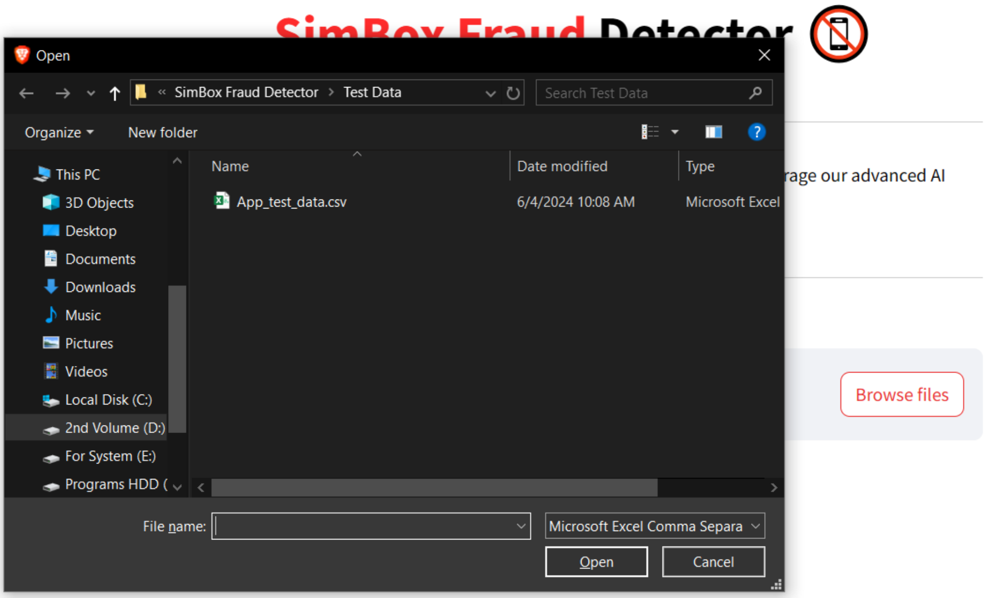


SimBox Fraud Detector 📵

AI-powered SIMbox Fraud Detector: Upload your data using the button below to leverage our advanced AI model and identify potential SIMbox fraud.

FIGURE 3.12: SimBox Fraud Detector Homepage

Users have the option to either drag and drop the dataset directly onto the designated area or click the 'Browse files' button to browse files on their device for upload **Figure [3.13]**.



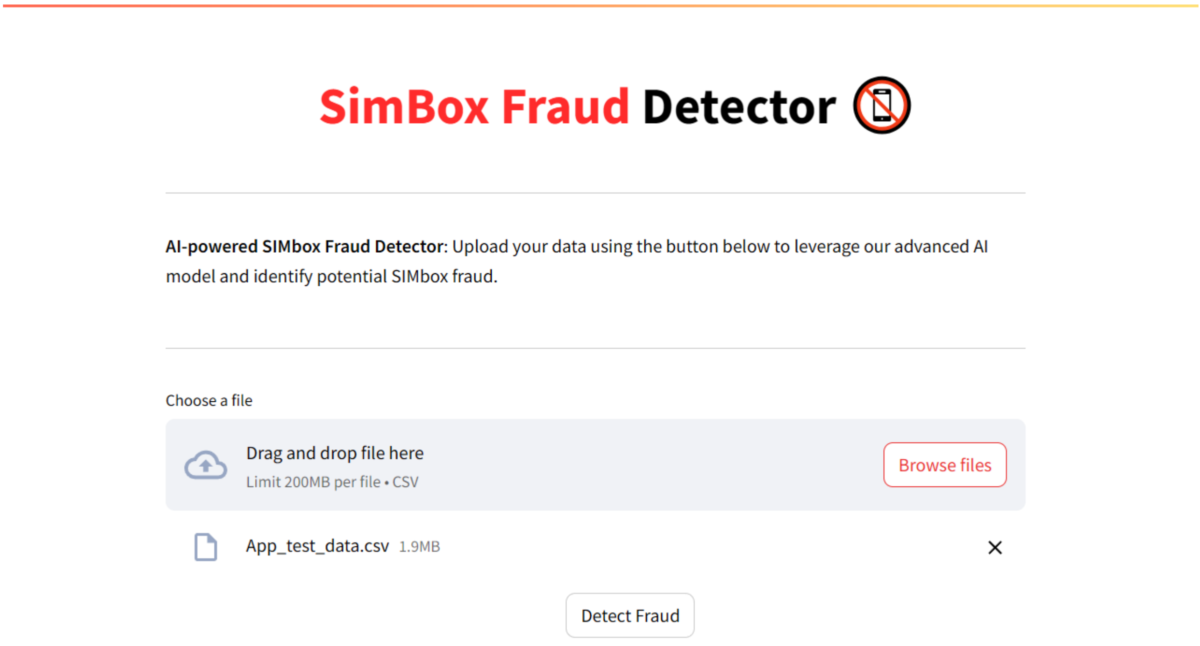
App_test_data.csv 6/4/2024 10:08 AM Microsoft Excel

File name:

Microsoft Excel Comma Separa

FIGURE 3.13: Uploading Data

After selecting the data, the name and size of the selected file will be displayed on the website interface. Additionally, users will see a 'Detect Fraud' button **Figure [3.14]**.



SimBox Fraud Detector

AI-powered SIMbox Fraud Detector: Upload your data using the button below to leverage our advanced AI model and identify potential SIMbox fraud.

Choose a file

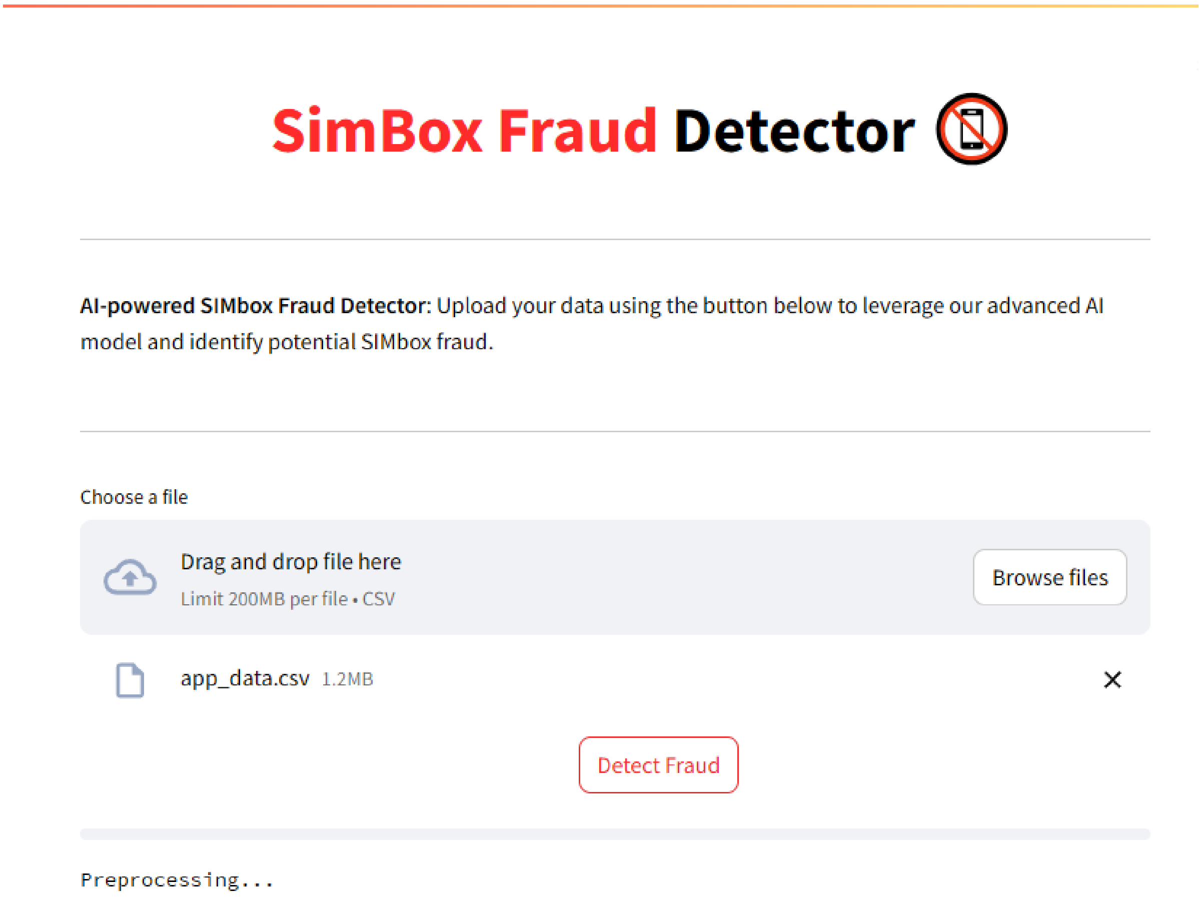
Drag and drop file here

Limit 200MB per file • CSV

App_test_data.csv 1.9MB

FIGURE 3.14: Pre-Detection Interface

Clicking on this 'Detect Fraud' button will initiate the procedure of detecting fraudulent numbers from the uploaded file. Prior to detection, the system conducts comprehensive preprocessing on the dataset. This preprocessing includes managing missing values, eliminating duplicates, standardizing data **Figure [3.15]**.

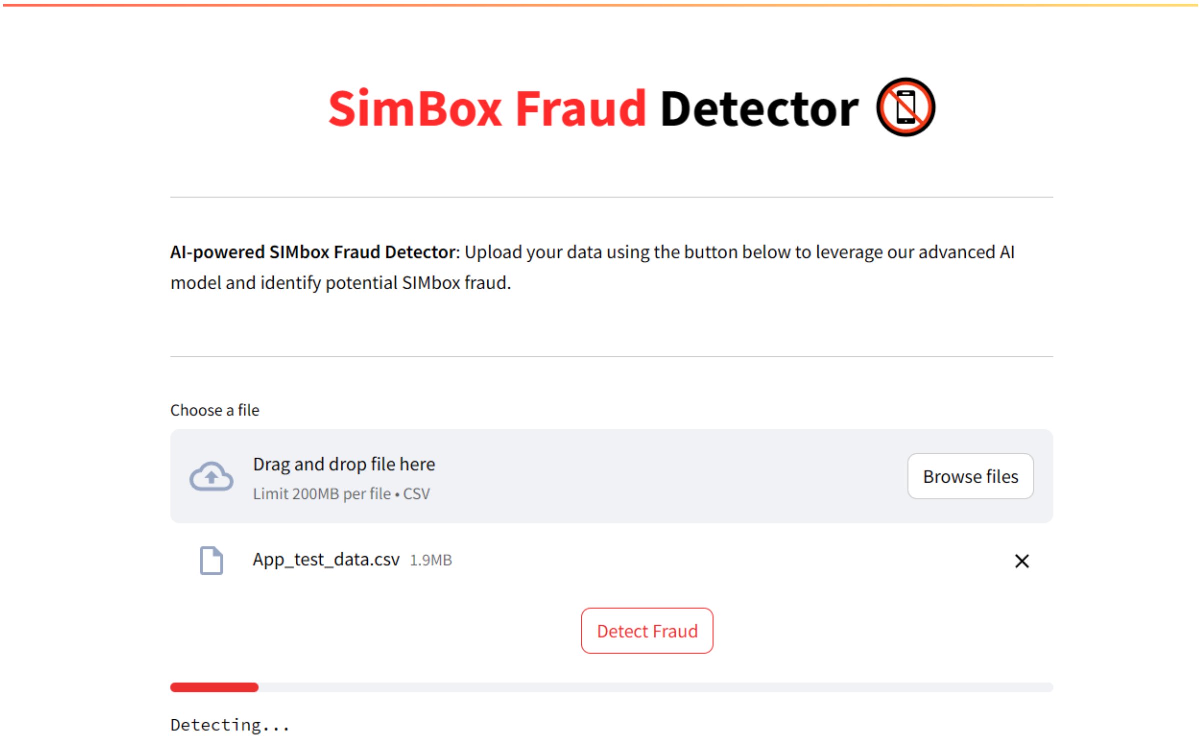


SimBox Fraud Detector

AI-powered SIMbox Fraud Detector: Upload your data using the button below to leverage our advanced AI model and identify potential SIMbox fraud.

```
Preprocessing...
```

(A) Preprocessing



SimBox Fraud Detector

AI-powered SIMbox Fraud Detector: Upload your data using the button below to leverage our advanced AI model and identify potential SIMbox fraud.

```
Detecting...
```

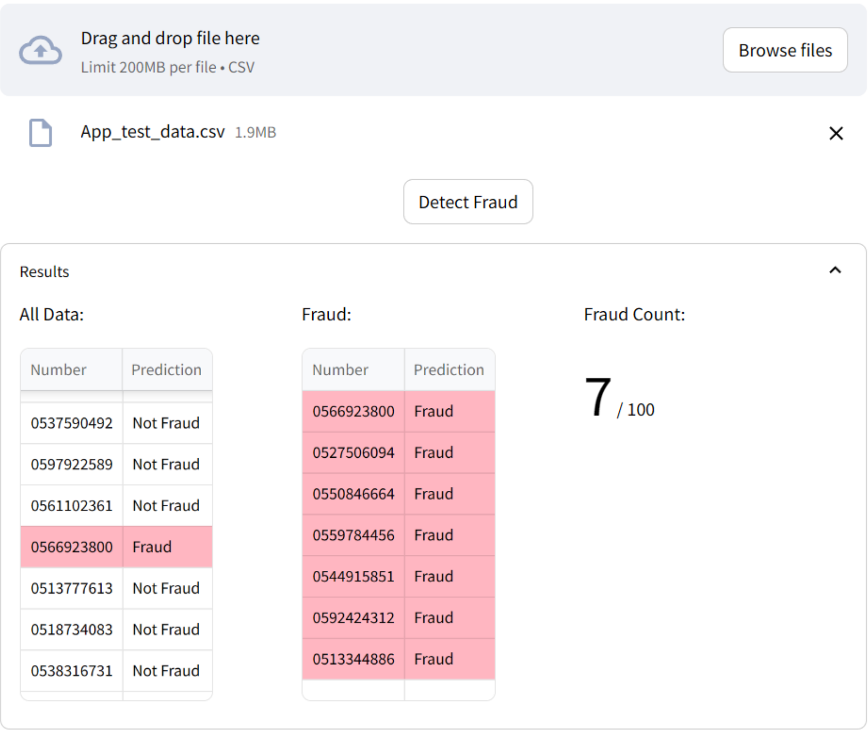
(B) Detection

FIGURE 3.15: Detection Process

After completing the detection process, the results are displayed in a container on the interface. This container includes two datasets that can be downloaded:

- The first dataset presents the full dataset, with fraudulent rows highlighted in red for easy identification.
- The second dataset exclusively contains the fraudulent numbers, also highlighted in red.

Additionally, the container prominently showcases the count of detected fraudulent numbers, providing users with a comprehensive overview of the fraud detection outcome **Figure [3.16]**.



All Data:

Number	Prediction
0537590492	Not Fraud
0597922589	Not Fraud
0561102361	Not Fraud
0566923800	Fraud
0513777613	Not Fraud
0518734083	Not Fraud
0538316731	Not Fraud

Fraud:

Number	Prediction
0566923800	Fraud
0527506094	Fraud
0550846664	Fraud
0559784456	Fraud
0544915851	Fraud
0592424312	Fraud
0513344886	Fraud

Fraud Count:

7 / 100

FIGURE 3.16: Detection Process

3.9 Conclusion

In this chapter, we explored the application of artificial intelligence methods for identifying SIM box fraud using a dataset provided by Djezzy, an Algerian telecommunications company. Our investigation into SIM box fraud detection is motivated by its significant financial risks and regulatory challenges faced by telecommunications companies globally. Throughout this chapter, we outlined a comprehensive approach consisting of several essential stages: preprocessing data, feature engineering, and training predictive models. Each stage was meticulously designed to address the unique challenges posed by SIM box fraud. We began by introducing the dataset, which comprised anonymized transactional telecommunication data, ensuring compliance with privacy regulations. Subsequently, we detailed the extensive preprocessing steps undertaken to clean and organize the data, followed by the deliberate development of features to highlight fraudulent activities. Finally, we discussed the selection of machine learning models to achieve accurate fraud prediction.

This section serves as a foundational component of the thesis, bridging theoretical methodologies from earlier sections to practical implementations and results. The insights gained from our experimentation pave the way for future advancements in SIM box fraud detection and contribute to the ongoing efforts to combat fraudulent activities in the telecommunications industry.

General Conclusion

In this Master's thesis, we have explored the application of artificial intelligence (AI) techniques for the detection of SIM box fraud in mobile telecommunications networks. Beginning with an examination of the structure and operations of mobile telephony networks, we identified SIM box fraud as a significant challenge facing telecommunications companies, posing financial risks and regulatory hurdles.

Through a comprehensive review of SIM box fraud detection techniques, including machine learning algorithms, deep learning approaches, and data balancing methods, we highlighted the complexity of detecting fraudulent activities in telecommunications data. Our investigation revealed the importance of fast and accurate detection methods to combat the evolving nature of fraud schemes.

The implementation and experimental results presented in this Master's thesis demonstrated the efficacy of AI in detecting SIM box fraud. By preprocessing call detail record (CDR) data and employing advanced machine learning models such as XGBoost with Tomek+RUS balancing technique, we achieved promising results in identifying fraudulent activities. Specifically, our model attained an accuracy of 96%, demonstrating its reliability in distinguishing between legitimate and fraudulent activities. Furthermore, the model achieved a recall of 96%, a precision of 96%, and an F1 score of 96%, indicating a high level of performance in both detecting fraud and minimizing false positives.

Our research contributes to the body of knowledge in the field of fraud detection and underscores the potential of AI technologies to enhance security and mitigate financial risks in mobile telecommunications networks. By leveraging the insights gained from this study, telecommunications companies can develop more robust fraud detection systems to safeguard their networks and protect against fraudulent activities.

Moving forward, further research is needed to explore emerging trends in fraud detection and to develop more sophisticated AI-based approaches capable of addressing the evolving tactics employed by fraudsters. Additionally, collaboration between industry stakeholders, regulatory bodies, and academic researchers is essential to ensure the continued advancement of fraud detection technologies and the protection of telecommunications networks worldwide.

Overall, this Master's thesis represents a significant step towards understanding and combating SIM box fraud, laying the groundwork for future advancements in fraud detection and prevention in the telecommunications industry.

Communication

Ahmed Rami Bouguettoucha and Aicha Aggoune. "Supervised Machine Learning Approaches for SIM Box Fraud Detection: Review of Current Works". In: *Advances in Telecommunication Electronics and Computer Engineering (ATECE'24)*. Algeria, Khenchela, 2024

Bibliography

[1] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.

[2] What is Anaconda? `https://domino.ai/data-science-dictionary/anaconda`. (Accessed: 2024-05-06).

[3] Bart Baesens et al. "robROSE: A robust approach for dealing with imbalanced data in fraud detection". In: *Statistical Methods & Applications* 30.3 (2021), pp. 841–861.

[4] Deepak Kumar Barik et al. "Design and Analysis of RF Optimization in 2G GSM and 4G LTE Network". In: *Innovation in Electrical Power Engineering, Communication, and Computing Technology: Proceedings of Second IEPCCT 2021*. Springer. 2022, pp. 11–18.

[5] Richard A Becker, Chris Volinsky, and Allan R Wilks. "Fraud detection in telecommunications: History and lessons learned". In: *Technometrics* 52.1 (2010), pp. 20–33.

[6] Ahmed Rami Bouguettoucha and Aicha Aggoune. "Supervised Machine Learning Approaches for SIM Box Fraud Detection: Review of Current Works". In: *Advances in Telecommunication Electronics and Computer Engineering (ATECE'24)*. Algeria, Khenchela, 2024.

[7] SIM box. `https://en.antrax.mobi/products/simbox/`. (Accessed: 2024-02-13).

[8] Leo Breiman and RA Cutler. "Random forests machine learning [J]". In: *journal of clinical microbiology* 2 (2001), pp. 199–228.

[9] SIM Card. `https://umutcanbolat.com/sim-card-basics/`. (Accessed: 2024-02-20).

[10] Hubert Cardot. *Recurrent neural networks for temporal data processing*. BoD–Books on Demand, 2011.

[11] Mayank Arya Chandra and SS Bedi. *Survey on SVM and their application in image classification*. 2021.

[12] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[13] Tianqi Chen et al. *Xgboost: extreme gradient boosting*. 2015.

[14] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. "Supervised learning". In: *Machine learning techniques for multimedia: case studies on organization and retrieval*. Springer, 2008, pp. 21–49.

[15] Eric Michel Deussom Djomadji et al. "Machine Learning-Based Approach for Identification of SIM Box Bypass Fraud in a Telecom Network Based on CDR Analysis: Case of a Fixed and Mobile Operator in Cameroon". In: *Journal of Computer and Communications* 11.2 (2023), pp. 142–157.

[16] Weijiang Feng et al. *Audio visual speech recognition with multimodal recurrent neural networks*. 2017. DOI: 10.1109/IJCNN.2017.7965918.

[17] Alberto Fernández et al. "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary". In: *Journal of artificial intelligence research* 61 (2018), pp. 863–905.

[18] Fraud 101: What Is Fraud? https://www.acfe.com/fraud-resources/fraud-101-what-is-fraud/. (Accessed: 2023-02-12).

[19] Palak Gupta et al. "Unbalanced Credit Card Fraud Detection Data: A Machine Learning-Oriented Comparative Study of Balancing Techniques". In: *Procedia Computer Science* 218 (2023), pp. 2575–2584.

[20] Basna Mohammed Salih Hasan and Adnan Mohsin Abdulazeez. *A review of principal component analysis algorithm for dimensionality reduction*. 2021.

[21] Haibo He et al. *ADASYN: Adaptive synthetic sampling approach for imbalanced learning*. Ieee, 2008.

[22] James B Heaton, Nick G Polson, and Jan Hendrik Witte. *Deep learning for finance: deep portfolios*. 2017.

[23] Chadi Helwe, Chloé Clavel, and Fabian Suchanek. *LogiTorch: A PyTorch-based library for logical reasoning on natural language*. 2022.

[24] Tomek Ivan. "Two modifications of CNN". In: *IEEE transactions on Systems, Man and Communications, SMC* 6 (1976), pp. 769–772.

[25] Christian Janiesch, Patrick Zschech, and Kai Heinrich. *Machine learning and deep learning*. 2021.

[26] What is Jupyter Notebook? `https://domino.ai/data-science-dictionary/jupyter-notebook`. (Accessed: 2024-05-06).

[27] AVVS Karunathilaka. "Fraud Detection on International Direct Dial Calls". PhD thesis. 2021.

[28] Mhair Kashir and Sajid Bashir. "Machine learning techniques for sim box fraud detection". In: *2019 International Conference on Communication Technologies (ComTech)*. IEEE. 2019, pp. 4–8.

[29] Anne Josiane Kouam. "Bypass frauds in cellular networks: Understanding and Mitigation". PhD thesis. Ecole Polytechnique (Palaiseau, France), 2023.

[30] Anne Josiane Kouam, Aline Carneiro Viana, and Alain Tchana. "SIMBox bypass frauds in cellular networks: Strategies, evolution, detection, and future directions". PhD thesis. 2021, pp. 2295–2323.

[31] Takio Kurita. *Principal component analysis (PCA)*. 2019.

[32] Fei-Fei Li, Andrej Karpathy, and Justin Johnson. *Convolutional neural networks for visual recognition*. 2015.

[33] Charles X Ling and Chenghui Li. "Data mining for direct marketing: Problems and solutions." In: *Kdd*. Vol. 98. 1998, pp. 73–79.

[34] Qing Lv, Suzhen Zhang, and Yuechun Wang. *Deep learning model of image classification using machine learning*. 2022.

[35] Oded Z Maimon and Lior Rokach. *Data mining with decision trees: theory and applications*. Vol. 81. World scientific, 2014.

[36] Giridhar Maji, Sharmistha Mandal, and Soumya Sen. *Identification of city hotspots by analyzing telecom call detail records using complex network modeling*. 2023.

[37] Abdullah Masrub and Mohamed Alahemar. "SIM Boxing Problem: ALMADAR ALJADID Case Study". In: *2020 International Conference on Electrical Engineering (ICEE)*. IEEE. 2020, pp. 1–5.

[38] Alicja Miniak-Górecka, Krzysztof Podlaski, and Tomasz Gwizdałła. *Using k-means clustering in python with periodic boundary conditions*. 2022.

[39] Tom M Mitchell. "Does machine learning really work?" In: *AI magazine* 18.3 (1997), pp. 11–11.

[40] Roweida Mohammed, Jumanah Rawashdeh, and Malak Abdullah. *Machine learning with oversampling and undersampling techniques: overview study and experimental results*. IEEE, 2020.

[41] Vladimir Nasteski. "An overview of the supervised machine learning methods". In: *Horizons. b* 4 (2017), pp. 51–62.

[42] Michael Paluszek et al. "What Is Deep Learning?" In: *Practical MATLAB Deep Learning: A Project-Based Approach* (2020), pp. 1–24.

[43] Putting Telecom Fraud Loss into Perspective. `https://cfca.org/putting-telecom-fraud-loss-into-perspective/`. (Accessed: 2024-02-12).

[44] What Is Matplotlib In Python? How to use it for plotting? `https://www.activestate.com/resources/quick-reads/what-is-matplotlib-in-python-how-to-use-it-for-plotting/`. (Accessed: 2024-05-06).

[45] David MW Powers. *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*. 2020.

[46] What is Pandas in Python? https://www.activestate.com/resources/quick-reads/what-is-pandas-in-python-everything-you-need-to-know/. (Accessed: 2024-05-06).

[47] What is Python? Executive Summary. https://www.python.org/doc/essays/blurb/. (Accessed: 2024-05-06).

[48] PyTorch. https://www.nvidia.com/en-us/glossary/pytorch/. (Accessed: 2024-05-31).

[49] D Ramyachitra and Parasuraman Manikandan. "Imbalanced dataset classification and solutions: a review". In: *International Journal of Computing and Business Research (IJCBR)* 5.4 (2014), pp. 1–29.

[50] Wojciech Samek et al. "Explaining deep neural networks and beyond: A review of methods and applications". In: *Proceedings of the IEEE* 109.3 (2021), pp. 247–278.

[51] Arthur L Samuel. "Some studies in machine learning using the game of checkers". In: *IBM Journal of research and development* 3.3 (1959), pp. 210–229.

[52] Matthias Schonlau and Rosie Yuyan Zou. *The random forest algorithm for statistical learning*. 2020.

[53] What is Scikit-learn? https://domino.ai/data-science-dictionary/sklearn. (Accessed: 2024-05-31).

[54] Saurabh Shukla, Arushi Maheshwari, and Prashant Johri. *Comparative analysis of ml algorithms & stream lit web application*. IEEE, 2021.

[55] Kristina P Sinaga and Miin-Shen Yang. *Unsupervised K-means clustering algorithm*. 2020.

[56] Amit Singh, Ranjeet Kumar Ranjan, and Abhishek Tiwari. "Credit card fraud detection under extreme imbalanced data: a comparative study of data-level algorithms". In: *Journal of Experimental & Theoretical Artificial Intelligence* 34.4 (2022), pp. 571–598.

[57] Evgeny I Tarmazakov and Dmitry S Silnov. "Modern approaches to prevent fraud in mobile communications networks". In: *2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*. IEEE. 2018, pp. 379–381.

[58] Diana C Toledo-Pérez et al. "Support vector machine-based EMG signal classification techniques: A review". In: *Applied Sciences* 9.20 (2019), p. 4402.

[59] Kanishka Tyagi et al. *Unsupervised learning*. 2022.

[60] Shivani Tyagi and Sangeeta Mittal. *Sampling approaches for imbalanced data classification problem in machine learning*. Springer, 2020.

[61] Arshdeep Singh Veer and Sushil Bhardwaj. *Enabling streamlined call recordings by integrating NAS with a GSM Gateway through an UC Platform.* IEEE, 2023.

[62] Raden Aurelius Andhika Viadinugroho. *Imbalanced classification in python: SMOTE-Tomek Links method combining SMOTE with Tomek Links for imbalanced classification in python*. 2023.

[63] Rikiya Yamashita et al. "Convolutional neural networks: an overview and application in radiology". In: *Insights into imaging* 9 (2018), pp. 611–629.

[64] Hira Zahid et al. *Big data analytics in telecommunications: literature review and architecture recommendations*. 2019.

8TH MAY GUELMA

Startup Annex

SUPERVISED BY:
DR. AICHA
AGGOUNE

PRESENTED BY:
AHMED RAMI
BOUGUATTOUCHA

2024

INFORMATION CARD

SUPERVISION TEAM

Main Supervisor	Specialty
Dr. Aicha Aggoune	Computer Science

PREOJECT TEAM

Student	Faculty	Specialty
Ahmed Rami Bouguettoucha	Mathematics, Computer Science and Material Sciences	Computer Science

Table of Contents

CONTENTS

1.PROJECT PRESENTATION

PROJECT IDEA

Simbox fraud, also known as GSM gateway fraud, poses a significant threat to telecommunications companies worldwide. It involves fraudsters using SIM boxes filled with multiple SIM cards to bypass international call rates by routing calls as local calls. This not only results in substantial financial losses for telecom operators but also degrades call quality and results in revenue losses for governments due to unpaid taxes.

To combat this pervasive issue, our project proposes the development of an advanced web application for Simbox Fraud Detection powered by Artificial Intelligence. The web application leverages state-of-the-art machine learning algorithms to analyze call data and detect potential simbox fraud.

Our AI-powered Simbox Fraud Detector web application offers a robust and efficient solution to one of the most challenging issues facing the telecommunications industry today. By harnessing the power of advanced analytics and machine learning, we can protect revenue streams, enhance service quality, and ensure regulatory compliance, paving the way for a more secure and reliable telecommunications landscape.

SUGGESTED VALUES

01

Advanced Fraud Detection Model

Utilizes sophisticated machine learning models trained to identify patterns and anomalies in call traffic data indicative of Simbox usage. These models detect unusual call durations, frequencies, and other irregularities that deviate from typical call behaviors.

02

User-Friendly Web Application

- A web-based platform where users can easily upload their call data for analysis. The interface is designed to be intuitive and accessible to users with varying technical expertise.

03

Accurate Fraud Detection

- The advanced machine learning model significantly improves the accuracy of detecting Simbox fraud, reducing financial losses and operational disruptions.

04

Cost-Effective Solution

- By offering a web-based tool, the solution minimizes the need for extensive hardware and software investments, providing a cost-effective means of combating simbox fraud.

SUGGESTED VALUES

05

Downloadable Results

Provides the option to download the analyzed data in a convenient format. Users can download the dataframe for further investigation and record-keeping

06

Accessibility and Ease of Use

The web application format ensures that telecom operators can quickly and easily utilize the fraud detection capabilities without the need for complex integrations or technical expertise.

07

Actionable Insights

The clear, color-coded results allow users to swiftly identify and act upon detected fraudulent activities, ensuring timely mitigation.

WORKING TEAM

Ahmed Rami Bouguettoucha: An intern at Djezzy, specializing in Artificial Intelligence, Machine Learning, and Deep Learning. Completed extensive coursework in these domains and has practical experience through multiple projects, focusing on developing solutions in telecommunications fraud detection.

PROJECT OBJECTIVES

SHORT AND MID-TERM

- Develop and Deploy Simbox Fraud Detection Web Application.
- Validate and Refine Machine Learning Model.
- Establish Partnerships and Collaborations.

LONG TERM

- Expand and Scale the Solution Globally.
- Enhance the application with additional features and capabilities to address evolving fraud tactics and regulatory requirements..
- Become a leader in AI-driven fraud detection solutions for the telecommunications industry.

IMPLEMENTATION SCHEDULE

Phases	1	2	3	4	5	6	7	OnGoing
Planning and Preparation	▨							
Development		▨	▨					
Deployment and initial testing				▨				
Testing and Optimization					▨	▨		
Full Deployment							▨	
Monitoring and Improvement								▨

2.INNOVATIVE ASPECTS

NATURE OF INNOVATIONS

01

AI-Driven Fraud Detection

- Utilizes advanced machine learning algorithms to identify simbox fraud. These models can detect subtle patterns and anomalies in call data that are often missed by traditional rule-based systems.

02

Real-Time Analysis

- Provides real-time data analysis and fraud detection, enabling telecom operators to act swiftly and mitigate potential losses immediately.

03

User-Friendly Web Application

- Offers an intuitive and accessible web interface that allows users to easily upload call data, view analysis results, and download detailed reports. This simplifies the process of fraud detection and makes it accessible to users with varying levels of technical expertise.

04

Scalability and Flexibility

- Designed to scale seamlessly to handle large volumes of call data from multiple telecom operators. The flexible architecture allows for easy integration with existing systems and customization based on specific user needs.

FIELDS OF INNOVATION

01

Machine Learning and Artificial Intelligence

Employs cutting-edge AI and machine learning techniques to detect Simbox fraud. These technologies enable the system to identify complex patterns and behaviors that traditional methods cannot easily detect.

02

Data Processing and Analysis

Innovates in the way call data is processed and analyzed. The system handles large datasets efficiently, providing fast and accurate fraud detection results. Advanced data preprocessing techniques ensure the integrity and quality of the input data.

03

User Experience and Interface Design

Focuses on creating a highly intuitive and user-friendly web application. The design prioritizes ease of use, allowing telecom operators to navigate the system effortlessly, upload data seamlessly, and interpret results quickly.

04

Real-Time Fraud Detection

Introduces real-time data analysis capabilities, allowing telecom operators to detect and respond to fraud as it happens. This proactive approach minimizes the impact of fraud and protects revenue more effectively.

3.STRATEGIC MARKET ANALYSIS

MARKET SEGMENT

TELECOM OPERATORS

The primary market for the Simbox Fraud Detector is Mobile Network Operators (MNOs). These operators face significant revenue losses and degraded call quality due to simbox fraud. By using the Simbox Fraud Detector, MNOs can detect and mitigate fraudulent activities in real-time, protecting their revenue streams and ensuring high service quality.

TELECOM INFRASTRUCTURE PROVIDERS

Another key market segment is Telecom Infrastructure Providers. These companies supply the necessary technology for telecom operators. By integrating the simbox fraud detection solution into their offerings, infrastructure providers can enhance network security and efficiency, offering added value to their telecom clients.

MEASURING COMPETITION INTENSITY

STRENGTHS

- Low Direct Competition: In Algeria, there are currently no providers offering dedicated simbox fraud detection services. This presents a significant opportunity to become the leading solution in the market.
- Market Entry Advantage: Being the first to market with a specialized simbox fraud detection solution allows for the establishment of a strong brand presence and customer loyalty.
- Potential for High Market Share: With no direct competitors, the Simbox Fraud Detector has the potential to quickly capture a large share of the market, especially among telecom operators looking for effective fraud detection solutions.

WEAKNESSES

- Indirect Competition: There may be indirect competition from general fraud detection services that could potentially add Simbox fraud detection features in the future.
- Threat of New Entrants: As the need for simbox fraud detection becomes more apparent, new competitors may enter the market, increasing the intensity of competition.

MARKETING STRATEGIES

DIRECT SALES

Implementing a direct sales approach to engage with key decision-makers within telecom operators. This involves personalized outreach, product demonstrations, and tailored presentations highlighting the benefits and ROI of our AI-powered fraud detection solution.

ONLINE MARKETING

Leveraging digital channels such as social media and targeted online advertising campaigns. These efforts will focus on creating awareness, driving traffic to our website, and generating leads interested in our fraud detection capabilities.

PARTNERSHIPS WITH TELECOM EQUIPMENT PROVIDERS

Forming strategic alliances with telecommunications equipment providers to integrate our fraud detection solution into their product offerings. This partnership will enhance market penetration and credibility, leveraging their established customer base and industry relationships.

4.PRODUCTION PLAN AND ORGANIZATION

PRODUCTION PROCESS

01

Requirements Gathering

- Collaborate with telecom operators and stakeholders to understand their fraud detection needs and operational requirements.

02

System Design and Architecture

- Develop a detailed system architecture, outlining the technical specifications and infrastructure required to support the web application.

03

Software Development

- Utilize agile development methodologies to build and integrate AI-powered machine learning algorithms for simbox fraud detection.

04

Testing and Quality Assurance

- Conduct rigorous testing phases, including unit testing, integration testing, and user acceptance testing, to ensure the reliability and performance of the application.

05

Deployment and Implementation

- Deploy the web application on secure servers, configure for optimal performance, and ensure compatibility with existing telecom infrastructure.

06

Maintenance and Support

- Provide ongoing maintenance, updates, and technical support to address any issues and optimize performance based on user feedback and industry developments.

SUPPLY CHAIN

HARDWARE AND INFRASTRUCTURE

- Servers and Hosting: Secure and reliable servers to host the web application, ensuring high availability and performance.
- Networking Equipment: Robust networking infrastructure to support seamless data transfer and real-time fraud detection capabilities.

SOFTWARE AND TOOLS

- Development Tools: Advanced software development tools and platforms to facilitate the coding, testing, and deployment of the web application.
- Machine Learning Frameworks: State-of-the-art machine learning frameworks and libraries to develop and refine fraud detection algorithms.

DATA SOURCES

- Telecom Data: Access to large volumes of call data from telecom operators to train and validate the machine learning models.
- Public Databases: Utilize public databases and datasets to supplement and enhance the training data for more accurate fraud detection.

SUPPLY CHAIN

THIRD-PARTY SERVICES

- Cloud Services: Utilize cloud computing services for scalable storage, processing power, and backup solutions.
- Security Services: Implement third-party security services to ensure data protection and compliance with industry standards.

HUMAN RESOURCES

- Development Team: Skilled software developers, data scientists, and machine learning experts to build and maintain the application.
- Support Staff: Technical support and customer service teams to assist users and ensure smooth operation of the web application.

LABOR

DEVELOPMENT TEAM

- Software Developers: Build and implement the web application.
- Data Scientists: Develop and refine machine learning models to detect simbox fraud.
- Machine Learning Engineers: Integrate and optimize machine learning models for real-time data processing.

QUALITY ASSURANCE (QA)

- QA Engineers and Testers: Ensure the application is reliable and performs well through rigorous testing.

CUSTOMER SUPPORT

- Technical Support Specialists: Assist users with troubleshooting and effective use of the application.

SALES AND MARKETING

- Sales Representatives: Engage with potential clients and drive sales.
- Marketing Specialists: Promote the Simbox Fraud Detector and generate leads.

KEY PARTNERS

01

Telecom Operators

Collaborate with major Mobile Network Operators (MNOs) to gather data, validate the fraud detection models, and pilot the solution.

02

Telecom Infrastructure Providers

Partner with providers of telecom infrastructure to integrate the fraud detection solution into their offerings, enhancing their value proposition and reaching a broader market.

03

Technology Providers

Partner with cloud and data storage companies for robust infrastructure and secure data handling, and collaborate with security providers to ensure compliance and protect sensitive information.

04

Academic and Research Institutions

Collaborate with universities and research centers specializing in AI and machine learning to stay at the forefront of technological advancements and continuously improve the fraud detection algorithms.

05

Industry Associations

Engage with telecommunications industry associations to promote the solution, gain insights into industry trends, and establish credibility.

COSTS AND CHARGES

DEVELOPMENT COSTS

- Investment in software development for creating and deploying the fraud detection platform.
- Costs for developing algorithms, user interface (UI), backend systems, and database integration.

INFRASTRUCTURE COSTS

- Expenses related to cloud-based hosting services to support the platform.
- Consideration for data storage, scalability, and security measures.

MARKETING AND PROMOTION

- Budget allocated for marketing efforts to attract telecom operators.
- Strategies may include digital marketing, content creation, and participation in industry events.

OPERATIONAL AND ADMINISTRATIVE COSTS

- Ongoing expenses for managing operations, including salaries, office rent, utilities, and general administrative costs.
- Provision for customer support, training, and operational maintenance.

TURNOVER

Year	Total Subs	Price (Month)	Total Revenue (Year)
1	1	400,000 DZD	4,800,000 DZD
2	2	400,000 DZD	9,600,000 DZD
3	3	400,000 DZD	14,400,000 DZD
4	4	400,000 DZD	19,200,000 DZD
5	6	400,000 DZD	28,800,000 DZD

EXPECTED RESULTS

YEAR 1:

- Total Revenue: 4,800,000 DZD

YEAR 2:

- Total Revenue: 9,600,000 DZD

YEAR 3:

- Total Revenue: 14,400,000 DZD

YEAR 4:

- Total Revenue: 19,200,000 DZD

YEAR 5:

- Total Revenue: 28,800,000 DZD

CASH FLOW PLAN

YEAR 1:

- **Inflows:**
  - Revenue: 4,800,000 DZD
- **Outflows:**
  - Development Costs
  - Infrastructure Costs
  - Marketing Costs
  - Operational Costs

YEAR 2:

- **Inflows:**
  - Revenue: 9,600,000 DZD
- **Outflows:**
  - Continued Development and Maintenance Costs
  - Increased Marketing and Sales Efforts
  - Operational and Administrative Costs

YEAR 3:

- **Inflows:**
  - Revenue: 14,400,000 DZD
- **Outflows:**
  - Scaling Infrastructure Costs
  - Enhanced Marketing and Expansion Efforts
  - Operational and Administrative Costs

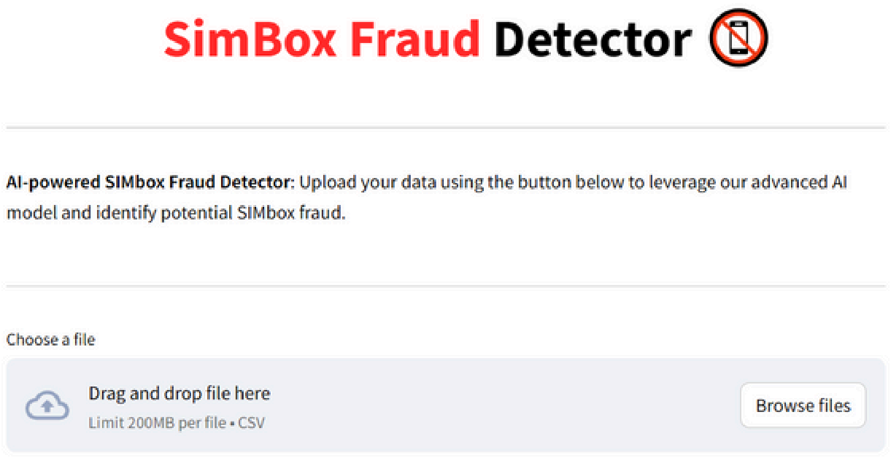
YEAR 4:

- **Inflows:**
  - Revenue: 19,200,000 DZD
- **Outflows:**
  - International Expansion Costs
  - Continued Marketing and Customer Acquisition
  - Operational and Administrative Costs

YEAR 5:

- **Inflows:**
  - Revenue: 28,800,000 DZD
- **Outflows:**
  - Further Scaling and Expansion Costs
  - Sustained Marketing and Customer Retention Efforts
  - Operational and Administrative Costs

6. PROTOTYPE

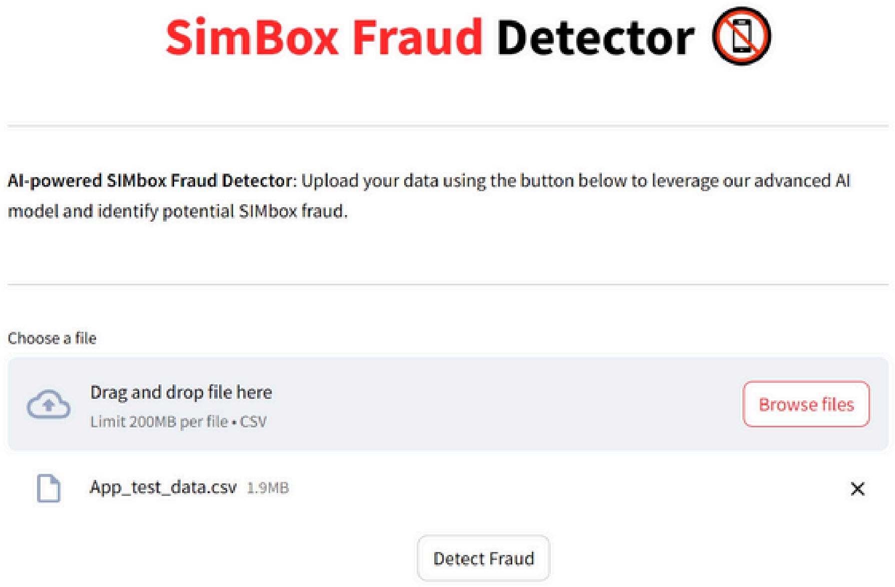


SimBox Fraud Detector

AI-powered SIMbox Fraud Detector: Upload your data using the button below to leverage our advanced AI model and identify potential SIMbox fraud.

Choose a file

Figure.1: HomePage

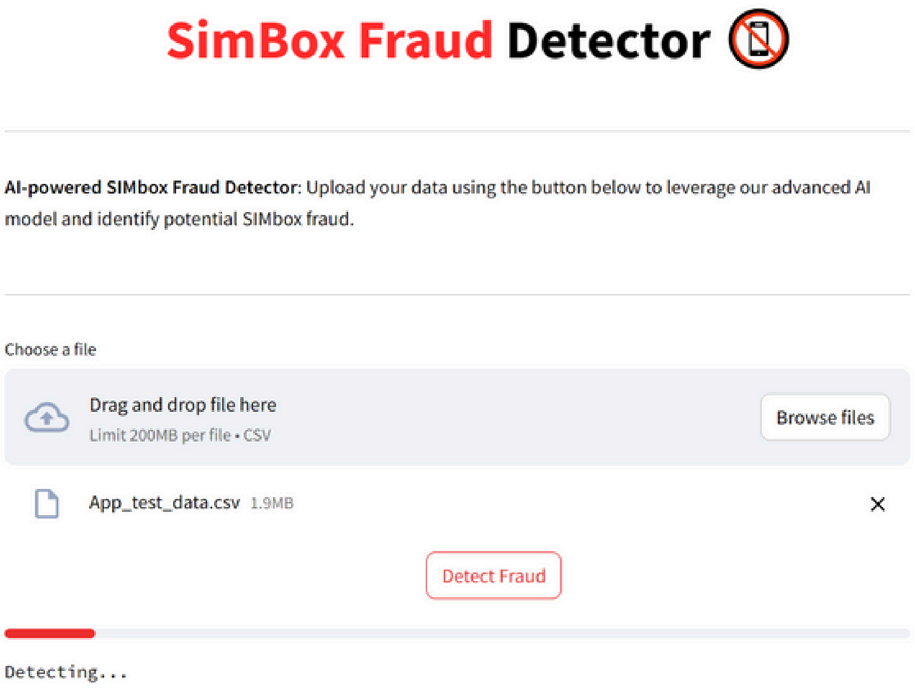


SimBox Fraud Detector

AI-powered SIMbox Fraud Detector: Upload your data using the button below to leverage our advanced AI model and identify potential SIMbox fraud.

Choose a file

Figure.2: Selecting Data



SimBox Fraud Detector 📵

AI-powered SIMbox Fraud Detector: Upload your data using the button below to leverage our advanced AI model and identify potential SIMbox fraud.

Choose a file

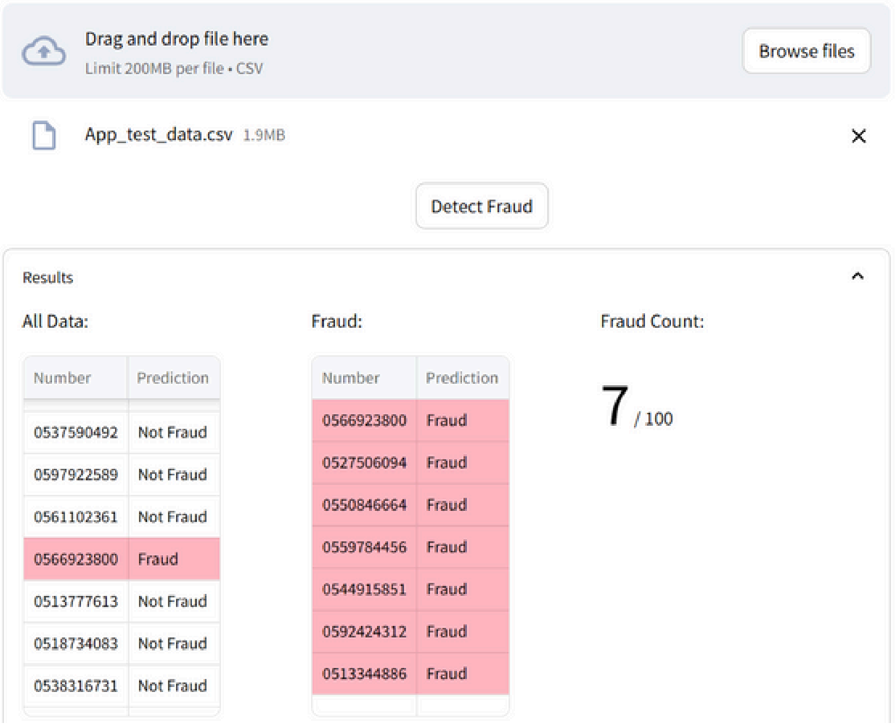
Drag and drop file here

Limit 200MB per file • CSV

App_test_data.csv 1.9MB

```
Detecting...
```

Figure.3: Detecting Fraud



Drag and drop file here

Limit 200MB per file • CSV

App_test_data.csv 1.9MB

Results

All Data:

Number	Prediction
0537590492	Not Fraud
0597922589	Not Fraud
0561102361	Not Fraud
0566923800	Fraud
0513777613	Not Fraud
0518734083	Not Fraud
0538316731	Not Fraud

Fraud:

Number	Prediction
0566923800	Fraud
0527506094	Fraud
0550846664	Fraud
0559784456	Fraud
0544915851	Fraud
0592424312	Fraud
0513344886	Fraud

Fraud Count:

7 / 100

Figure.4: Detection Results