SIM BOX DETECTION USING MACHINE LEARNING TECHNIQUES

Weerakkodi Pathirathnage Tharindu Darshana Dissanayake

(188252k)

Degree of Master of Science

Department of Electronic and Telecommunication Engineering

University of Moratuwa

Sri Lanka

January 2022

SIM BOX DETECTION USING MACHINE LEARNING TECHNIQUES

Weerakkodi Pathirathnage Tharindu Darshana Dissanayake

(188252k)

Dissertation submitted in partial fulfillment of the requirements for the degree Master of Science

Department of Electronic and Telecommunication Engineering

University of Moratuwa

Sri Lanka

January 2022

DECLARATION

Candidate:

I declare that this is my work to the best of my knowledge, and this report does not incorporate any material previously submitted for a degree or diploma in any university or institute of higher learning without acknowledgment. This report does not contain any material previously published or written by another person except where the acknowledgment is made in the text.

I also hereby grant to the University of Moratuwa the non-exclusive right to reproduce and distribute my report, in whole or in part in print, electronic, or any other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

…….…… …….

Dissanayake W. P. T. D Date

Supervisor:

I certify that the declaration above by the candidate is true to the best of my knowledge, and the candidate has carried out research under my supervision.

…….…… ..…….

Dr. Ranga Rodrigo Date

Abstract

Subscriber Identity Module (SIM) box fraud causes severe International Direct Dialing (IDD) revenue losses to network operators and also causes harm to their brand names due to its low Quality of Service (QoS). Therefore, it is very important to prevent this fraud by disconnecting SIM numbers used for SIM box fraud with smaller call attempts to hinder fraudsters from generating revenue. However, detection of SIM box fraud numbers for disconnection consumes lots of work hours as it is necessary to minimize the impact on genuine subscribers. Call Detail Record (CDR) analysis is one of the methods to detect SIM numbers used in the SIM box. Machine learning techniques for fast data analysis are suitable for detecting SIM box numbers using CDR. This research was carried out to accurately identify suitable machine learning methods to detect SIM box fraud numbers. Subscriber profiles are created using CDR to train and evaluate machine learning models. Support vector classifier, decision tree classifier, and K-nearest neighbors classifier algorithms show better results from different types of machine learning models that were trained and evaluated to identify suitable algorithms to detect accurately. A weighted linear combination of these algorithms was used to get the best result with a macro F1-score value of 0.89 using the ensemble voting classifier algorithm. The result confirms that machine learning models can be used to detect SIM box fraud numbers accurately.

Keywords: International direct dialing, SIM box fraud, bypass fraud, machine learning, neural networks, support vector machine algorithms, decision tree algorithms, K-nearest neighbors algorithms, ensemble voting algorithm.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to Dr. Ranga Rodrigo for the guidance, supervision, and advice given to complete this research. Then, I would also like to extend my gratitude to Dr. Tharaka Samarasinghe, Dr. Kasun Hemachandra, and other lecturers who provide guidance and support to complete the M.Sc. in Telecommunication and the research. Next, I would like to thank my supervisors from Dialog Axiata PLC Mr. Dasun Kehelwala, and Mr. Sampath Ilesinghe for providing help and approval to complete the research successfully. Finally, I would like to thank my family members and friends who helped me a lot in completing the research.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CDR	Call Detail Record
CFCA	Communications Fraud Control Association
CLI	Calling Line Identification
IDD	International Direct Dialing
IMEI	International Mobile Equipment Identity
IMSI	International Mobile Subscriber Identity
IP	Internet Protocol
IRSF	International Revenue Share Fraud
MO	Mobile Originated
MT	Mobile Terminated
PBX	Public Branch Exchange
QoS	Quality of Service
SIM	Subscriber Identity Module
SIP	Session Initiation Protocol
SMS	Short Message Service
USD	United States Dollars
VoIP	Voice over Internet Protocol

1. INTRODUCTION

1.1. Background

International Direct Dialing (IDD) is one of the main revenue streams of telecommunication network operators. IDD calls that terminated in a network generate high revenue for the network operator and generate higher tax revenue for the government compared to local calls. In Sri Lanka, the total IDD termination was 393,330,936.2 minutes in 2020. Even though it is significantly lower than 61,464,764,054 minutes of local incoming usage, it is a sizable portion of the revenue due to high IDD termination charges [1]. However, there are massive revenue losses that occur from IDD calls due to different frauds.

Subscriber Identity Module (SIM) box fraud is one of the frauds that directly causes IDD termination revenue losses. According to a survey conducted by the Communications Fraud Control Association (CFCA), the estimated global loss due to SIM box fraud was around 3.28 billion United States Dollars (USD) in 2019, which was the second-largest loss incurred by a fraud type in the telecommunication industry as it was only behind International Revenue Share Fraud (IRSF) which incurred 5.04 billion USD in loss [2]. Figure 1.1 shows the estimated fraud losses of the most prevalent fraud types including SIM box fraud. SIM box fraud is also called bypass fraud as its bypasses the normal IDD call route through illegal routes that terminate the IDD call as a local call with the help of a SIM box device. The SIM box which contains several local on-net or off-net SIMs can route IDD calls via these SIMs as local calls, so the network operator of the terminating network only observes the calls as local calls causing them to miss out on the IDD call charges. Therefore, it is necessary to prevent SIM box fraud to mitigate this IDD revenue loss.

Figure 1.1: Estimated fraud losses by type (in USD billions) in 2019 as per the CFCA survey [2].

Network operators prevent SIM box fraud by detecting SIMs used by SIM boxes to make calls and then blocking these SIM numbers from making any calls to the network. The detection of the SIM box fraud is mainly done via two methods which are active test call, and Call Detail Record (CDR) analyzing methods. In the active test call method, test IDD calls are made from outside the network to several test SIM numbers to check whether the received calls are through legitimate IDD routes or fraudulent local SIM boxes. This method can identify mobile numbers which are used in SIM boxes accurately. However, the volume of the detected fraudulent mobile numbers is directly proportional to the volume of the test calls. Therefore, this method still cannot identify all the active SIM box numbers. Therefore, CDR analyzing method is important to detect a higher volume of fraudulent mobile numbers, if not all the active SIM box numbers. In CDR analyzing method, CDRs are analyzed under different criteria to identify the mobile numbers which are used for SIM box fraud. This is possible as the mobile numbers used in SIM boxes show different behavior compared to the normal subscriber behavior. However, the identification of SIM box numbers via CDR analysis has issues with accuracy compared to the active test call method that detects SIM box numbers with full accuracy. The inaccuracy is there as sometimes genuine subscribers tend to show behavior close to SIM box numbers.

Therefore, it is very important to use correct methods to analyze CDRs and detect SIM box numbers to avoid the detection of genuine subscribers as SIM box numbers (false alarms).

CDRs can be analyzed using the usual data analysis methods. Therefore, machine learning algorithms that are employed in different fields for data analysis can also be applied for the detection of SIM Box fraud.

1.2. Problem Statement

CDR analyzing method is the most suitable method to detect active SIM box fraud numbers. Accurate and faster detection of fraud numbers is essential to avoid customer dissatisfaction due to incorrect number disconnection or revenue loss due to delay in SIM box number disconnection. Therefore, CDR analyzing method that can detect SIM box fraud numbers accurately and efficiently is necessary. Past behavior of SIM box fraud can be used to identify new SIM box fraud numbers. However, with the development in technology, SIM box devices provide many options which can help fraudsters to change SIM box number behavior to hide SIM box numbers from detection in the absence of analysis. Therefore, machine learning algorithms that can provide accurate and fast detection with adaptability to changing parameters are more suitable for SIM box fraud detection via CDR analysis.

There are several research papers available regarding SIM box fraud number detection using machine learning algorithms. These research papers use accuracy to determine the success of the SIM box detection. However, accuracy is not a suitable method to measure machine learning algorithm success for fraud detection as SIM box fraud numbers that are active in a given period are less than 1%. Therefore, more than 99% accuracy can be archived by considering all the numbers as non-fraud numbers. Most of these research papers do not provide a method to arrange CDR data before feeding it to machine learning algorithms. Therefore, it is not possible to recreate research results with a new dataset. These research papers are also a few years old, so they are

not suitable to detect current SIM box numbers that use advanced SIM box devices compared to past SIM box devices.

Therefore, the problem that this research address is how to identify a practical method to detect SIM box fraud numbers from CDR details using machine learning algorithms. In other words, the problem is the identification of a suitable machine learning algorithm that can successfully detect current SIM box fraud numbers, without only considering the accuracy and selection of a practical method to process CDR details via machine learning algorithms.

1.3. Objectives

The objectives of this research can be listed as follows:

- Research machine learning algorithms suitable for SIM box fraud detection.
- Determine useful parameters that can be extracted from CDR details.
- Develop programs to detect SIM box fraud using machine learning.
- Conduct comprehensive performance studies on the selected machine learning method to determine the feasibility of the solution.

1.4. Outline

This research report contains five chapters including Chapter 1 which is the introduction of this research. Chapter 2 describes the literature review related to fraud types observed in telecommunication networks and SIM box fraud detection using machine learning algorithms. Chapter 3 presents the design of the machine learning method to detect SIM box fraud. Chapter 4 includes the experimental setups and results observed from the proposed machine learning method. Chapter 5 summarizes the research, points out limitations of the research, and presents directions to improve the research in the future.

2. LITERATURE REVIEW

2.1. Fraud Types

The use of any telecommunication service without paying service charges is defined as telecommunication fraud [3], [4]. There are different types of fraudulent activities observed in telecommunication networks, which cause losses to both subscribers and network operators. This section will briefly introduce top telecommunication fraud types including SIM box fraud that can be observed in a telecommunication network. These frauds mainly cause losses related to IDD calls.

2.1.1. IRSF

IRSF is the fraud, according to the CFCA survey, incurs the highest losses when considering overall fraud losses to network operators [2]. IRSF occurs when a fraudster makes a high volume of calls to a high-cost destination via a hacked Public Branch Exchange (PBX), stolen SIM cards, or dialer malware like 2Faced [5]. Initially, this fraud occurred when fraudsters make calls to international premium-rate numbers using stolen SIM cards in roaming [6]. This fraud mainly occurs due to the use of low-cost off-the-shelf carriers by the calling party network operators to reduce the cost of routing the calls. Fraudsters can make revenue-sharing agreements with these off-the-shelf carriers to route the calls to their desired destination to false answer the calls, so they can make a profit from these high-cost calls. Figure 2.1 shows an example of how IRSF is performed using hacked PBX, stolen SIM cards, or dialer malware. The red dotted line shows how a fraudster makes revenue when network operators use a low-cost carrier (T1) to route traffic. A carrier (T2) in the route that has a revenue-share agreement routes the calls to premium-rate numbers belonging to the fraudster to generate revenue from the termination fee [7].

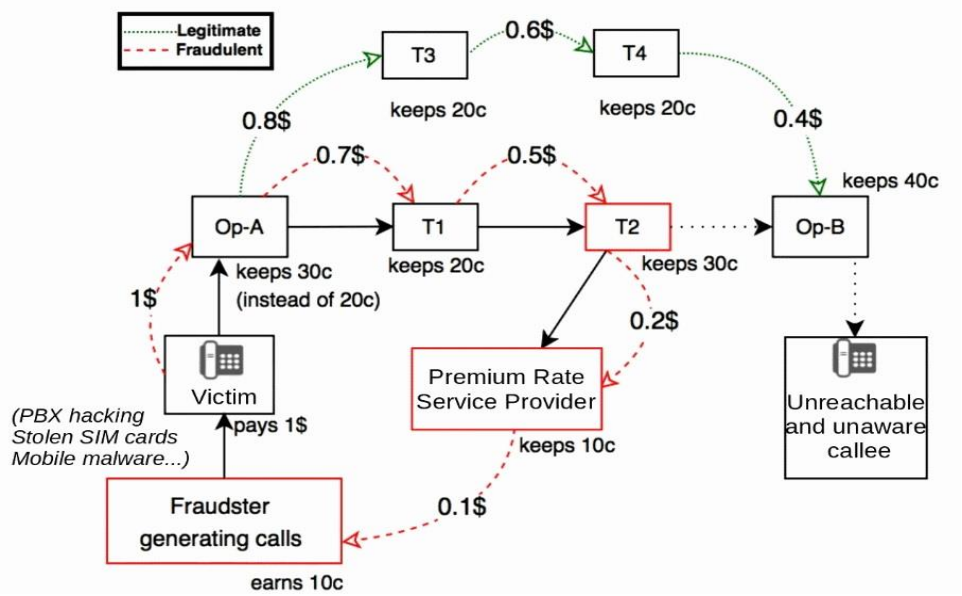

Figure 2.1: Example of IRSF hacked PBX, stolen SIM cards, or dialer malware [7].

2.1.2. Arbitrage

Arbitrage is a fraud that exploits the difference between IDD call charges between countries. In this fraud, calls between two countries are routed via a third country which has lower calling charges compared to direct calls between two counties. This fraud often causes issues to subscribers due to lower call quality. The network operators can mitigate this fraud by using reputable carriers instead of off-the-shelf carriers with lower costs to route calls and monitoring IDD call traffics regularly. According to CFCA 2019 survey, arbitrage fraud caused 2.71 billion USD of revenue loss in 2019 to the telecommunication industry [2].

2.1.3. Premium Rate Service Fraud

The premium rate service is a service provided by the network operators to some companies under the agreement to share revenue generated by the traffic to the assigned numbers which are premium-rate numbers. This service is mainly used by TV shows like reality shows, contests, and entertainment services. The calling party will have to pay a high premium rate when making calls to these premium-rate numbers.

Premium rate service fraud occurs when fraudsters that have premium-rate numbers entice normal subscribers who are mainly in different countries to call these premium-rate numbers by making missed calls or messages to the subscribers. Fraudsters receive a share of the revenue generated by the calls from normal subscribers [7], [8].

2.1.4. SIM Box Fraud

SIM box fraud uses SIM box devices to terminate IDD calls as local calls. Fraudsters act as carriers that route IDD traffic and charge for IDD calls that it receives. However, fraudsters only need to pay local call charges in return to connect these received IDD calls through the SIM box with local SIM numbers. Section 2.2 elaborates on SIM box fraud in more detail.

2.2. SIM Box Fraud in Detail

SIM box fraud can be mainly observed in destinations where IDD call charges are very high compared to local call charges. Fraudsters earn revenue from this difference in IDD call charges and local call charges [9]. They use SIM cards installed in the SIM box device to generate local calls related to the IDD calls that they received from the gateway that they used to act as a proper carrier. Fraudsters can provide lower calling charges compared to other carriers for the IDD calls they receive due to the lower cost they incur in terminating these calls as local calls. Therefore, they can also receive large amounts of IDD traffic due to their lower charges.

Routing traffic through this fraudulent method also has downsides regarding the Quality of Service (QoS) aside from the revenue loss to the destination network operator. The most obvious issue is the Calling Line Identification (CLI) issue observed by the call receiving party. The call receiving party observes a local CLI which is from the SIM number in the SIM box that generates the call. Therefore, the call receiving party cannot identify the call as an IDD call let alone the calling party from the CLI, and they cannot also call back to the actual call generating party. These fraudulent calls most of the time have low voice quality and connectivity as well. One

of the reasons for that is most fraudsters do not use a suitable Internet Protocol (IP) route to connect to other carriers, and telecommunication networks to receive IDD calls. Therefore, packet drops, and delays occur in the routing path causing voice quality issues. Another reason is that the SIM box contains up to hundreds of SIM numbers, and it uses them to simultaneously connect calls, so most of the times base stations in the area of the SIM box will be overloaded causing call drops and voice quality issues. Therefore, SIM box fraud causes dissatisfaction to both call generating and receiving parties regarding their connected networks.

There are two types of SIM box fraud observed according to SIM numbers used in the SIM box. They are on-net bypass and off-net bypass.

2.2.1. On-Net Bypass

On-net bypass occurs when the SIM box fraud uses a SIM of a network that is the same as the call receiving party subscribed network to connect the IDD call. This method is mostly used by fraudsters as calls between the same network numbers have lower call charges compared to out-of-network calling charges. Figure 2.2 shows the On-net bypass calling path.

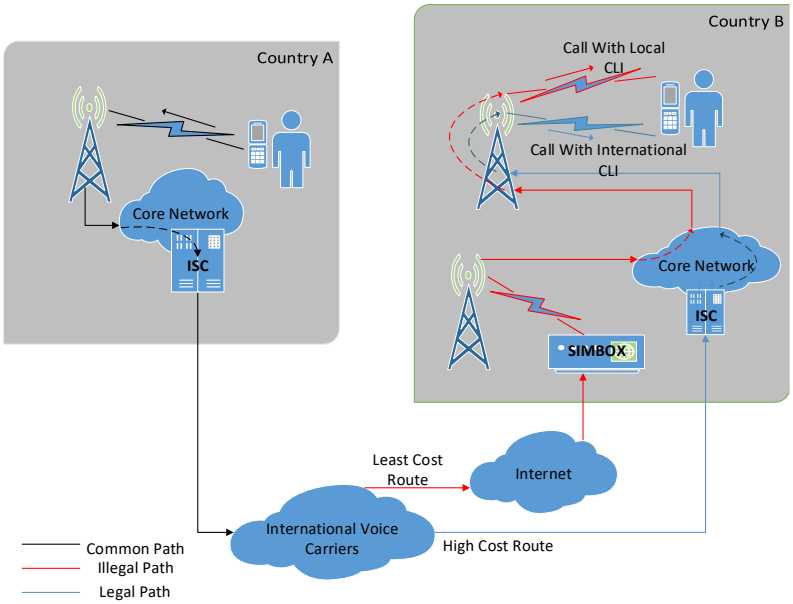

Figure 2.2: On-net bypass [10].

2.2.2. Off-Net Bypass

Off-net bypass occurs when the SIM box fraud uses a local SIM of a network that is different from the call receiving party subscribed network to connect the IDD call. This method is not used commonly as calls between different networks, even in the same country have higher charges even though they are lower than IDD charges. Therefore, this method is mostly used when fraudsters need to avoid being detected, calling fixed network numbers, or in an event of a promotion that may provide lower call charges. Figure 2.3 shows the off-net bypass calling path.

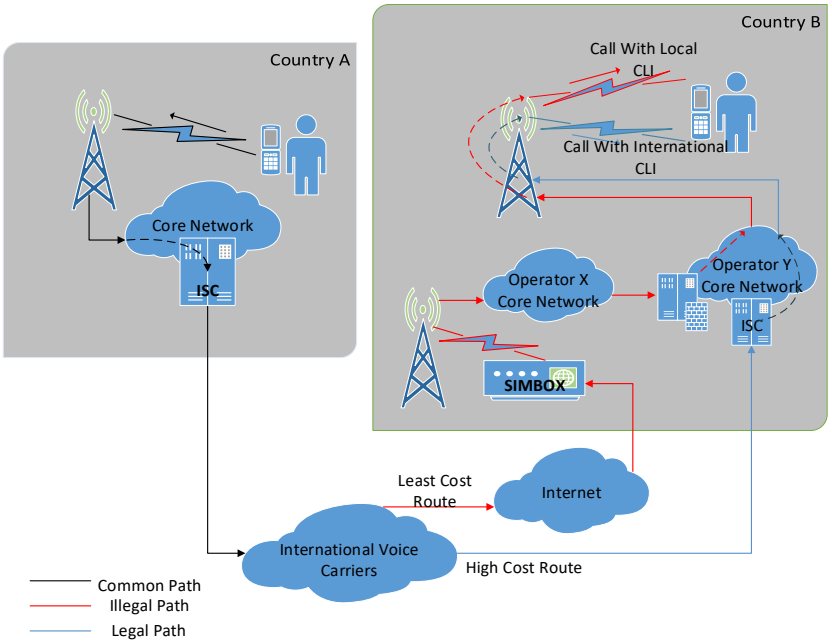

Figure 2.3: Off-net bypass [10].

2.3. SIM Box Device

The SIM box device is normally connected to a Voice over Internet Protocol (VoIP) gateway which fraudsters use to create the Session Initiation Protocol (SIP) trunks with IDD carriers and telecommunication networks, so IDD traffic is received into the SIM box through the gateway. SIM box can contain several SIM numbers of different mobile operators, and they can be used to make calls when the SIM box receives calls from the gateway. Fraudsters use this capability to get IDD calls from other connected

parties to bypass international routes to connect IDD calls as local calls using local SIM numbers in the SIM box.

Modern SIM box devices have some additional advanced technological capabilities like IMEI swapping and intelligent SIM swaps algorithms which greatly help fraudsters. This makes it hard to identify SIM numbers used in the SIM box as they show behaviors close to normal subscribers. IMEI swapping helps to avoid being blocked by network operators using the IMEI of the SIM box while intelligent SIM swaps allow fraudsters to manage the call amount that SIM numbers in the SIM box generate [11]. Fraudsters can choose between a high volume of calls or a lower volume of calls via SIM numbers. The high volume of calls via SIM numbers makes them easy to detect. However, fraudsters can get the maximum value out of SIM numbers before they are detected for SIM box fraud. The low volume of calls via SIM numbers makes them hard to detect, so fraudsters can use SIM numbers longer time. Therefore, these additional capabilities help the fraudsters to use the SIM box devices and SIM numbers used in the device for longer periods.

2.4. SIM Box Fraud Detection

Use of the SIM box for SIM box fraud is illegal by law in most countries. However, government involvement in actively seeking out SIM box fraud is very limited. Detection of the SIM box fraud is mostly done by network operators to prevent revenue losses. Network operators prevent SIM box fraud by detecting SIM numbers used in the SIM box and blocking these identified SIM numbers from making calls to the network. However, fraudsters usually use a new set of SIM numbers when old SIM numbers are blocked similar to Short Message Service (SMS) spam fraud [12]. Therefore, detection of the SIM box fraud is a continuous process. There are mainly two methods used by the network operators to detect SIM numbers used in the SIM box. There are active test calls, and CDR analyzing methods.

2.4.1. Active Test Call Method

The active test call method uses test IDD calls to detect SIM numbers used in the SIM box. Network operators can use their test call generator or can get IDD test calls from test call service providers. These test calls are routed via different IDD routes according to the preference of the network operator and received by the test SIM numbers dedicated to receiving these test calls. Network operators identify SIM numbers used for fraud when a test call is received with local CLI. However, sometimes calls received from the IDD route have local CLI due to CLI manipulation of a carrier [13]. Therefore, most of the time network also confirms whether the call is received from the local path to avoid disconnection of normal or invalid subscribers due to CLI manipulation. The active test call method has several advantages like detection of SIM box numbers with 100% accuracy and detection of SIM box numbers in real-time. However, this method cannot be used to detect a high volume of SIM box numbers. The volume of the detection of SIM box numbers is directly proportional to the volume of the test call. However, as it is not practical to use a very high volume of test calls, the use of the active test call method alone is not sufficient to keep the telecommunication network free of SIM box fraud.

2.4.2. CDR Analyzing Method

The CDR analyzing method analyses CDRs generated by the telecommunication network to detect SIM numbers used in SIM boxes. The SIM box numbers show somewhat different behaviors compared to a normal subscriber. This method is useful in detecting all the SIM box numbers used in the network. However, this method, unlike the active test call method, cannot detect SIM box numbers in real-time, and the detected number cannot label as SIM box numbers with 100% confidence. This is because this method sometimes detects genuine subscribers that show behaviors close to SIM box numbers. Therefore, it is very important to use suitable CDR data analyzing method that can detect SIM box numbers with higher accuracy and a low false alarm rate to prevent the disconnection of genuine subscribers. However, on the flip side, the analyzing method needs to detect the SIM box numbers as quickly as

possible with few attempts from the SIM box number. If the analysis method takes a long time to detect the SIM box number, the fraudsters can cover the cost of the detected SIM number before it is detected. Then they can just change the SIM numbers used in the SIM box periodically to keep the fraud operation going as they can buy SIM numbers very easily at a low cost in most instances. Therefore, machine learning algorithms which provide high accuracy, low false alarm rate, and fast response from a small set of data can be very suitable for detecting SIM box numbers. Section 2.5 elaborates on the use of machine learning techniques to detect SIM box fraud in more detail.

2.5. Machine Learning for SIM Box Fraud Detection

There are two main types of machine learning methods called supervised and unsupervised. The main difference between supervised and unsupervised learning is that supervised learning needs training data that have correctly identified input and output values to learn the function and build the model. However, unsupervised learning does not need any training data with labeled output. It uses inputs without outputs to learn the relationships and structure of the data [14]. There are training data obtained from the active test call method and other CDR analyzing methods for SIM box fraud detection. Therefore, the supervised method is suitable to detect SIM box numbers.

2.5.1. Research on SIM Box Fraud Detection Using Machine Learning

Most of the research approaches regarding the use of machine learning algorithms to detect frauds using CDR details used neural networks or support vector machine supervised learning algorithms. A paper on the use of a Bidirectional Artificial Neural Network (Bi-ANN) machine learning algorithm to detect generic mobile phone fraud using real-time CDR data showed that it is possible to predict fraud behavior from the time series of call duration attribute of subscribers using Bi-ANN with a 90% success rate [15].

Another research [16] discussed three machine learning methods to detect mobile network fraud. The methods were not focused to detect SIM box fraud in the mobile network but instead to detect fraud in mobile networks in general. The first method is a supervised learning neural network method that uses a discriminative function to detect fraud from a summary statistic of mobile subscribers. The second method is an unsupervised learning probability density estimation method that uses the past behavior of mobile subscribers to generate a probability density function to calculate the current behavior of the mobile subscribers. The third method is the Bayesian network method which uses statistics of different mobile subscriber behaviors including fraud behaviors to identify the probability that a mobile subscriber is a fraud given the behavior of the mobile subscriber. According to this research, these three methods could detect over 85% of the fraud without false detection from the test dataset.

Research on SIM box fraud detection using artificial neural networks show that analyzing nine attributes like 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, and average minutes using artificial neural network algorithm could detect SIM box numbers with 98.71% accuracy [17]. These nine attributes were obtained from CDRs. Same researchers with added support from Zain [18] present research on SIM box fraud detection using both support vector machine and neural networks. In this research same nine attributes were used. However, different CDRs were used, and the results show that the support vector machine algorithm could detect SIM box numbers with 99.06% accuracy while the artificial neural network algorithm could only provide 98.69% accuracy. This research also showed that the neural network algorithm takes more than three times the time to analyze the CDRs compared to the support vector machine algorithm confirming that the support vector machine is better for SIM box fraud detection compared to the neural network algorithm. However, another research on machine learning techniques for SIM box fraud detection analyzed twenty-five attributes that could be obtained from CDR details, using both support vector machine and neural networks under various evaluation parameters showed different results. The best and worst

performance of the support vector machine algorithm under different evaluation parameters were accuracy of 99.24% and accuracy of 95.18% respectively. These values are lower compared to the best and worst performance of the neural network under different evaluation parameters. The neural network algorithm showed the best and worst performance of 99.87% accuracy and 99.53% accuracy respectively [19].

Some other research approaches [20] use machine learning algorithms like alternating decision trees, functional trees, and random forests to detect SIM box fraud using CDR. The research used CDR details from mobile network operators in United State for analysis after transforming them into IMEI profiles with forty-eight features. According to the analysis, the following characteristics were identified to differentiate SIM boxes from genuine users and probing devices:

- The stationary nature of SIM boxes as their connection to the same base stations for most of the Mobile Originated (MO) or Mobile Terminated (MT) calls.
- SIM boxes generate a very high number of MO calls with significantly longer call durations to mostly different called party numbers while generating very low or zero MT calls with shorter call durations.
- Some SIM boxes generate a high number of international MO calls.
- SIM boxes use a smaller number of International Mobile Subscriber Identity (IMSI) during the longer period due to the disconnection of SIM box numbers by network operators.

These characteristics were used in detecting SIM box numbers using machine learning techniques. The machine learning technique presented in this research used a linear combination of three classifiers which were alternating decision tree, functional tree, and random forest to detect SIM box fraud. The linear combination of these classifiers was done with weight coefficients which were calculated to detect SIM box fraud with minimum prediction error for the training dataset which contained five hundred SIM box IMEI and about ninety-three thousand legitimate IMEI numbers. According to the research, the random forest method showed the lowest false positive and the highest

false negative while the functional tree method showed the lowest false negative rates and the highest false positive rates with the lowest accuracy of 99.90%. An optimal method combining the three classifiers showed the highest accuracy of 99.95%. However, IMEI-wise SIM box number detection that was presented in the research paper is not possible to use in the current mobile network because of the fraudulent IMEI used in some models of mobile phones and IMEI masking facilities are available in modern SIM box devices.

Therefore, it shows that different machine learning algorithms like alternating decision trees, functional trees, random forests, neural networks, and support vectors can be used to analyze CDR to detect SIM box numbers as they have around 99% accuracy under the best conditions. However, under different conditions and different training data different machine learning algorithms showed different accuracy levels. Therefore, it is better to choose an algorithm that is better suited to the telecommunication network and available data. Also, accuracy was used to determine the impact of the solution present in the research papers. However, SIM box fraud detection generally works with unbalanced distribution as an exceedingly small percentage of the numbers available are SIM box numbers compared to all the subscribers. Therefore, the F1-score value is a better measurement to identify the impact of the machine learning solution in detecting SIM box numbers.

3. METHODOLOGY

3.1. CDR Processing

The machine learning solution to detect SIM box fraud numbers use CDR data as input. However, raw CDR data cannot be optimally analyzed via a machine learning algorithm to identify SIM box fraud numbers. Therefore, it is necessary to create a method to process CDR before feeding it to a machine learning algorithm. SIM numbers which can be categorized as SIM box fraud or genuine numbers can be used as an index to create profiles related to each subscriber using CDR details.

3.1.1. CDR

CDR is a record generated by specific telecommunication devices to indicate attached user equipment usage. CDR is generated not only for calls generated in fixed and mobile networks but also for SMS and mobile data. It mainly contains attributes given in Table 3.1 for call CDR. However, the format of the CDR can change with telecommunication equipment, the service provided, and the equipment vendor as some equipment also have the facility to configure CDR format with needed attributes. CDR is mainly used for billing purposes. However, it is also used for troubleshooting issues, identifying resource usage, and fraud detection via analysis. CDR is also useful for detecting SIM box fraud numbers.

SIM box frauds use SIM numbers to generate IDD calls as local calls, so the SIM numbers used for the fraud show abnormal behavior compared to a normal user. The abnormalities such as generation of calls to many different numbers, showing previously identified fraud behavior, call generation from previously identified fraud locations, low mobility, and call generation from user equipment that was previously identified for fraud. Subscriber profiles that use SIM numbers as the indexes can be created using the CDR to better analyze the SIM number behavior to find abnormalities.

Table 3.1: Attributes available in a CDR.

Attribute	Description
Origination Date and Time	Date and time when the call reached the system
Answer Date and Time	Answered date and time when the call was answered
Disconnect Date and Time	Disconnected date and time when the call disconnected
Release Date and Time	Date and time when the call was released by the system
Calling Party ID	Subscriber Identity Number of the user who originated the call
Called Party ID	Subscriber Identity Number of the user intended to receive the call
Disconnected Party	The party that disconnected the call
Call Duration	Answer duration of the call
Release Cause	Reason for the call release in code form
Type of Call	Type of calls like national, international
Calling Party Cell ID	Cell number that the calling party was attached to
Called Party Cell ID	Cell number that the called party was attached to
IMEI	IMEI number of the user equipment

3.1.2. Creation of Subscriber Profiles Using CDR

SIM box fraud numbers have different behaviors compared to genuine subscribers, so these identifiable behaviors needed to be included accurately when creating subscriber profiles to allow a machine learning algorithm to identify SIM box fraud numbers. Local and international CDR as well as known SIM box number CDR can be used to get a good number of parameters to train machine learning algorithms. Table 3.2 contains subscriber profiles created using local, international, and past SIM box detected number CDR for the research.

Table 3.2: Subscriber profile created using CDR and other details.

Parameter Name	Description
Subscriber number	The SIM number which was used as the identity (index) field
Total calls	Total outgoing calls for a given period by the subscriber
Total distinct numbers called	Total outgoing calls to different numbers
Total calls to different operators	Total outgoing calls to other operator numbers
Total distinct calls to different operators	Total outgoing calls to different other operator numbers
Total calls to IDD receiving	Total outgoing calls to different numbers that receive IDD calls
Total calls to IDD generating	Total outgoing calls to different numbers that generate IDD calls
IDD received	The subscriber receives IDD calls (YES/NO)
IDD generated	Subscriber generates IDD calls (YES/NO)
Cell count	Total different cells used to make calls for a given period by the subscriber
Total duration	Total outgoing call duration
Grey cell	Subscriber used previously SIM box call generated cell (YES/NO)
Grey behavior	Similarities compared to past SIM box detected numbers
Subscriber number use days	Number of days the Subscriber number is used to make calls
Total received call duration	Total incoming call duration
Maximum received call duration	Maximum incoming call duration
Grey IMEI	Subscriber used previously SIM box call generated IMEI (YES/NO)
IMEI count	Number of different IMEI used by the subscriber

Total calls, total distinct numbers called, total distinct numbers called, total calls to different operators, cell count, total duration, and IMEI count parameters obtained from local CDRs. However, these parameters should be periodically calculated with fresh CDRs to detect SIM box numbers instantly. In this research, the sliding window method was used to generate above mentioned parameters from local CDRs. Figure 3.1 shows the sliding window methods used for the research. Three hours and twenty-four hours CDR time range sliding windows were used considering CDR processing times for different CDR time ranges. With the device specification used for the research, CDR processing with thirty minutes CDR range took around eight to twelve minutes while CDR range of three hours only took around ten to fifteen minutes, so

three hours CDR range was used for the research as it provided a high number of call count to analysis with a reasonable CDR processing time. Three hours CDR range was used to detect SIM box fraud numbers fast to prevent fraudsters that generate a high volume of fraud calls during a short period. Due to insufficient device spec, CDR ranges higher than twenty-four hours could not be processed. Therefore, twenty-four hours CDR range was used to detect SIM box fraud that generates a low volume of fraud calls.

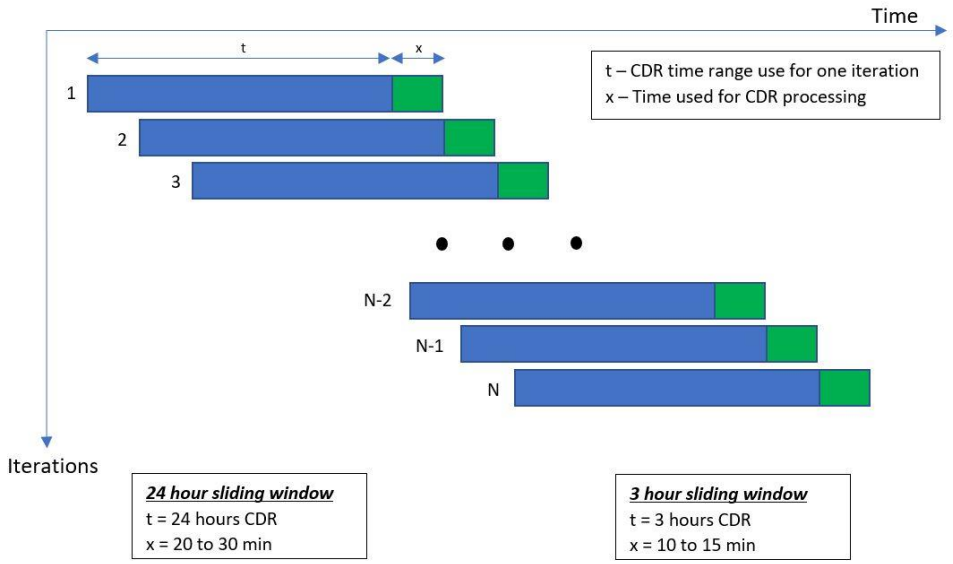

Figure 3.1: Three hours and twenty-four hours period sliding windows.

Total calls to IDD receiving, total calls to IDD generating, IDD received, and IDD generated parameters could be obtained from IDD CDRs. While grey cell, grey behavior, and grey IMEI parameters could be obtained from known SIM box number CDRs. These parameters do not change rapidly with time and the total details that are needed to contain are small. Therefore, the parameters could store for a longer period like three to six months. Then these stored data could be used to update the subscriber profiles. However, to update these parameters except for IDD received and IDD generated in the profile of a subscriber number, the stored data was needed to match with the subscriber, so the same local CDR range chosen from sliding windows was used to match the stored data with subscriber and update the parameters in the profile.

Subscriber number use days, total received call duration, and maximum received call duration details were stored from the local CDRs since the start of the subscriber activity and updated profile accordingly.

Therefore, local CDR, IDD CDR, old SIM box number CDR and subscriber details stored from local CDR were used to generate two subscriber profile datasets. One subscriber profile contained the past three hours of local CDR data while another subscriber profile contained the past twenty-four hours of local CDR data that could analyze using machine learning algorithms. However, parameters used to create profiles contained different value ranges, so the parameters were normalized between 0 to 1 range, and YES or NO values were marked as 1 or 0 respectively to improve the machine learning analysis.

3.2. SIM Box Detection Using Machine Learning

SIM box fraud detection method was created with machine learning methods using subscriber profiles that were already identified as SIM box fraud numbers or genuine subscriber numbers. Therefore, supervised machine learning methods were used that could identify SIM box fraud from subscriber profiles. However, a mobile network has millions of subscribers with only around hundreds of the subscribers are identified as SIM box fraud numbers. Therefore, subscriber profiles used to analyze via machine learning methods were an imbalanced dataset. Therefore, machine learning programs were created to detect SIM box frauds using an imbalanced dataset.

3.2.1. Machine Learning Model to Evaluate Imbalance Dataset

To properly analyze an imbalanced dataset, oversampling was used to boost the SIM box fraud number count in the dataset. The dataset was first separated into training and testing datasets. Then SIM box number count and genuine numbers were balanced in the training dataset using oversampling. The testing dataset was not balanced to test the machine learning models in the natural state to select the most suitable model to detect SIM box fraud numbers. Due to this imbalance in the testing dataset, the

accuracy-based evaluation method of the machine learning model was not possible as high accuracy could obtain by predicting all the subscribers as genuine subscribers. Therefore, evaluation metrics such as confusion matrix, precision, recall, or F1-score were used to evaluate machine learning models.

Machine learning programs were created using the Python programming language. A simple Python machine learning program is shown in Appendix A. In the program, the Python imblearn library oversampling module is used for the oversampling training datasets and the sklearn library is used for machine learning algorithms and to evaluate the models. Two separate programs were created to generate models using imbalanced datasets containing CDR details from three-hour and twenty-four hours periods. According to the evaluation of the training dataset, the support vector classifier showed higher accuracy. However, the decision tree classifier showed better performance in detecting SIM box fraud numbers whereas the best overall performance was observed via the K-nearest neighbors classifier. Figure 3.2 shows the evaluation result of the support vector classifier and decision tree classifier models. Section 4.1.1 elaborates on the performance of different machine learning models created to detect SIM box fraud numbers.

	Support Vector Classification Model	Decision Tree Classifier Model
3 hour sliding window	Accuracy: 99.69% Macro F1-score: 0.89 Confusion Matrix:	Accuracy: 99.61% Macro F1-score: 0.88 Confusion Matrix:
24 hour sliding window	Accuracy: 98.48% Macro F1-score: 0.87 Confusion Matrix:	Accuracy: 98.86% Macro F1-score: 0.91 Confusion Matrix:

		Prediction	
		Genuine (0)	SIM Box (1)
Reference	0	17394	34
	1	20	99

		Prediction	
		Genuine (0)	SIM Box (1)
Reference	0	17365	63
	1	5	114

		Prediction	
		Genuine (0)	SIM Box (1)
Reference	0	10097	125
	1	34	239

		Prediction	
		Genuine (0)	SIM Box (1)
Reference	0	10107	115
	1	4	269

Figure 3.2: Performance of support vector classifier and decision tree classifier models for imbalanced dataset.

According to the confusion matrix, the machine learning models were selecting a high number of SIM box numbers as genuine subscribers. This shows that models are better at detecting genuine subscribers while inaccurate when detecting SIM box numbers. This is mainly due to the imbalance of the dataset. Even with the oversampling, machine learning models were struggling to differentiate between SIM box numbers and genuine subscribers like communication centers, and telemarketing companies that show subscriber profiles close resemble SIM box number profiles. Therefore, machine learning programs were developed to detect SIM box numbers from subscriber profiles that show characteristics close to SIM box number characteristics by filtering out genuine subscribers that showed obvious characteristics different from SIM box characteristics. The filtering was done automatically using a simple Python program. This filtering also made the dataset balanced compared to the previous instance.

3.2.2. Machine Learning Model to Evaluate Balance Dataset

SIM box fraud numbers show distinct characteristics that could separate them from most of the genuine subscribers. The below list shows characteristics that SIM box fraud numbers have that could be used to filter out most genuine subscribers before analyzing subscriber profiles via machine learning models.

- Most of the calls are generated to different numbers (total distinct numbers called count is close to total call count)
- Sometimes most of the calls are generated to other operator numbers (total calls to different operators is close to total call count)
- The cell count is very low
- Most of the calls are generated to numbers that generate or receive IDD calls
- Do not generate or receive IDD calls
- Generate calls from grey cell location
- Generate calls using grey IMEI
- Very low total received call duration

- Very low maximum received call duration

Datasets were generated after filtering out most of the genuine numbers. These generated datasets were balanced, so the machine learning programs could be created without balancing the training datasets. Therefore, accuracy-based evaluation was also possible to use for the evaluation of machine learning models. Two separate programs were also created to generate models using the datasets containing CDR details from three hours and twenty-four-hour time periods. According to the evaluation of the training dataset, the decision tree classifier showed both higher accuracy and better performance in detecting SIM box fraud numbers. Figure 3.3 shows the evaluation result of the support vector classifier and decision tree classifier models. Section 4.1.2 elaborates on the performance of different machine learning models created to detect SIM box fraud numbers.

<table>
<tr><th></th><th>Support Vector Classification Model</th><th>Decision Tree Classifier Model</th></tr>
<tr><td>3 hour sliding window</td><td>Accuracy: 80.54%
Macro F1-score: 0.80
Confusion Matrix:
<table><tr><td></td><td></td><td>Prediction</td><td></td></tr><tr><td></td><td></td><td>Genuine (0)</td><td>SIM Box (1)</td></tr><tr><td>Reference</td><td>0</td><td>56</td><td>12</td></tr><tr><td></td><td>1</td><td>24</td><td>93</td></tr></table></td><td>Accuracy: 87.56%
Macro F1-score: 0.87
Confusion Matrix:
<table><tr><td></td><td></td><td>Prediction</td><td></td></tr><tr><td></td><td></td><td>Genuine (0)</td><td>SIM Box (1)</td></tr><tr><td>Reference</td><td>0</td><td>55</td><td>13</td></tr><tr><td></td><td>1</td><td>10</td><td>107</td></tr></table></td></tr>
<tr><td>24 hour sliding window</td><td>Accuracy: 82.28%
Macro F1-score: 0.82
Confusion Matrix:
<table><tr><td></td><td></td><td>Prediction</td><td></td></tr><tr><td></td><td></td><td>Genuine (0)</td><td>SIM Box (1)</td></tr><tr><td>Reference</td><td>0</td><td>231</td><td>24</td></tr><tr><td></td><td>1</td><td>63</td><td>173</td></tr></table></td><td>Accuracy: 87.98%
Macro F1-score: 0.88
Confusion Matrix:
<table><tr><td></td><td></td><td>Prediction</td><td></td></tr><tr><td></td><td></td><td>Genuine (0)</td><td>SIM Box (1)</td></tr><tr><td>Reference</td><td>0</td><td>231</td><td>24</td></tr><tr><td></td><td>1</td><td>35</td><td>201</td></tr></table></td></tr>
</table>

Figure 3.3: Performance of support vector classifier and decision tree classifier models for the balanced dataset.

Even though the decision tree classifier showed better performance than other machine learning algorithms, voting-based algorithms were used to get inputs from different machine learning algorithms including the decision tree classifier to create a better machine learning model to detect SIM box fraud numbers with minimum prediction error. Python sklearn library ensemble voting classifier module was used to get a linear combination of different machine learning algorithm inputs with weight coefficients

to create a final machine learning model with better performance. In the research, support vector classifier, decision tree classifier, and K-nearest neighbors classifier algorithms were used with weight coefficients to create a better performing ensemble voting classifier model. Figure 3.4 shows the evaluation result of the ensemble voting classifier model.

	Ensemble Voting Classifier Model
3 hour sliding window	Accuracy: 88.10% Macro F1-score: 0.87 Confusion Matrix:

		Prediction	
		Genuine (0)	SIM Box (1)
Reference	0	54	14
	1	8	109

24 hour sliding window	Accuracy: 88.59% Macro F1-score: 0.89 Confusion Matrix:

		Prediction	
		Genuine (0)	SIM Box (1)
Reference	0	233	22
	1	34	202

Figure 3.4: Performance of ensemble voting classifier for balance dataset.

4. EXPERIMENTS AND RESULTS

4.1. Evaluation of Machine Learning Algorithms

Several machine learning algorithms were used to create SIM box fraud detection programs. However, the performance of some of the machine learning algorithms was not suitable for the detection of the SIM box fraud numbers. Therefore, different machine learning algorithms were evaluated with different parameters to determine the most suitable algorithm and the parameters to create a SIM box fraud detection program.

Support vector classifier, decision tree classifier, and K-nearest neighbors classifier algorithms were mainly used in this research to create machine learning models. These machine learning algorithms with optimal parameters showed good results in detecting SIM box fraud from both datasets categories which were unfiltered subscriber profiles (imbalance dataset) and filtered suspicious subscriber profiles (balance dataset after removing some of the genuine subscribers). However, machine learning programs were separately evaluated according to the dataset as measures like accuracy were less effective when detecting optimal machine learning solutions with an imbalanced dataset. The evaluation was carried out for machine learning algorithms with optimal parameters for each dataset. In the dataset, SIM box fraud numbers were labeled as 1 and genuine subscriber numbers were labeled as 0. Therefore, they were represented as 1 and 0 in the performance evaluation as well.

4.1.1. Evaluation of Machine Learning Algorithms Using Imbalance Dataset

For imbalance datasets, performance evaluation was carried out for models created using support vector classifier, decision tree classifier, and K-nearest neighbors classifier algorithms. All the models showed high accuracy values due to the imbalanced nature of the datasets. The algorithms showed more than 99% accuracy when detecting genuine subscribers. However, they showed around 70% accuracy when detecting SIM box fraud numbers. As the number of genuine subscribers in

datasets was also more than 99%, overall accuracy was also around 99%. Therefore, to evaluate the performance of the machine learning models, confusion matrix and macro F1-score values were used.

4.1.1.1. Evaluation of Decision Tree Classifier Model Using Imbalance Dataset

The decision tree classifier was the machine learning algorithm that detected the most SIM box fraud numbers. The algorithm had a lower macro F1-score due to inaccuracies in genuine subscriber detection compared to other algorithms. Therefore, the decision tree classifier model is not suitable for SIM box fraud number detection using the imbalance dataset if the benchmark is to have less impact on genuine subscribers due to SIM box fraud number detection. Table 4.1 and Figure 4.1 show evaluation results of decision tree classifier models for both three hours and twenty-four hours CDR time ranges.

Table 4.1: Performance of decision tree classifier model for imbalance dataset.

	3-hour sliding window				24-hour sliding window			
	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
Genuine (0)	1.00	1.00	1.00	17428	1.00	0.99	0.99	10222
SIM Box (1)	0.64	0.96	0.77	119	0.70	0.99	0.82	273
Macro avg.	0.82	0.98	0.88	17547	0.85	0.99	0.91	10495
Weighted avg.	1.00	1.00	1.00	17547	0.99	0.99	0.99	10495

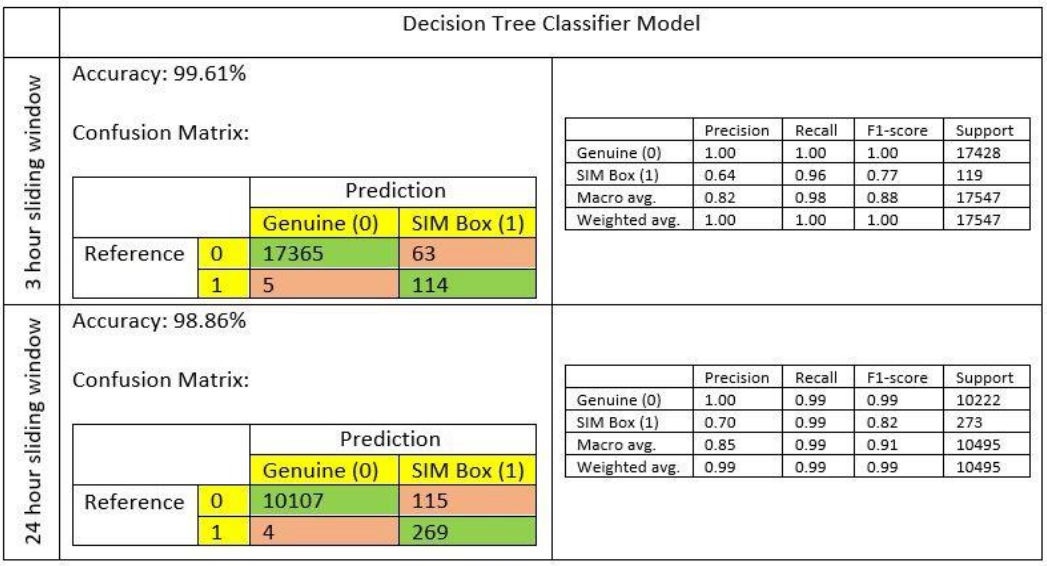

Figure 4.1: Performance of ensemble voting classifier for balance dataset.

4.1.1.2. Evaluation of K-Nearest Neighbors Classifier Model Using Imbalance Dataset

K-nearest neighbors classifier was the machine learning algorithm that shows the highest macro F1-score for the imbalance dataset in this research. The algorithm showed better results in detecting SIM box fraud numbers and genuine subscribers than other algorithms. Therefore, based on the results that we have observed, K-nearest neighbors classifier model is the most effective model for SIM box fraud number detection using the imbalanced dataset. Table 4.2 and Figure 4.2 show evaluation results of K-nearest neighbors classifier models for both three hours and twenty-four hours CDR time ranges.

Table 4.2: Performance of K-nearest neighbors classifier model for imbalance dataset.

	3-hour sliding window				24-hour sliding window			
	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
Genuine (0)	1.00	1.00	1.00	17428	1.00	0.99	0.99	10222
SIM Box (1)	0.68	0.91	0.78	119	0.76	0.94	0.84	273
Macro avg.	0.84	0.95	0.89	17547	0.88	0.96	0.92	10495
Weighted avg.	1.00	1.00	1.00	17547	0.99	0.99	0.99	10495

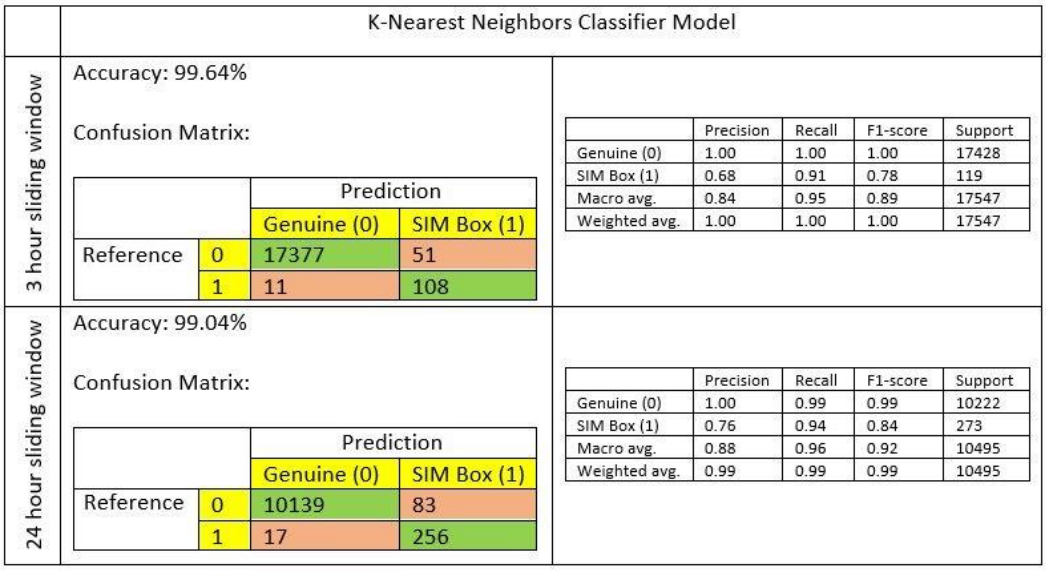

Figure 4.2: Performance of K-nearest neighbors classifier model for imbalance dataset.

4.1.1.3. Evaluation of Support Vector Classifier Model Using Imbalance Dataset

Support vector classifier was the machine learning algorithm that detects genuine subscribers more accurately. However, it was poor at detecting SIM box fraud numbers. The algorithm had better evaluation results for three hours CDR time range for the imbalance dataset compared to other algorithms. Table 4.3 and Figure 4.3 show evaluation results of support vector classifier models for both three hours and twenty-four hours CDR time ranges.

Table 4.3: Performance of support vector classifier model for imbalance dataset.

	3-hour sliding window				24-hour sliding window			
	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
Genuine (0)	1.00	1.00	1.00	17428	1.00	0.99	0.99	10222
SIM Box (1)	0.74	0.83	0.79	119	0.66	0.88	0.75	273
Macro avg.	0.87	0.91	0.89	17547	0.83	0.93	0.87	10495
Weighted avg.	1.00	1.00	1.00	17547	0.99	0.98	0.99	10495

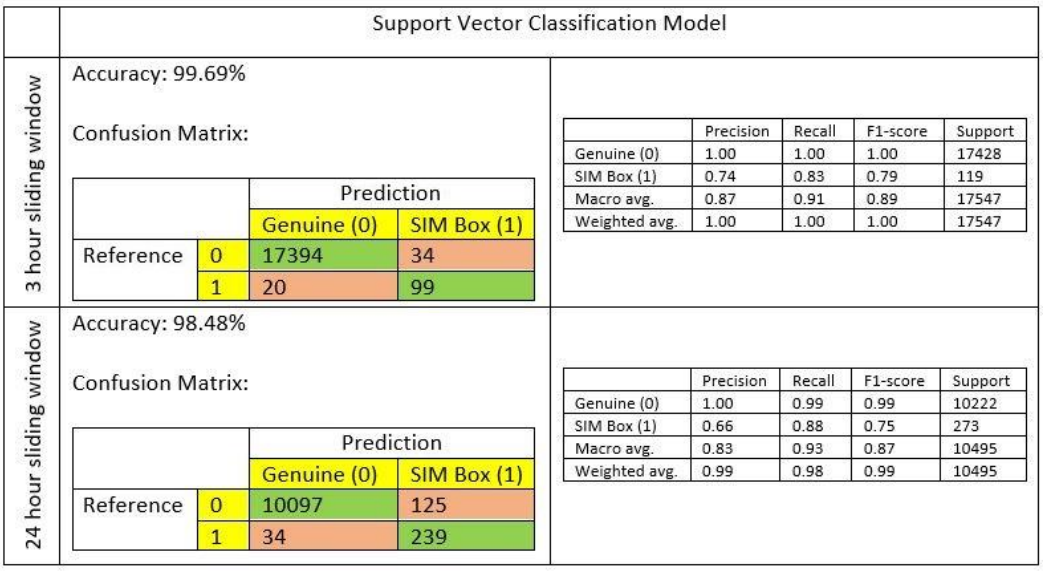

Figure 4.3: Performance of support vector classifier model for imbalance dataset.

4.1.2. Evaluation of Machine Learning Algorithms Using Balance Dataset

For balance datasets, performance evaluation was carried out for models created using support vector classifier, decision tree classifier, K-nearest neighbors classifier, and ensemble voting classifier algorithms. All the models showed lower accuracy values compared to models created from imbalance datasets. This is mainly due to the incorrect assumption of equal misclassification error cost for imbalance dataset model accuracy and significantly low data points in balance datasets compared to imbalance datasets. Even though accuracy was sufficient to evaluate the models for balance datasets, measurements like confusion matrix and macro F1-score values were still used for consistency.

4.1.2.1. Evaluation of Decision Tree Classifier Model Using Balance Dataset

The decision tree classifier was the machine learning algorithm that showed the highest accuracy and macro F1-score for balance datasets in this research without considering the ensemble voting classifier algorithm. It also used the decision tree classifier algorithm as an input to make the prediction. Therefore, using balanced datasets, the decision tree classifier model is the most efficient model for SIM box fraud number detection. Table 4.4 and Figure 4.4 show evaluation results of decision tree classifier models for both three hours and twenty-four hours CDR time ranges.

Table 4.4: Performance of decision tree classifier model for balance dataset.

	3-hour sliding window				24-hour sliding window			
	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
Genuine (0)	0.85	0.81	0.83	68	0.87	0.91	0.89	255
SIM Box (1)	0.89	0.91	0.90	117	0.89	0.85	0.87	236
Macro avg.	0.87	0.86	0.87	185	0.88	0.88	0.88	491
Weighted avg.	0.87	0.88	0.88	185	0.88	0.88	0.88	491

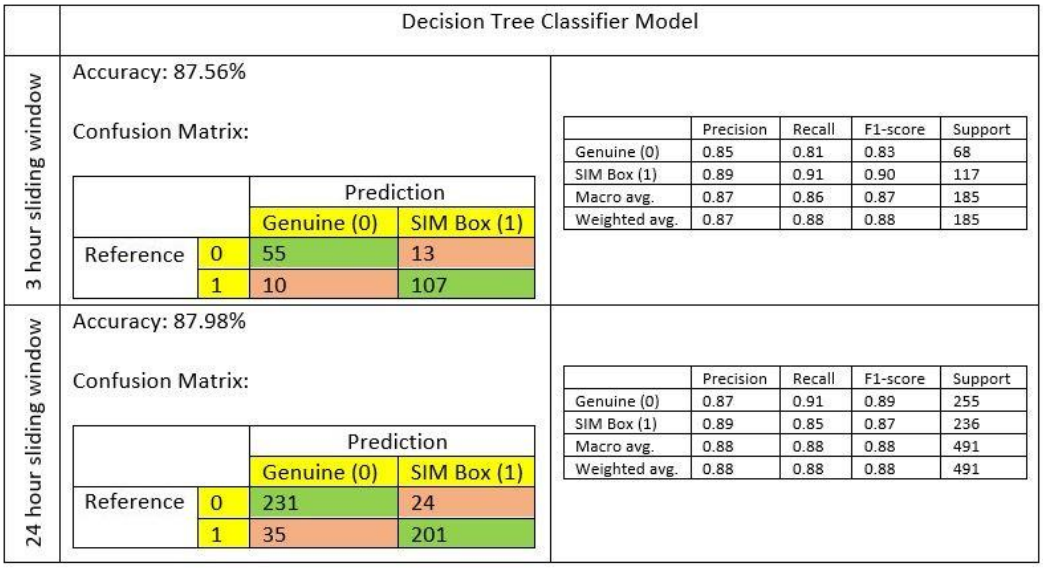

Figure 4.4: Performance of decision tree classifier model for balance dataset.

4.1.2.2. Evaluation of K-Nearest Neighbors Classifier Model Using Balance Dataset

K-nearest neighbors classifier was the machine learning algorithm that showed the lowest accuracy and macro F1-score for balance datasets in this research unlike for imbalance datasets where the algorithm showed the highest macro F1-score. This was mainly due to a reduction in data points in balance datasets. K-nearest neighbors classifier uses the nearest training data points as a reference to determine the state of validating data points. Therefore, the decrease in data points in balanced datasets due to the removal of obvious genuine subscriber numbers caused a reduction in accuracy and macro F1-score. Therefore, increasing available data points in datasets can increase the accuracy and macro F1-score in the K-nearest neighbors classifier model. Table 4.5 and Figure 4.5 show evaluation results of K-nearest neighbors classifier models for both three hours and twenty-four hours CDR time ranges.

Table 4.5: Performance of K-nearest neighbors classifier model for balance dataset.

	3-hour sliding window				24-hour sliding window			
	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
Genuine (0)	0.75	0.60	0.67	68	0.82	0.91	0.86	255
SIM Box (1)	0.79	0.88	0.83	117	0.88	0.78	0.83	236
Macro avg.	0.77	0.74	0.75	185	0.85	0.84	0.84	491
Weighted avg.	0.78	0.78	0.77	185	0.85	0.85	0.84	491

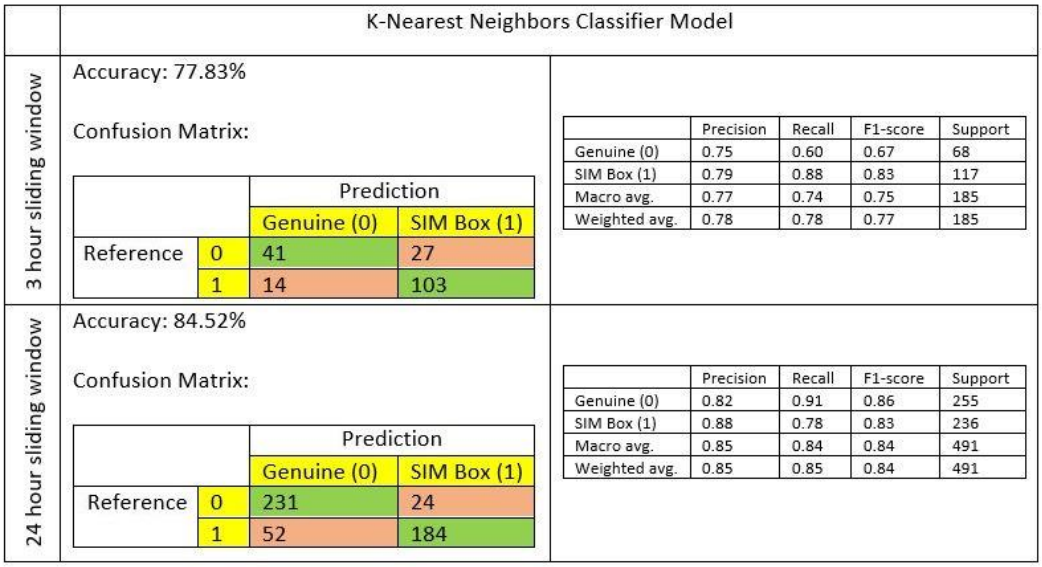

Figure 4.5: Performance of K-nearest neighbors classifier model for balance dataset.

4.1.2.3. Evaluation of Support Vector Classifier Model Using Balance Dataset

Support vector classifier model showed good accuracy and macro F1-score value compared to K-nearest neighbors classifier. However, it could not match up to the decision tree classifier for balanced datasets. Table 4.6 and Figure 4.6 show evaluation results of support vector classifier models for both three hours and twenty-four hours CDR time ranges.

Table 4.6: Performance of support vector classification model for balance dataset.

	3-hour sliding window				24-hour sliding window			
	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
Genuine (0)	0.70	0.82	0.76	68	0.79	0.91	0.84	255
SIM Box (1)	0.89	0.89	0.84	117	0.88	0.73	0.80	236
Macro avg.	0.79	0.81	0.80	185	0.83	0.82	0.82	491
Weighted avg.	0.82	0.81	0.81	185	0.83	0.82	0.82	491

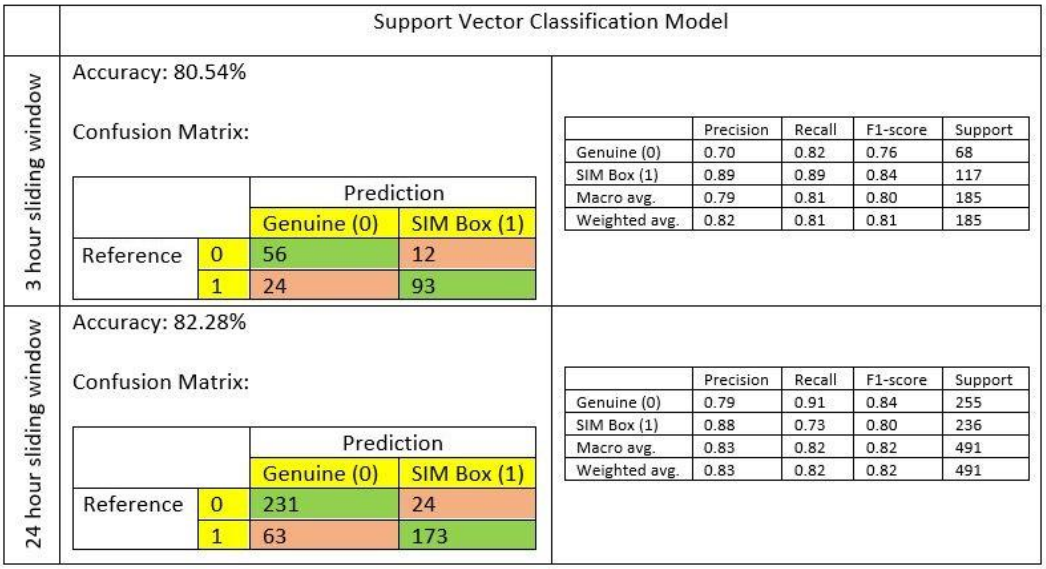

Figure 4.6: Performance of support vector classification model for balance dataset.

4.1.2.4. Evaluation of Ensemble Voting Classifier Model Using Balance Dataset

Ensemble voting classifier algorithm used a linear combination of decision tree classifier, K-nearest neighbors classifier, and support vector classifier inputs to generate a better machine learning model to detect SIM box fraud numbers. Ensemble voting classifier allows different voting parameters for each algorithm, so the algorithm inputs carry weight coefficients. Voting parameters are allocated according to machine learning algorithm individual performance for the same dataset so in this research, the highest weight coefficient was allocated to the decision tree classifier input followed by the support vector classifier input, and finally, the lowest weight coefficient was allocated to K-nearest neighbors classifier. Table 4.7 and Figure 4.7

show evaluation results of ensemble voting classifier models for both three hours and twenty-four hours CDR time ranges.

Table 4.7: Performance of ensemble voting classifier model for balance dataset.

	3-hour sliding window				24-hour sliding window			
	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
Genuine (0)	0.87	0.79	0.83	68	0.87	0.91	0.89	255
SIM Box (1)	0.89	0.93	0.91	117	0.90	0.86	0.88	236
Macro avg.	0.88	0.86	0.87	185	0.89	0.88	0.89	491
Weighted avg.	0.88	0.88	0.88	185	0.89	0.89	0.89	491

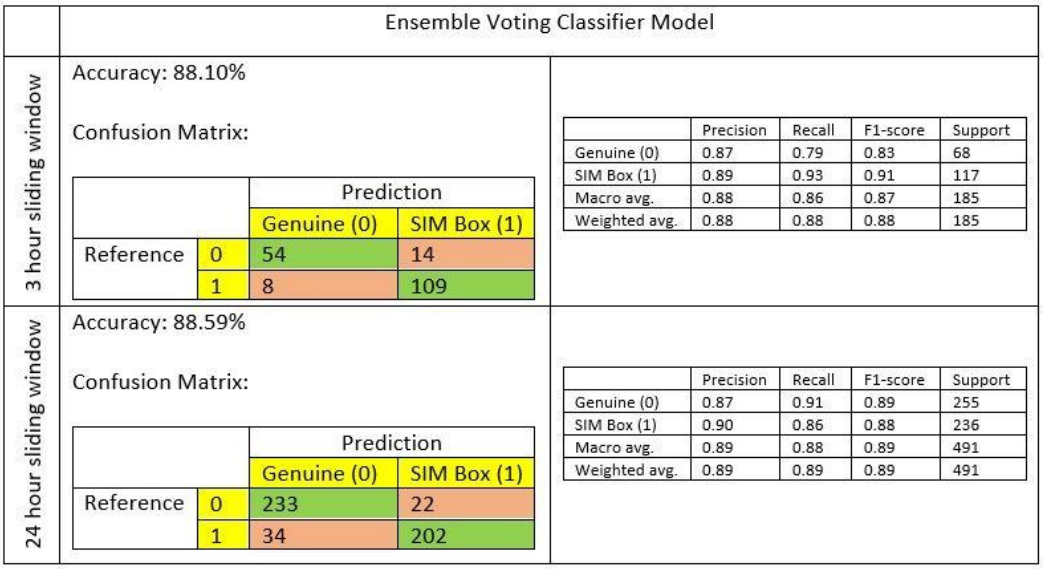

Figure 4.7: Performance of ensemble voting classifier model for balance dataset.

4.2. Parameter Importance Score for Different Subscriber Profile Datasets

Parameter importance score can be logged using the decision tree classifier algorithm. It shows the weight coefficient of each parameter in a dataset. The total of the parameter weight coefficients is equal to one. In this research, the parameter importance score showed the influence of each subscriber parameter in detecting SIM box fraud numbers accurately. Parameter importance score was used to create four graphs for each dataset which are imbalance three hours CDR time range, imbalance twenty-four hours CDR time range, balance three hours CDR time range, and balance twenty-four hours CDR time range datasets.

4.2.1. Parameter Importance Score for Imbalance Dataset

Parameter importance scores for both imbalance datasets show similar characteristics. Both graphs give around 95% weight to the maximum received call duration. The main reason for this is that SIM box numbers normally do not receive calls while genuine subscribers usually receive a high number of calls. However, the reason for the high weight coefficient for maximum received call duration instead of the total received call duration is due to the intentional answering of calls via SIM numbers used for SIM box frauds to avoid detection due to zero total received call duration. Therefore, maximum received call duration is used to identify SIM box fraud numbers that have very small, or zero maximum received call duration. Table 4.8 contains parameter importance scores for imbalance three hours CDR and twenty-four hours CDR datasets. Figure 4.8 shows the parameter importance score graph for the imbalance three hours CDR time range dataset and Figure 4.9 shows the parameter importance score graph for the imbalance twenty-four hours CDR time range dataset.

Table 4.8: Parameter importance scores for imbalance three hours CDR and twenty-four hours CDR datasets.

Parameter Name	Parameter Importance Score	
	Three Hours Range	Twenty-four Hours Range
Total calls	0.00000	0.00023
Total distinct calls	0.02480	0.00131
Calls to different operators	0.00011	0.00000
Distinct calls to different operators	0.00093	0.00117
Calls to IDD receiving	0.00024	0.00725
Calls to IDD generating	0.00006	0.00055
IDD received	0.00000	0.00000
IDD generated	0.00000	0.00000
Cell count	0.00000	0.00046
Total duration	0.00003	0.00019
Grey cell	0.00000	0.00365
Subscriber number use days	0.00096	0.00184
Grey behavior	0.01696	0.01807
Total received call duration	0.00017	0.00129
Maximum received call duration	0.95553	0.96399
Grey IMEI	0.00000	0.00000
IMEI count	0.00021	0.00000

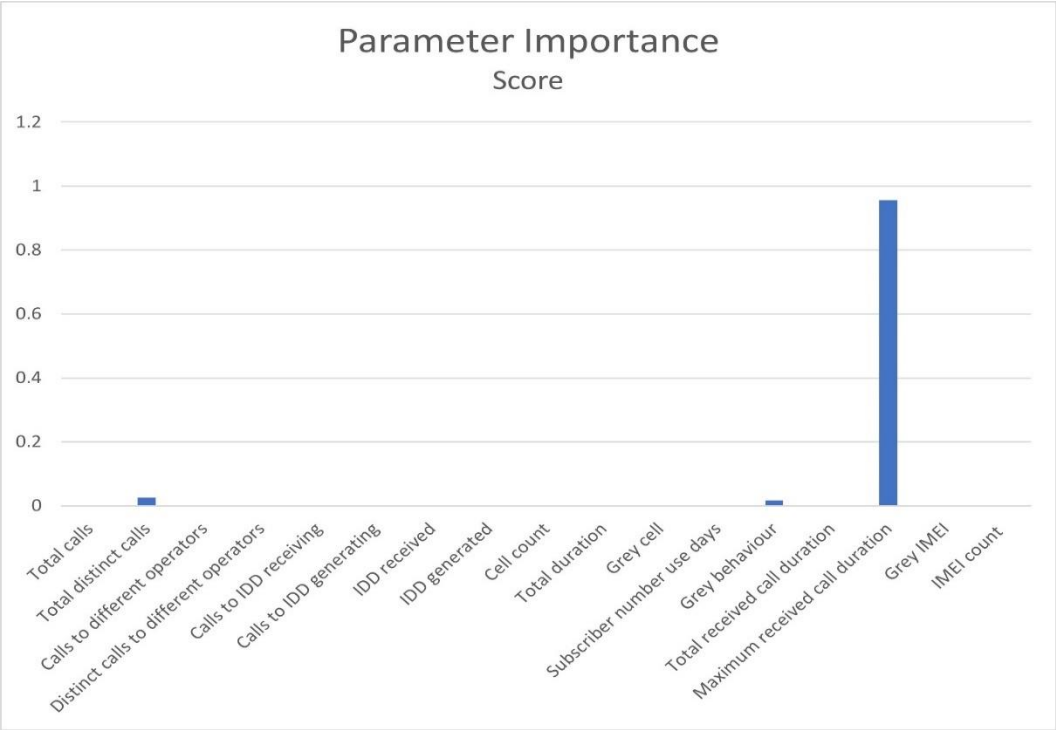

Figure 4.8: Parameter importance score graph for imbalance 3-hour sliding window dataset.

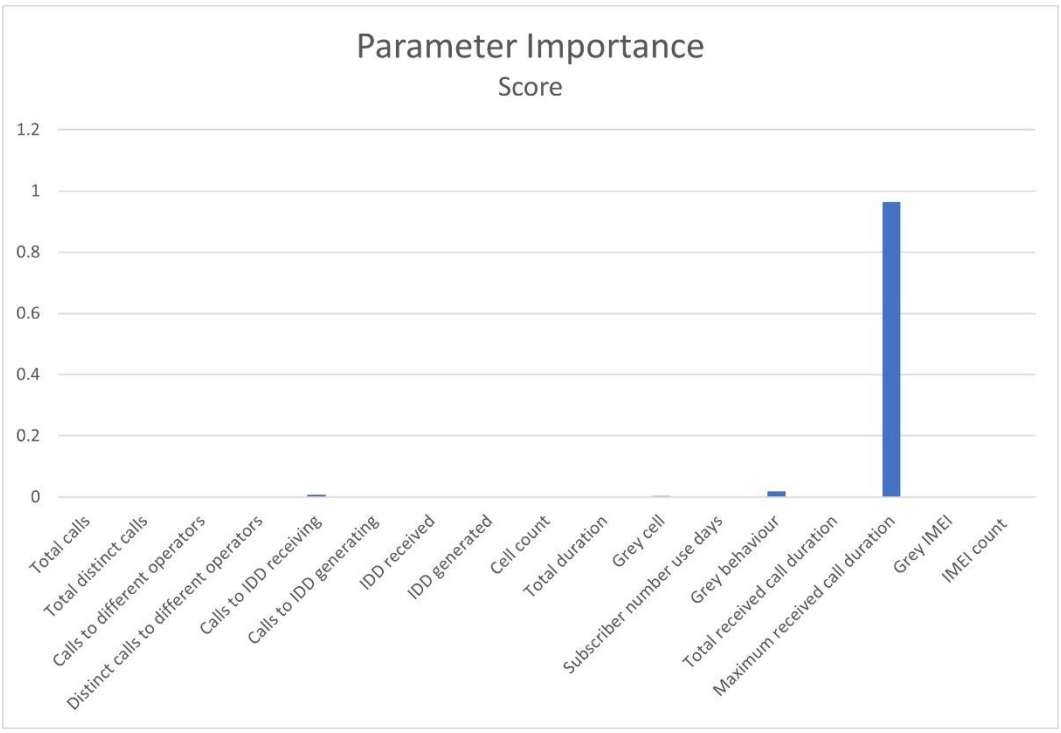

Figure 4.9: Parameter importance score graph for imbalance 24-hour sliding window dataset.

4.2.2. Parameter Importance Score for Balance Dataset

Parameter importance scores for balance datasets are different from each other unlikely for imbalance datasets. Balance datasets contain SIM numbers with highly suspicious characteristics for SIM box fraud. Therefore, received call duration is not a major factor in determining whether a SIM number is used for SIM box fraud as most of the suspicious numbers have very low received call duration.

Subscriber number use days is the parameter with the highest weight coefficient for the balance three hours CDR time range dataset while grey behavior is the parameter with the highest weight coefficient for the balance twenty-four hours CDR time range dataset. The main reason for this difference is due to the lower count of local CDR in three hours CDR range compared to the twenty-four hours CDR range. Unlike in the case of grey behavior which uses the subscriber local CDR range to match past SIM box behavior with the subscriber behavior, subscriber number use days are not dependent on the local CDR range as it only shows the age of the subscriber number. Therefore, the decision tree classifier algorithm chooses subscriber number use days parameter as the parameter with the highest weight coefficient for balance three hours CDR time range dataset while grey behavior parameter as parameter with highest weight coefficient for balance twenty-four hours CDR time range dataset.

Both subscriber number use days and grey behavior parameters are very important in detecting SIM box fraud. SIM box fraud numbers are mainly contained new SIM numbers as SIM box fraud numbers are detected frequently and disconnected by the network operators. Therefore, SIM box fraud numbers are mostly new and have lower subscriber number use days. As for grey behavior, SIM box frauds mainly route the same IDD traffic via their SIM box device, and they replace disconnected SIM numbers with new SIM numbers, so the new SIM numbers route the same IDD traffic as disconnected SIM numbers. Therefore, the behaviors of new SIM numbers are similar to disconnected SIM box numbers. Even though three hours CDR time range dataset has CDR for three hours maximum for each subscriber making grey behavior hard to observe, the grey behavior parameter is still the parameter with the second most

weight coefficient after subscriber number use days. Table 4.9 contains parameter importance scores for balance three hours CDR and twenty-four hours CDR datasets. Figure 4.10 shows the parameter importance score graph for the balance three hours CDR time range dataset and Figure 4.11 shows the parameter importance score graph for the balance twenty-four hours CDR time range dataset.

Table 4.9: Parameter importance scores for balance three hours CDR and twenty-four hours CDR datasets.

Parameter Name	**Parameter Importance Score**	
	Three Hours Range	**Twenty-four Hours Range**
Total calls	0.02195	0.00199
Total distinct calls	0.05188	0.07517
Calls to different operators	0.02693	0.00264
Distinct calls to different operators	0.08789	0.04699
Calls to IDD receiving	0.02084	0.08942
Calls to IDD generating	0.01272	0.01247
IDD received	0.01120	0.00000
IDD generated	0.00602	0.00000
Cell count	0.06669	0.02455
Total duration	0.03239	0.02485
Grey cell	0.09690	0.21328
Subscriber number use days	0.29856	0.03086
Grey behavior	0.15973	0.34930
Total received call duration	0.06449	0.02511
Maximum received call duration	0.00000	0.07177
Grey IMEI	0.03468	0.02368
IMEI count	0.00714	0.00789

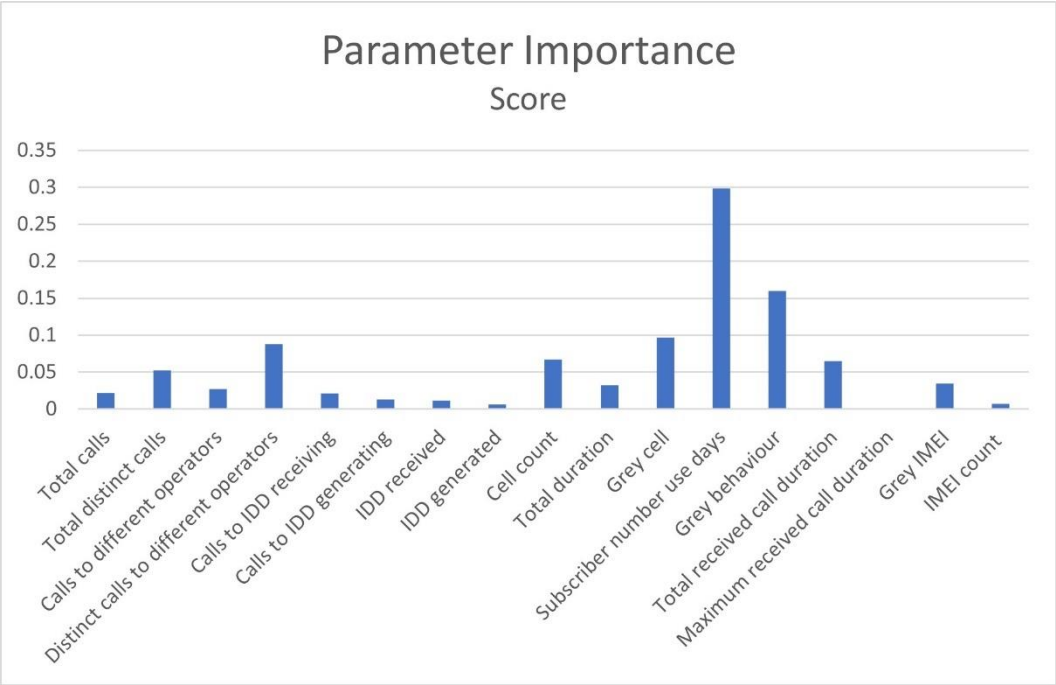

Figure 4.10: Parameter importance score graph for balance 3-hour sliding window dataset.

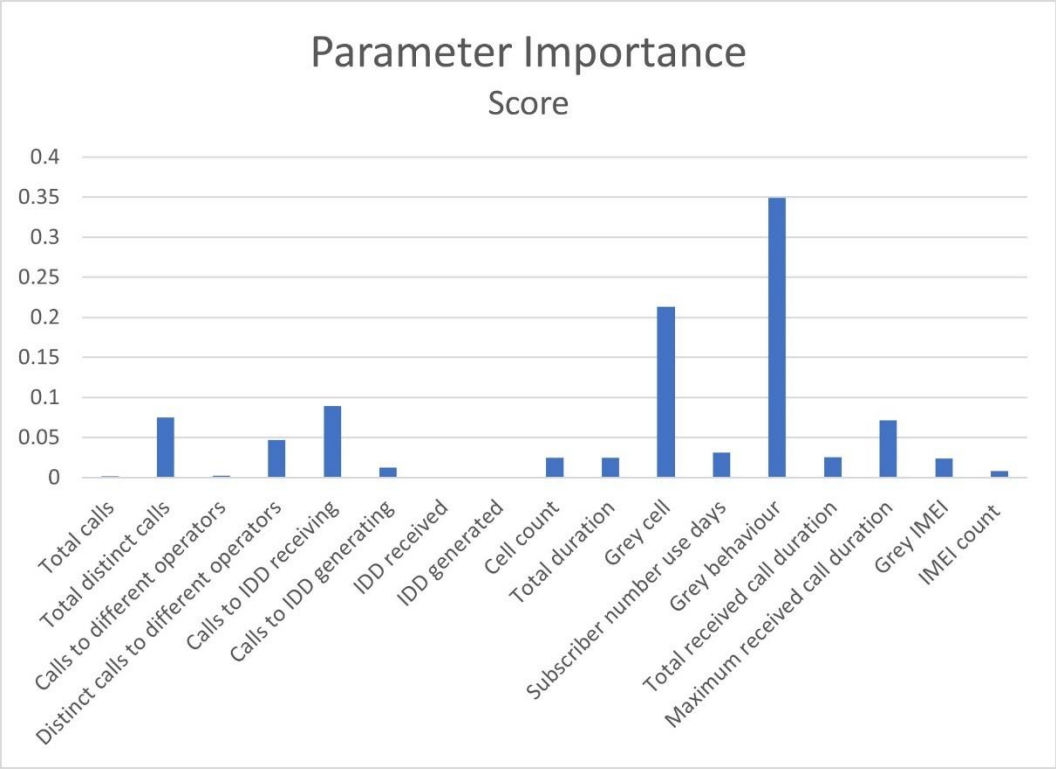

Figure 4.11: Parameter importance score graph for balance 24-hour sliding window dataset.

4.2.3. Comparing Evaluation Results with Other Published Research Findings

There are several research on SIM box fraud detection using machine learning algorithms. However, unlike the proposed solution in this research which only used a maximum of 24 hours of CDR data, most of the published research used several months of data to create machine learning models. The CDR data also contain mostly an equal number of normal subscribers and SIM box numbers unlike in normal mobile networks at a given time. Therefore, the detection of SIM box fraud using these models takes a longer period causing losses to the relevant network operators. The solution presented in this research detects SIM box numbers fast with less CDR data. Therefore, a direct comparison between the result observed in this research with other published research results is not possible.

To compare this research result with the other published research results, machine learning models of these other research were recreated using the balance dataset used in this research. There are several research that use artificial neural network algorithms to detect SIM box fraud numbers [17], [18], [19]. The research titled "Machine learning techniques for sim box fraud detection" [19] proposed the best solution. Therefore, the artificial neural network machine learning model recreated using the balance dataset was compared with the ensemble voting classifier model which is the best solution proposed for the balance dataset in this research.

Figure 4.12 shows the performance comparison between the ensemble voting classifier and artificial neural network models for the balance dataset. According to the evaluation result models created using the ensemble voting classifier show better performance compared to artificial neural network models. Ensemble voting classifier models have accuracies of 88.10% and 88.59% with macro F1 score values of 0.87 and 0.89 for the 3-hour CDR dataset and 24-hour CDR dataset respectively. However, artificial neural network models only have accuracies of 82.70% and 81.67% with macro F1 score values of 0.81 and 0.81 for the 3-hour CDR dataset and 24-hour CDR dataset respectively. Therefore, the proposed model in this research is better when detecting SIM box fraud numbers for the balance dataset.

<table>
<tr><th></th><th>Ensemble Voting Classifier Model</th><th>Artificial Neural Network Model</th></tr>
<tr><td>3 hour sliding window</td><td>Accuracy: 88.10%
Macro F1-score: 0.87
Confusion Matrix:<table><tr><td colspan="2" rowspan="2"></td><td colspan="2">Prediction</td></tr><tr><td>Genuine (0)</td><td>SIM Box (1)</td></tr><tr><td rowspan="2">Reference</td><td>0</td><td>54</td><td>14</td></tr><tr><td>1</td><td>8</td><td>109</td></tr></table></td><td>Accuracy: 82.70%
Macro F1-score: 0.81
Confusion Matrix:<table><tr><td colspan="2" rowspan="2"></td><td colspan="2">Prediction</td></tr><tr><td>Genuine (0)</td><td>SIM Box (1)</td></tr><tr><td rowspan="2">Reference</td><td>0</td><td>47</td><td>21</td></tr><tr><td>1</td><td>11</td><td>106</td></tr></table></td></tr>
<tr><td>24 hour sliding window</td><td>Accuracy: 88.59%
Macro F1-score: 0.89
Confusion Matrix:<table><tr><td colspan="2" rowspan="2"></td><td colspan="2">Prediction</td></tr><tr><td>Genuine (0)</td><td>SIM Box (1)</td></tr><tr><td rowspan="2">Reference</td><td>0</td><td>233</td><td>22</td></tr><tr><td>1</td><td>34</td><td>202</td></tr></table></td><td>Accuracy: 81.67%
Macro F1-score: 0.81
Confusion Matrix:<table><tr><td colspan="2" rowspan="2"></td><td colspan="2">Prediction</td></tr><tr><td>Genuine (0)</td><td>SIM Box (1)</td></tr><tr><td rowspan="2">Reference</td><td>0</td><td>238</td><td>17</td></tr><tr><td>1</td><td>73</td><td>163</td></tr></table></td></tr>
</table>

Figure 4.12: Performance comparison between ensemble voting classifier and artificial neural network models for the balance dataset.

5. CONCLUSIONS

5.1. Summary

There are different types of frauds observed in operator networks. One of the main frauds that cause IDD call revenue losses to network operators is SIM box fraud which bypasses the IDD calling path to generate international calls as local calls using local SIM numbers. The objective of this research is to create a suitable machine learning program to detect SIM box fraud numbers fast and accurately using CDR.

To process CDR data via machine learning algorithms, the data is needed to be arranged in a suitable format. In this research, local CDR, IDD CDR, and past SIM box fraud number details were used to create a profile for each SIM number (subscriber) with seventeen parameters. However, storing and processing a very large duration of CDR was not possible, so CDR was stored for twenty-four hours, and a subscriber profile dataset was created from the twenty-four hours of CDR. However, the processing of twenty-four hours of CDR and analyzing them via machine learning algorithm took around 20 to 30 minutes making the detection of some SIM box fraud numbers delayed for more than 30 minutes. Therefore, another subscriber profile dataset was created using CDR from the past three hours CDR as this dataset only took around 10 minutes to process and analyze via machine learning algorithm making it ideal to detect SIM box fraud numbers that make SIM box fraud calls fast.

All the machine learning algorithms such as support vector classifier, decision tree classifier, and K-nearest neighbors classifier showed an accuracy of 99.99% for both datasets. However, the macro F1-score value was low. K- nearest neighbors classifier algorithm showed the highest macro F1-score value of 0.89 and 0.92 respectively for three hours CDR dataset and twenty-four hours CDR dataset. The main reason for this high accuracy and lower macro F1-score value was due to an imbalance of the datasets where they contain more than 99% of non-SIM box numbers (genuine subscribers). Therefore, even the best-performing K- nearest neighbors classifier algorithm detected around 20% of genuine subscribers as SIM box fraud numbers.

To correct this issue, another two datasets were created by removing obvious genuine subscribers from three hours CDR dataset and the twenty-four hours CDR dataset. This makes the datasets more balanced. The decision tree classifier algorithm showed the highest accuracy of 87.56% and 87.98% respectively for three hours CDR dataset and twenty-four hours CDR dataset. The decision tree classifier algorithm also showed the highest macro F1-score value of 0.87 and 0.88 respectively for three hours CDR dataset and twenty-four hours CDR dataset. Even though accuracy and macro F1-score values were reduced due to the removal of obviously genuine subscribers, detection of SIM box numbers was improved as the best performing decision tree classifier algorithm detected only around 10% of genuine subscribers as SIM box fraud numbers. Ensemble voting classifier algorithm was used to further increase the performance of the machine learning program and it improved the accuracy to 88.10% and 88.59%, and the macro F1-score value to 0.87 and 0.89 respectively for three hours CDR dataset and twenty-four hours CDR dataset. Ensemble voting classifier models also show better performance compared to other published research models recreated using the balance dataset.

5.2. Limitations of the Research

The main limitation observed during the research was hardware limitations. A device with fifteen cores with a processor base frequency of 2.66 GHz was used for the research. However, analyzing datasets with more than a hundred thousand data points with complex machine learning programs took several hours, and sometimes it was also crash the system. This limited the amount of data that could analyze and also limited the time to work on the machine learning programs. Due to privacy policies, CDR details could not be released to outside devices with better processing powers as well, so research was carried out on the operator premise with limited resources.

5.3. Future Enhancements

A major weakness in this research was the inability of close to 100% detection of SIM box fraud numbers. The programs introduced in this research could not identify around 10% of SIM box numbers. This is mainly due to the capabilities of modern SIM box

devices that make it possible to replicate genuine subscriber characteristics by SIM box fraud numbers. Improvement in SIM box number detection that replicates genuine subscriber characteristics is a future enhancement direction that can take in this research. Machine learning technology is developing fast. Therefore, these developments in the machine learning field can be used to improve machine learning methods to detect SIM box frauds more accurately.

Also in this research, there were two separate SIM box fraud detection processes which were the detection of SIM box fraud from all the subscriber profiles (imbalance dataset) using machine learning and the detection of SIM box fraud from suspicious subscriber profiles (balance dataset after removing some of the genuine subscribers). Integration of these two processes into one process is a possible future enhancement direction. This integration will allow improvement in SIM box fraud detection as the first step will detect suspicious subscribers from all the subscriber profiles while the following step will detect SIM box fraud numbers from suspicious subscribers. Finally, these machine learning programs can be also introduced to detect not only SIM box fraud, but also other types of fraud observed in telecommunication networks.

REFERENCES

[1] S. Bandusir, "Consumer Information Statistical Overview Report Q4 2021," Telecommunications Regulatory Commission of Sri Lanka, Colombo, Sri Lanka, Apr. 27, 2021. Accessed: Jun. 20, 2021. [Online]. Available: https://trc.gov.lk/images/pdf/StatisticalOverViewReportQ1202105052021abc.pdf

[2] J. Howell, "CFCA 2019 fraud loss survey," Communications Fraud Control Association, Version 1.0, 2019. Accessed: Jun. 20, 2021. [Online]. Available: https://cfca.org/wp-content/uploads/2021/02/CFCA-2019-Fraud-Loss-Survey.pdf

[3] C. Zhou and Z. Lin, "Study on fraud detection of telecom industry based on rough set," in *2018 IEEE 8th Annu. Comput. Commun. Workshop and Conf. (CCWC)*, 2018, doi: 10.1109/CCWC.2018.8301651.

[4] P. Gosset and M. Hyland, "Classification detection and prosecution of fraud on mobile networks," in *Proc. ACTS Mob. summit*, Sorrento, Italy, Jan. 1999, no. 1, pp. 2–4.

[5] Y. Nadji, J. Giffin and P. Traynor, "Automated remote repair for mobile malware," in *Annu. Comput. Secur. Appl. Conf. (ACSAC)*, Orlando, FL, USA, Dec. 5-9, 2011, pp. 413–422.

[6] D. Baker, "International Revenue Share Fraud: Are We Winning the Battle Against Telecom Pirates?," in *Black Swan Telecom Journal*, Nov. 2012. [Online]. Available: http://bswan.org/revenue_share_fraud.asp

[7] M. Sahin, A. Francillon, P. Gupta, and M. Ahamad, "Sok: Fraud in telephony networks," in *2017 IEEE Eur. Symp. Security Privacy (EuroSP)*, 2017, pp.235–250, doi: 10.1109/EuroSP.2017.40.

[8] J. S. Taylor, K. Howkev and P. Burge, "Detection of Fraud in Mobile Telecommunications," *Inf. Secur. Tech. Rep.*, vol. 4, no. 1, pp. 16–28, 1999.

[9] "Alternative calling procedures: Background and trends," in *ITU World Radiocommunication Conf. (WRC)*, Jan. 2012.

[10] N. Alghawi, "A study on sim box or interconnect bypass fraud," M. S. thesis, Dept. Inf. Tech. Eng., The British Univ., Dubai, UAE, 2019.

[11] B. Reaves, E. Shernan, A. Bates, H. Carter and P. Traynor, "Boxed out: Blocking cellular interconnect bypass fraud at the network edge," in Proc. 24th USENIX Secur. Symp., Washington, DC, USA, Aug. 12–14, 2015, pp. 833–848.

[12] I. Murynets and R. P. Jover, “Analysis of SMS Spam in Mobility Networks,” *Int. J. Adv. Comput. Sci. (IJACSci)*, vol. 3, no. 7, pp. 330–343, Jul. 2013.

[13] R. Clayton, “Can {CLI} be trusted?,” *Inf. Secur. Tech. Rep.*, vol. 12, pp. 74–79, 2007.

[14] D. Soni, “Supervised vs. unsupervised learning,” towardsdatascience.com. https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d (accessed Jun. 20, 2021).

[15] A. Krenker, M. Volk, U. Sedlar, J. Bester, and A. Kos, “Bidirectional artificial neural networks for mobile-phone fraud detection,” *Electronics Telecommunication Research Inc. (ETRI) J.*, vol. 31, no. 1, pp. 92–94, 2009.

[16] M. Taniguchi, M. Haft, J. Hollmkn, and V. Tresp, “Fraud detection in communication networks using neural and probabilistic methods,” in *1998 IEEE Int. Conf. Acoust. Speech Signal Process (ICASSP)*, vol. 2, 1998, pp. 1241–1244, doi: 10.1109/ICASSP.1998.675496.

[17] A. H. Elmi, S. Ibrahim, and R. Sallehuddin, “Detecting sim box fraud using neural network,” in *IT Convergence and Security 2012*, vol. 215, K. Kim and K Chung, Eds., Dordrecht, Netherlands: Springer, 2013, pp. 575–582.

[18] A. H. Elmi, R. Sallehuddin, S. Ibrahim, and A. M. Zain, “Classification of sim box fraud detection using support vector machine and artificial neural network,” *Int. J. Innov. Comput.*, vol. 4, no. 2, pp. 19–27, 2014.

[19] M. Kashir and S. Bashir, “Machine learning techniques for sim box fraud detection,” in *2019 Int. Conf. Commun. Technol. (ComTech)*, 2019, pp. 4–8, doi: 10.1109/COMTECH.2019.8737828.

[20] I. Murynets, M. Zabarankin, R. P. Jover, and A. Panagia, “Analysis and detection of sim box fraud in mobility networks,” in *2014 IEEE Conf. Comput. Commun.*, 2014, pp. 1519–1526, doi: 10.1109/INFOCOM.2014.6848087.

APPENDIX A: SIMPLE MACHINE LEARNING PROGRAM

```
# Load Python library
from pandas import read_csv
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import SMOTE
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.metrics import f1_score
from sklearn.svm import SVC # sklearn Support Vector Machines library
from sklearn.tree import DecisionTreeClassifier # sklearn Decision Trees library
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import VotingClassifier
import pickle

# Load dataset
url = "dataset.csv"
names = ['a', 'tot', 'dtot', 'otot', 'dotot', 'didd', 'iddb', 'iddai', 'iddao', 'cell',
'dur', 'grcell', 'f', 'grb', 'ansd', 'ansg', 'grimei', 'imeic', 'bypass']
dataset = read_csv(url, names=names)

# Split input (X) and output (y)
array = dataset.values
X = array[:,1:18]
y = array[:,18]

# Split training and testing datasets
X, X_validation, y, Y_validation = train_test_split(X, y, test_size=0.20,
random_state=1)

# Oversampling training data
oversample = SMOTE()
X_train, Y_train = oversample.fit_resample(X, y)

# Create machine learning modle using training dataset
model = DecisionTreeClassifier(criterion='gini', max_depth=6, min_samples_split=10,
random_state=10)
model.fit(X_train, Y_train)

# Validate model using validation dataset
predictions = model.predict(X_validation)

# Evaluate predictions
print(accuracy_score(Y_validation, predictions))
print(confusion_matrix(Y_validation, predictions))
print(classification_report(Y_validation, predictions))
print(f1_score(Y_validation, predictions, average='macro'))

# Save the model to a file
filename = 'model.pkl'
pickle.dump(model, open(filename, 'wb'))
```