International Revenue Sharing Fraud (IRSF) Detection Using Data Mining Techniques: The Case of ethio telecom

BY: ALAMAREW DIGAFE

ADVISOR: EPHREM TESHALE (PhD)

A Thesis submitted to
School of Electrical and Computer Engineering
Addis Ababa Institute of Technology

in Partial Fulfillment of the Requirements for the Degree of Master of Science
(Telecommunication Engineering)

Addis Ababa University
Addis Ababa, Ethiopia
January 28, 2022

Declaration

I, the undersigned, declare that the thesis comprises my own work in compliance with internationally accepted practices; I have fully acknowledged and referred all materials used in this thesis work.

Alamarew Digafe

_____ _____

Name Signature

Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

This is to certify that the thesis prepared by **Alamarew Digafe**, entitled *International Revenue Sharing Fraud (IRSF) Detection Using Data Mining Techniques: The Case of ethio telecom* and submitted in partial fulfillment of the requirements for the degree of Master of Science (Telecommunication Engineering) complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

Examiner 1 _____ Signature _____ Date _____

Examiner 2 _____ Signature _____ Date _____

Advisor Ephrem Teshale (PhD) Signature _____ Date _____

_____

Dean, School of Electrical and Computer Engineering

ABSTRACT

One of the most often expressed concerns in the telecom industry is international revenue sharing fraud. Fraudsters are motivated to generate traffic for their services without paying the originating network. Existing detection methods include monitoring call patterns and blocking high-risk range lists obtained from various telecom associations. However, these approaches have issues that make them ineffective, time-consuming techniques, leading to financial losses before the call is blocked as they are frequently bypassed, renew high-risk range list numbers repeatedly, and make lists out of date.

The goal of this paper is to develop an international revenue sharing fraud model for real-time detection of missed call fraud generation schemes using an international voice call detail record in an hourly manner to minimize detection time as well as revenue loss. To do so, data is collected, data preprocessing is performed, and relevant attributes are determined. As classifiers, support vector machines, artificial neural networks, and random forests are utilized to classify the data set for fraud and normal call transactions.

The findings indicate that random forest techniques outperform in terms of f-measure, accuracy, receiver operating characteristic curve, the time required for building, and inference time in both training modes (percentage split and 10-cross-validation). It performed with 97.00% accuracy and also achieved comparable accuracy to the state of the art. Consequently, the support vector machine and neural network multilayer perceptron classification algorithms are found to be the next best after the random forest in terms of overall performance metrics. However, the support vector machine classifier in both test modes exceeds the acceptable detection delay in the classification process.

KEYWORDS

International revenue sharing fraud, International premium rate number, Classification, Telecom fraud, Fraud detection, Missed calls.

ACKNOWLEDGMENTS

I would like to express my gratitude to Dr. Ephrem Teshale for his great guidance, as well as for his patience and help throughout my thesis research. I would also like to thank my examiners, Dr. Yalemzewd Negash and Dr. Fitsum Assamnew, for their constructive feedback during the session. Not only that, but I would like to express my gratitude to Hanna Bizuwork, my wife, and my friends. They have been inspiring me, especially when I am experiencing difficulties. Finally, I want to thank God for providing me with the ability to complete my thesis at the Addis Ababa Institute of Technology.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ANN	Artificial Neural Networks
ARFF	Attribute Relation File Format
ASCII	American Standard Code For Information Interchange
CDR	Call Detail Record
CFCA	Communications Fraud Control Association
CFS	Correlation based Feature Selection
CRISP-DM	Cross Industry Standard Process for Data Mining
DMT	Data Mining Technique
DT	Decision Tree
EM	Expectation Maximization
F1H	Fixed 1-Hour
FMS	Fraud Management System
FN	False Negative
FP	False Positive
GSMA	Global System for Mobile communication Association
IPRN	International Premium Rate Numbers
IQR	Inter Quartile Range
IRSF	International Revenue Share Fraud
KDD	Knowledge Discovery in Databases
MLP	Multi layer Perceptron
PBX	Private Branch Exchange
RF	Random Forest

ROC	Receiver Operating Characteristic
SEMMA	Sample Explore Modify Model and Assess
SIM	Subscriber Identity Module
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
VoIP	Voice over Internet Protocol
WEKA	Waikato Environment for Knowledge Analysis

1

INTRODUCTION

International revenue sharing fraud is a kind of telecommunications fraud in which the perpetrator artificially inflates traffic by using certain parts of international number ranges without intending to pay for services or convincing others to dial those numbers[1]. As per the global system for mobile communication association (GSMA) definition, the fraudster obtains a portion of the revenue generated by the number range holder's termination charges for incoming traffic. It is the most common revenue diversion fraud, and due to high termination or destination rates, provides a strong motivation for fraudsters to commit such fraud.

Telecom revenue losses are due to telecom fraud, which is continuously altering and dynamic [2]. Many telecom operators have serious concerns about number-resource misuse, a practice whereby calls never reach the destination designated by the international country code[3]. International revenue sharing fraud is carried out as a consequence of number-resource misuse. They are terminated early through a carrier or content provider agreement for revenue-generating content services.

One of the most common types of interconnect fraud is international revenue sharing fraud, in which fraudulent operators hijack calls to certain destinations and divert them to so-called international premium rate services[4]. The fraudster teams up with a local operator who charges high call termination rates in exchange for a share of the revenue generated by the fraudster's traffic.

The fraudster aims to get revenue from the termination charges on international premium numbers. Fraudsters take advantage of the telecom operator's infrastructure to fraudulently increase traffic to high-risk foreign destinations with the intention of non-payment[5]. It is common across geographies, and it is one of the five main types of fraud and the highest fraud loss contributor, as per the communications fraud control association (CFCA) global fraud loss survey, 2019[6].

Despite telecom operators' efforts to secure services, a fraud survey report published in 2019 claimed that revenue losses have increased by 37%, with a financial estimate of 28.3 billion dollars in global fraud losses, or 1.74% of global revenues[6, 7].

The amount of international revenue sharing fraud losses as a proportion of the total losses globally from 2015 to 2019 is shown below.

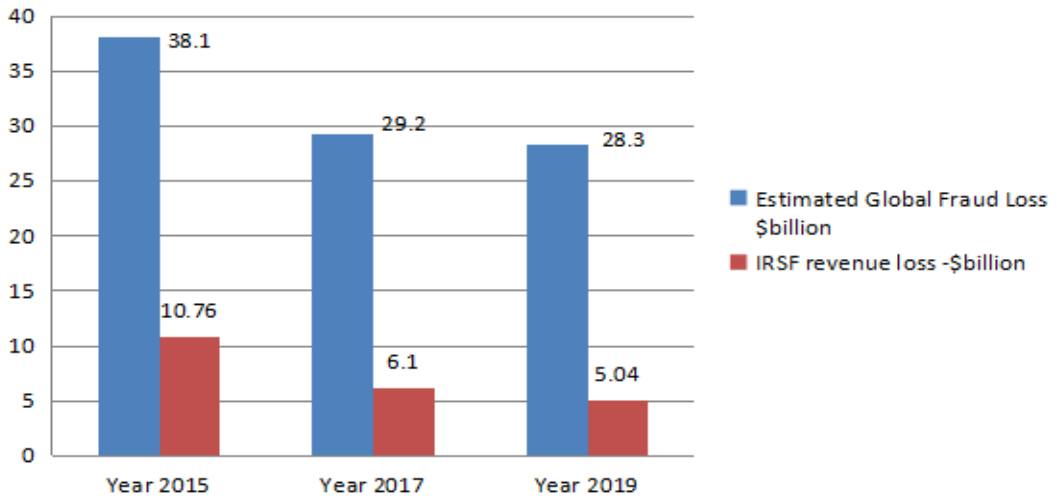

Figure 1.1: International revenue sharing fraud causes revenue losses globally[6, 7]

Furthermore, in the ethiotelecom system, from March 1 to April 30, 2021, a total of 3923 total fraud international numbers with a 142784 second duration were lost due to an international revenue sharing fraud type from a missed call scheme.

International revenue sharing fraud typically includes various aspects (for example, a fraudulent carrier in collaboration with a premium rate service provider), which collect and share the revenue from calls, and are often combined with different fraudulent methods to generate traffic calls without payment [4].

International revenue sharing fraud is committed when a fraudulent operator or third-party operator's service provider advertises many phone numbers as international premium rate numbers (IPRNs) around the world. Their target numbers, which unfortunately fall within the range advertised, often come from developing countries or small satellite operators with high termination fees. The fraudsters receive payment for every minute of traffic generated [7].

An international call initiated from a cellular or IP network travels over transit operators before reaching its destination/termination country[8]. Each of these transit operators gets a share of the call revenue for passing over the call, and the local operator in the terminating country receives a termination fee for the service.

In telecommunications, the carrier's business interconnection agreement, known as carrier trust, is bypassed by such fraudulent activities. Operators need to negotiate interconnection agreements and rely on transit operators to provide worldwide services; hence, the interconnection business is mostly for international and long-distance calls[7]. They don't have a direct link to reason, and such interconnected agreements define interconnected pricing, policy, and conflict resolution. There are also exchange platforms where operators can buy and sell minutes directly or anonymously.

Fraudsters are interested in generating as much call traffic as possible. They make a lot of calls to international premium rate numbers and share part of the profits with the IPRN provider. Fraudsters may use a variety of methods

to generate call traffic, including fraudulently obtained SIM cards, hacking private branch exchanges (PBX), missed calls, and mobile malware (malicious mobile applications) and others[9].

The missed call scam, for example, involves a one-touch technique before hanging up[4]. Fraudsters utilize this tactic to get users to call back and dial international premium rate numbers. A fraudster will construct a device to dial many random phone numbers (for example, using auto-dialers). Each call rings once and then hangs up, leaving the recipients with a missed call. Users are tempted to return, but the call is international. As a result, international revenue sharing fraud is a critical and growing issue in the telecommunications industry that must be detected and addressed. To minimize detection time and financial losses, a near real time international revenue sharing fraud detection system should be developed.

1.1 STATEMENT OF THE PROBLEM

As the number of services grows, fraudsters employ a variety of methods to gain an illegal revenue share from various types of services. As a result, a variety of methods for detecting international revenue sharing fraud have been used, the most notable of which is rule-based. Second, telecommunications industry organizations usually exchange fraud information (recommendations, best practices) between operators (for example, ethio telecom uses the GSMA high-risk range list); third, detection via a call center and bill process (where the center reports unfamiliar international numbers, the domain expert investigates suspicious and then determines the necessary action)[10]. Fourth, telecom fraud has recently been detected using a data mining technique.

In traditional systems, rule-based fraud detection often requires a threshold or identifies fraudulent behavior based on a previous set of known fraud patterns to discover anomalous data [4]. Thus, such an approach is one in which fraud patterns are predefined by a set of many conditions. Consequently, fraud detection efficiency is often limited, relying on the knowledge of domain experts, ineffective in providing early warnings, and exposed to unknown and abnormal fraud patterns and numbers of false positives[11].

Telecom industry associations such as the GSMA and CFCA and others share international revenue sharing fraud high-risk list numbers with their members, but fraudsters can easily circulate or renew international premium rate numbers frequently, making the lists obsolete[4, 7].

The summary of the existing detection system is shown in Table 1.1

Table 1.1: Existing international revenue sharing fraud detection method lists

Ref.	**Detection method**	**Techniques**	**Shortcomings**
[11]	Rule-based system	List of predefined indicators	Generate many false positives, and takes time.
[4, 7]	Monitor high-risk and fraudulent ranges list.	Hot IPRN lists numbers.	Renew IPRNs frequently; incorrectly block normal traffic, and analysis time.
[10]	Find out a report on the call center.	Manual investigation	Feedback delay

The data mining technique has become increasingly popular in the telecommunications business and other fields in recent years. Telecom fraud has been detected using data mining methods. There is, however, a trade-off between detection time and accuracy. Operators are being impacted by in-

ternational revenue sharing fraudsters. It should be discovered as soon as possible before it causes a major revenue loss. As a result, evaluating the capabilities of the fastest detection systems is crucial.

The goal of this research is to find an answer to the following question:

1. What data features or attributes can be utilized to identify international revenue sharing fraud?
2. Which classification technique is more appropriate for detecting potential international revenue sharing fraud?

1.2 OBJECTIVES

The main goal of this thesis is to develop an international revenue sharing fraud detection model in near-real-time by analyzing international voice traffic usage data for ethio telecom users using a data mining technique based on missed call fraud generation schemes.

The specific objectives of this research are:

- To select the best classification-learning algorithm for detecting international revenue sharing fraud.
- To select relevant attributes for the international revenue sharing fraud detection model.
- To detect international revenue sharing fraud in near-real-time by analyzing fraudulent behavior based on offline user data.
- To build models and evaluate their performance.

1.3 SCOPE OF THE STUDY

International revenue sharing fraud is a major source of concern for telecom operators, as it can occur on both local (home) and roaming networks. All network users, including prepaid, postpaid, and private branch exchange subscribers, are affected. International revenue sharing fraud can be detected using a variety of data mining techniques and fraud generation schemes. However, this study will focus on international revenue sharing fraud detection in local networks (i.e., domestic users) international voice calls, as well as a missed call fraud generation scheme for mobile users. Furthermore, the detection model evaluation will only compare three well-known classification algorithms: random forest, support vector machine, and artificial neural networks, as well as the Weka data analysis tool. We used ethio telecom international voice data as a case study.

1.4 CONTRIBUTIONS OF THE RESEARCH

By assessing relevant research on telecommunication fraud detection and data mining techniques for detecting international revenue sharing fraudulent calls, we discovered the first limitation of the reviewed literature was the lack of studies on the detection of international revenue sharing fraud in ethio telecom. Even though in the state of the art, multiple studies have been undertaken to analyze international revenue sharing fraud detection in various scheme scenarios, such as call generation schemes via malicious mobile software and exploiting private branch exchange breaches, no specific study on international revenue sharing fraud detection using a fixed one-hour data aggregation mode and data mining technique has been done to our knowledge for call generating schemes from missed calls. Furthermore, it detects with comparable accuracy to the state-of-the-art.

The findings of this research will help telecom operators:

- To detect fraud involving illegal international incoming and outgoing calls.
- To give operators a near-real-time fraud detection capability, enabling them to prevent fraudsters before they cause harm.

1.5 LITERATURE REVIEW

Several scientific research papers and studies on detecting telecom fraud related to international revenue sharing fraud detection using data mining techniques have been published. The below literature is closely related to the title of the study.

There are small and medium-sized operators in several countries that can only resell the services of other carriers. In order to obtain advantages, they frequently attempt to hijack phones originally operated by other operators. In this context, Sahin et al.[9] investigated the international revenue share fraud ecosystem and suggested a detection approach. They investigate internet IPRN resellers and their test portals, which are commonly used by scammers. From over 3M test IPRNs (Distinct), telephony hot spot (10K), 206K test call logs were obtained to investigate international revenue sharing fraud from a different perspective, along with an additional 689k call records from a telecom operator's real-time data set. They investigated the IPRN provider ecosystem in terms of coverage, different providers' collaboration (overlapping IPRNS), and the spread of the IPRN providers, and based on this, they discovered eleven international revenue-shared fraud detection features (related to call destination and historical call records). Analyze call logs from a victim operator (869K from small European operators) and validate them against the specified feature. As a result, the ran-

dom forest algorithm can detect fraudulent calls with a 98% accuracy rate and a 0.28% false-positive rate, which is better than the naive algorithm, which is 88.3%.

Gopal et al.[12] discussed a rule-based approach to detecting fraudulent subscriber usage. A sampled estimation from historical (study) and current (test) data was collected for three months. The study period holds call records of customers' non-anomalous behavior. The data was first categorized based on the behavior of the users. For each group, they developed a probabilistic model to describe customer usage. Next, with the help of maximum likelihood estimation, they estimate the parameters of the calling behavior, compared against thresholds. Then, any change above that threshold triggers an alert.

Sahin et al.[7] tried to figure out a complete framework for better understanding and assessing telecom fraud. The aim is to clarify inconsistencies in current fraud terminology and increase awareness among users and operators that aren't members of telecom associations such as the GSMA, CFCA, and other partnerships. A fraud taxonomy was used to identify root causes, vulnerabilities, exploitation methods, fraud types, and, finally, how fraught benefits fraudsters. Data was also collected through interviews and surveys with a range of subject experts, as well as participation in industry forums.

N.Jiang et al.[13] developed a Markov clustering-based method to detect fraud activities on voice calls in a cellular network, using 6-month international voice calls collected in the UMTS network's MSCs by large mobile operators. These phone calls to international terminating numbers are initiated by mobile users on the cellular network. The IRSF list and online feedback serve as ground truth on fraudulent destination numbers, and they compare the proposed model's effectiveness against the two sources. In general, the approach is successful in detecting the destination numbers for 78% of the IRSF calls in the dataset. But, this method appears to produce

many false positives: Only 9.3% of the 24K potential phone numbers they found suspicious in the ground truth dataset were involved directly with IRSF or other fraud activities.

A. Prakasam et al. [10], used with diverse mobile customers, assume a lower probability of contacting the identical set of international terminating numbers frequently. Based on this key assumption, they proposed a data mining approach for identifying international revenue sharing fraud activities on large-scale cellular networks initiated by malware fraud generation schemes. The model was evaluated and tested on a one-year dataset using three steps: time-series analysis (Expectation-Maximization algorithm) and a change detection algorithm to identify emerging popular international terminating numbers, Apriori mining was used to identify correlated foreign phone numbers, and then they associated them with billing information to confirm the detection results. Two sources of ground truth are used to evaluate the high-rate IRSF number of candidates. 10% came from the IRSF list, and 40% came from internet feedback on popular forums.

To detect calls from unknown subscribers and inspire the local subscriber to call back unknowingly to premium numbers, which is often the revenue-generating part of the IRSF schemes. To find fraudulent call behavior, M. Arafat et al. [14], have used a process model. They used three ensembles learning data mining algorithms, such as random forests, AdaBoost, and boosting methods, and evaluated those ensemble classifiers. Extreme gradient performance is best among the other two algorithms.

A. Mehadi[15], focuses on the detection of toll fraud committed through enterprise private branch exchanges in the ethio telecom network. The methodology they used was 4998 private branch exchange users by analyzing data traffic, and additional synthetic call detail records were also used in the research. Findings were conducted on both the k-means and EM al-

gorithms. The results show that the k-means algorithm performs 97.96% in 0.09 sec, which is higher than the EM algorithm, which is at 93.88%.

To investigate the IRSF ecosystem for online IPRN providers in order to better understand how they work and how they misuse international phone numbers, Festor [4], collected the CDR of 517,319 unique test numbers from ten websites advertising as well as commercial numbering plans databases to extract further information on the test numbers and verify their validity. Thus, there is a broad variety of ranges that can potentially be used for both fixed and mobile numbers by the IRSF, including invalid length, unallocated number ranges, and different operator types are used as IPRNs.

Meijaard et el.[16], aims to detect calls at the time of their initiation, preventing IRSF from occurring. The model they used is an isolation forest, which is an unsupervised data mining approach. The data traffic network from and to private branch exchanges consists of over 10K VoIP data. The result is a pre-call detection success of at least 45% up to 87% false-positive rate of 2% to 5% respectively. But using such an approach does not allow for enforcing which features should be considered and has too high false-positive rates.

Table 1.2: Summary of related works

Ref.	Objective	Techniques	Data set	Key Findings
[9]	Explore the IRSF ecosystem from multiple angles and propose IRSF features	Random Forest(10-fold cross validation)	Unused range, 689K real CDR, 3.14M test IPRNS, and 206K test call(PBX and stolen SIM cards)	98% accuracy with a 0.28% FP over Naive method which is 88.3%
[7]	Holistic framework for telecom fraud	Taxonomy	Causes, weaknesses, techniques, schemes, and benefit	Clarify the inconsistencies
[12]	Detect anomalous call usage	Two periods (study, test) and estimation (calling behavior)	Study data(90 days) and test data(10 days)	Identify 90% accuracy, with less than 1% FP
[10]	Detect IRSF activities	EM, Apriori, and confirm with billing	Incoming calling traffic (malware scheme traffic generation)	Validate:from IRSF list (10%), online feedback (40%)
[13]	Isolate fraudulent activities	Markov Clustering (MCL)	6-month voice calls and online complaints (IRSF list and online feedback)	low accuracy and high false positive rates.
[15]	Detection of toll fraud	K-means, EM, and Rule-based	5K PBX users real and synthetic data	k-means: higher performance
[14]	Detect missed calls	Gradient Boosting, AdaBoost and RF	The data set has used a total of 2M calls	XGBoost: higher performance
[16]	Identify IRSF calls as anomalies	Isolation Forests	Traffic (PBX), data set over 10K CDR	45% valuable scenario (has too high FP)

1.5.1 *Summary*

In summary, literature that detects international revenue sharing fraud in various fraud generation scenarios, such as call generation schemes via malicious mobile software, exploiting private branch exchange breaches, missed calls, and stolen SIM cards from larger data aggregation periods, such as four-hours, daily, and so on, etc. While this may improve accuracy, it also opens up a wider window for fraudsters to operate and profit. As a result, adopting a near-real-time detection scheme necessitates a narrow data granularity level. However, with fewer aggregation levels of data, there are fewer distinguishing patterns for detection.

1.6 METHODOLOGY

The methodologies used to achieve the general and specific goals of this thesis are as follows:

- A literature review on international revenue sharing fraud and data mining techniques, namely, SVM, RF, and ANN-MLP, are conducted.
- A Cross-industry Standard Process for Data Mining (CRISP-DM) is used as a process model for experimentation.
- Raw CDR and labeled international numbers are collected from the mediation and fraud management systems. All raw data is collected for two months, after which it is pre-processed, aggregated to a fixed one-hour, and relevant features are chosen.
- We used performance evaluation metrics such as accuracy, f-measure, time taken to build a model and inference time, and receiver operat-

ing characteristic curve to assess the classification algorithm's performance.

1.7 THESIS ORGANIZATION

The remaining content is organized as follows. Chapter 2 discusses what international revenue sharing fraud is and how it works. It also describes some data mining techniques and tools that were used in this study. The experimental design used is described in chapter 3, and the results of the experiments are presented in chapter 4. Finally, Chapter 5 contains the conclusions and future works.

2

INTERNATIONAL REVENUE SHARING FRAUD AND DETECTION METHOD

This chapter discusses the theoretical background of international revenue sharing fraud and detection methods, as well as process models. The first section defines international revenue sharing fraud and explains how it works. The second section is a literature review of state-of-the-art detection methods published on the subject of fraud detection.

2.1 INTERNATIONAL REVENUE SHARING FRAUD

International revenue sharing fraud can affect all users of the telephone networks and targets a wide range of countries with varying call termination costs [17]. To do this, fraudsters can generate illegal traffic calls in various ways, for example, by using fraudulent SIM cards, missing calls (tricking fixed/mobile users), compromising the company's telephone system (private branch exchange) to make calls, or via mobile malware that invisibly calls international premium numbers.

The absence of a centralized numbering plan database that keeps up-to-date information on all operators in every country, as well as a least-cost routing policy to increase profit, are major enablers of international revenue sharing fraud[4].

A fraud agreement plan often involves numerous parties collecting and sharing call income, and it is frequently paired with traffic generation methods to create calls without payment[4]. Value-added services (e.g., international premium rate services) are often manipulated for revenue share fraud. Thus, the main fraud actors described are as follows:

- Fraudsters purchase international numbers from international premium rate number providers for a revenue share that offers revenue in exchange for generating traffic to these numbers.
- Fraudsters generate a huge number of international calls.
- Calls terminate through the international carrier, eventually through the IPRN provider who monitors the termination phone numbers and billing flow process.
- And then, the IPRN providers share revenue with the fraudsters.

In international revenue sharing fraud, a fraudulent operator advertises a range of phone numbers as IPRN in various parts of the world. This victim number range often belongs to a small, developing country, or a satellite operator with a high interconnect termination fee.

The paper [4] classifies different kinds of fraud schemes, and these are:

- Legitimate terminating operator (owner of the victim number range) resells its numbers as IPRNs to a premium rate provider and terminates the illegal traffic on its own network.
- Operator is terminating the calls illegitimately: a transit operator can make an agreement with a premium rate service provider to misroute the hijacked phone calls.

2.1.1 *International Revenue Sharing Fraud Scenario*

A domestic (country A) caller, as indicated in Figure 2.1 with a green broken line, will pay a fee to his or her operator to call the called party in country B [18]. Due to the lack of a direct link between these two operators, the call must be routed through numerous transit operators, with each operator, such as operator A, routing the call through several transit operators, each with varying quality and pricing. If operator A selects T3 as the transit operator, T3 will present many options, one of which may be T4, which will pay the destination operator's international termination costs, and calls will be terminated at the destination operator.

However, in the case of international revenue sharing fraud, as depicted in Figure 2.1 with the red broken line, fraudsters use various fraud techniques to make many calls on behalf of others, and the call route has a shady, rogue transit operator to some extent. Instead of forwarding the call to the intended recipient, the operator can make a contract with a premium service provider to pick up the call and forward it to the premium service provider [7, 9, 17]. They do not have to pay money to operator B in this case; instead, they can keep the money and pay a portion of the earnings to the fraudster for each minute of the call he generates.

The most common fraudulent call generation techniques include[9]:

- Malware is installed on mobile phones, which can make phone calls without the caller's knowledge.
- Automated calls are made using SIM cards that have been stolen or hacked.
- Hacking company networks and gaining control of private branch exchanges to make calls to premium numbers.

- Missed call scammers make short calls to users from international premium phone numbers in the hope that they will call back. For this research, we have used this scheme as a case study for experimentation and analysis.

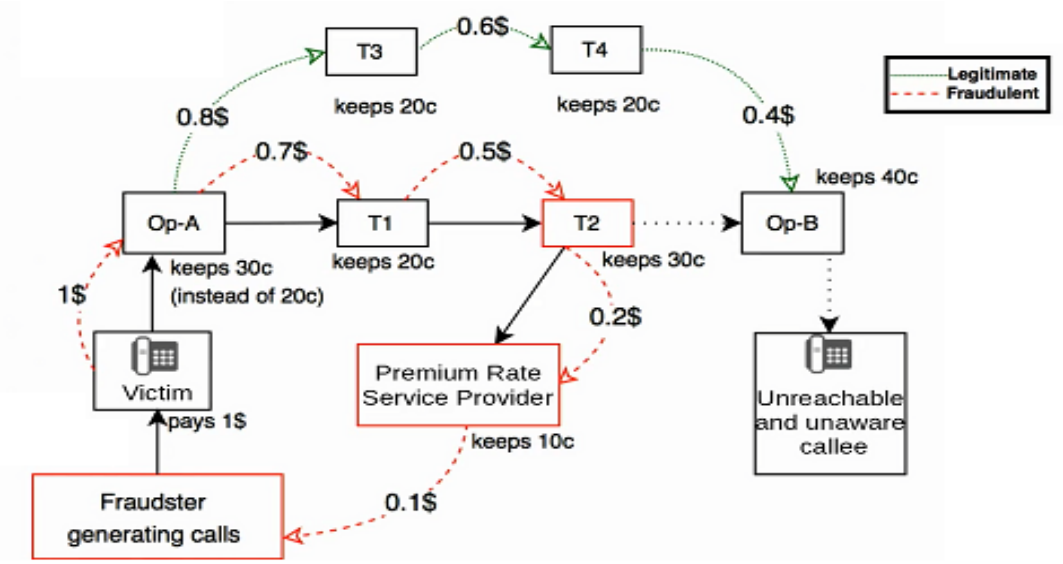

Figure 2.1: Regular call and international revenue sharing fraud call scenario[17, 18]

2.1.2 *Missed Call Scam properties*

The fraudster then attempts to artificially increase traffic to the required number ranges after routing arrangements are in place from IPRN providers. Among several strategies, one of the fraud schemes is to mislead many legitimate subscribers to generate traffic to the target number range, such as by tricking recipients of missed call notifications into calling fraud numbers from their subscriber cellphones [19].

The fraud generation scheme technique for tricking victims into dialing fraud numbers is as follows[7, 19].

- Fraudsters will use a variety of ways to place thousands of calls, frequently to random target numbers (e.g., auto-dialers, which are capable of making thousands of calls per minute).
- The fraudster will end the call immediately.
- The victim returns a call to the initiating number, and the scammers keep them online as much as possible.

2.1.3 *International Premium Rate Numbers*

A fraudulent operator or third-party service provider promotes a range of phone numbers as premium numbers worldwide. This target range is frequently associated with a small, developing country or a satellite operator that charges a high connection termination fee[7]. Fraudsters frequently employ illegal means to gain access to route traffic into international numbers obtained from an IPRNs provider. For example, the number of advertisements on every continent in April 2020 has roughly tripled as a percentage of those advertised in April 2019[20].

IPRNs that were in the market from April 2019 to April 2020 on a continental basis are presented in Figure 2.2

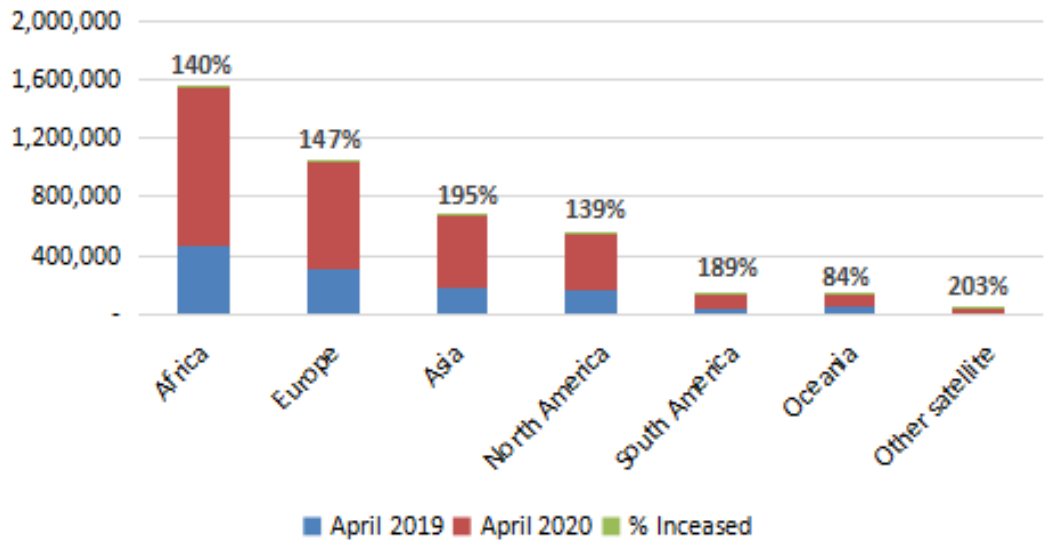

Figure 2.2: International premium rate numbers lists are advertised globally[20]

The IPRNs provider provides a test call interface for fraudsters to perform the following activities [7, 9, 17]:

- To check if the calls they initiate are routed via the involved transit operator.
- Advertise/publish test numbers, so fraudsters can check the list of successful test calls,
- Maintain real-time access to call logs and keep track of call detail records for calls made to the test numbers.

Many international premium rate number service providers offer promises of quick, easy money pay-back guarantees for the call traffic generated by fraudsters. Besides, they offer a range of payment methods and pricing plans for their customer[4].

2.2 DATA MINING TECHNIQUES

In telecommunication fraud detection, data mining is critical in the analysis of numerous forms of data to make strategic decisions[21]. Due to the large records of the data sets, the sequential and temporal features of the data, and the requirement to predict occurrences such as customer fraud and network failures in real-time. Telecommunication fraud detection techniques based on data mining methods help automate, speed up, and successfully detect abnormal call behavior.

Data mining techniques are used for discovering patterns and converting raw data into user-understandable information. It has recently been used in science and technology to extract a large amount of information. The practice of discovering knowledge from large volumes of structured data sources is known as data mining. It gives insights that are statistically reliable and previously unknown from data. According to [22], data mining techniques include clustering, classification, regression, prediction, outlier identification, and visualization. We focus on data mining techniques for three classifiers in this research; namely, support vector machines, random forests, and artificial neural networks. These classification techniques are widely utilized in fraud detection and the state of the art[23].

2.2.1 *Data mining Techniques in Fraud Detection Capability*

Data mining techniques are being more widely used in a variety of sectors, and they have numerous uses. [24, 25] identify telecom fraud, computer intrusion, and credit card payment fraud trends. Furthermore, the authors[23], give a well-organized and exhaustive literature analysis (from

2004 to 2015) on identifying financial fraud utilizing data mining techniques.

2.3 CLASSIFICATION TECHNIQUE

In this research, we discussed three classifiers learning data mining techniques that have been widely used in the literature for detecting fraud sectors. These classifications are random forest, support vector machine, and artificial neural network. All three classifiers are described in the upcoming subsections.

2.3.1 *Support Vector Machine*

The support vector machine is an effective classifier-building method. Its purpose is to establish a decision boundary (hyperplane) between classes that allows labels to be classified by one or more feature vectors. The concept of a margin on either side of a decision boundary that separates data classes is important for support vector machines. Increasing the margin and therefore generating the biggest possible space between the separating hyperplane and the instances on either side of it reduces the expected classification error[26, 27].

Given a labeled training data set (linearly separable):

$$g(x) = wx_i^T + b \tag{2.1}$$

Where,
g(x): linear classifier function
x : feature vector

w : adjustable weight vector to control direction of the hyper plane
b: bias which control the hyperplane position

For all input vectors belonging to class one, this function returns values greater than one. In other cases, for all values in class two, it will also return values that are less than minus one. The classifier represented by the normal vector w and bias b of the hyperplane is developed using the SVM training algorithm. This hyperplane maintains as much distance between classes as possible.

The objective of training an SVM model is to find the w and b so that the hyperplane separates the data and maximizes the margin $\frac{1}{\|w\|^2}$

The kernel method, which is capable of modeling higher dimensional, non-linear issues, is used to separate high-dimensional features input and mapped to a high-dimensional feature space[26]. The most popular types of kernels used in support vector machine are linear kernels, polynomial kernels, radial basis function (RBF) kernels, and sigmoid kernels. The RBF, for example, is a popular kernel function.its value depends on the distance from the origin. The data is modeled using RBF networks in the form of circles (radial shapes). Mathematically,

$$KRBF(x, y) = e^{-\gamma\|x-y\|^2} \tag{2.2}$$

$$\|x - y\| = \text{Euclidean distance between } x \text{ and } y$$

2.3.2 *Random Forest*

Decision tree learning creates a tree-based structure with class conditional probabilities at the tree's branches' ends[28]. It begins with a root node and grows into subtrees with internal nodes connected by emanating branches, ultimately ending in terminal nodes known as leaves. Each branch is a

binary partition of the test attribute, and each internal node represents a feature test.

Decision tree classifiers are easy to understand, can handle both numerical and nominal input, and are simple to build. Choosing a split attribute for each internal node and determining the appropriate number of levels for each tree branch, however, are two key downsides. Ensemble learning methods generate numerous classifiers and combine their output. Given the same amount of training data, it is widely accepted that the achievement of a group of many weak classifiers is usually better than a single classifier [29]. Boosting, bagging, and, more recently, random forests are all well-known ensemble approaches.

The notion of "at random" refers to the fact that each tree is sampled equally[30]. A random forest is a tree-based ensemble strategy that uses a majority vote or average to classify new instances[29]. It creates several decision trees. During the training phase, each decision tree is built using a subset of independent attributes. Each of these trees represents a single classifier that forms an ensemble classifier when combined.

Random forest design methods combine the concepts of bagging and random feature selection[31]. By sampling with replacements from the original training set and ensuring that the bootstrap duplicates have an identical number of instances as the original instance set, a bootstrap technique is used to build numerous randomized training sets.

The random decision tree learner has a parameter $1 \leqslant v < d$. A tree partition is formed by recursively partitioning the instance space R^d, such instance space corresponds to the root of a random tree. v variables are chosen uniformly at random from the d candidates x(1),..., x(d) at each phase of the tree building (d)[29]. If a majority vote is used in each cell, a leaf node split along one of these v variables is chosen to reduce the number of

miss-classified training points. The process is repeated until all the cells are zero.

Each node in the decision tree-building process considers only a small portion of the available variables and handles huge data sets with higher dimensions[29]. The classification error of a forest and its trees is determined by the strength of individual trees in the forest and the association between them.

2.3.3 *Artificial Neural Network*

For the goal of data mining learning, many algorithms have been created and deployed. Neural networks have played an essential role in the task of algorithms[32]. Artificial neural networks are built up of many of simple processing units connected by a complex communication network. Each node or unit is a simplified model of a true neuron. The strength of these connections can be adjusted to allow the network to perform various tasks based on node patterns.

A neuron is a fundamental processing unit in an artificial neural network. Each neuron has a single output as well as a single output[33]. A weight, also known as a factor or parameter, is assigned to each input. The related weight is multiplied by the input signal to each neuron, and the result is aggregated and communicated by a transfer function. If the summed result exceeds a certain threshold, the neuron output is activated.

$$\text{Input layer} \rightarrow \text{Output layer} \Rightarrow \begin{cases} 1 & \text{if } \sum w_i x_i > \theta \\ 0 & \text{otherwise} \end{cases}. \tag{2.3}$$

The perceptron can be employed with a variety of activation functions, the most popular of which are the step, sign, linear, and sigmoid functions. The summing function, for example, calculates the outputs of all neurons in the hidden layer is given a set of inputs x_j and matching weights w_j between the input and hidden neurons as shown in Equation 2.4. An example of a sigmoid function shows a simple perceptron.

$$y_i = f(\sum_{j=1}^{n} wjxj + b) \tag{2.4}$$

where:
y is the output, x is the input, f function (sigmoid), w is the weight and b is the bias.

Thus, the sum of these products is then supplied into the transfer function using the formula Equation 2.5.

$$f(x) = \frac{1}{1 + e^{-s}} \tag{2.5}$$

Where:
f(x) generate the output

Another type of neural network method is the multilayer perceptron (MLP). It solves a wide range of classification problems [32, 33]. The three levels of an artificial neural network are the input, hidden, and output layers. All layers are connected to all the inputs in the subsequent layer. The input layer receives the initial data, while the hidden layer computes a series of interim values that are utilized to construct output values in the output layer.

Recurrent neural and feed-forward are types of neural networks based on how the nodes in a network are connected[34]. Feed-forward networks cannot be interconnected recurrently, whereas recurrent networks can [32].

Feed-forward networks include MLPs. An input layer, several hidden layers, and an output layer are all present in an MLP. The back-propagation learning technique is a key aspect of MLPs, and it can handle several problems, including non-linear separability that is impossible to handle with perceptrons.

The back-propagation learning technique and the sigmoid activation function in the hidden and output neurons are used to build artificial neural networks models in this research. The back-propagation algorithm has two phases: a forward phase in which activation is propagated from the input layer to the output layer by combining all the weighted inputs and then process result using a sigmoid threshold [32]. In a backward phase in which the difference between the observed actual and the requested value in the output layer is transmitted backward.

2.4 PROCESS MODEL

In order to implement classification techniques properly and efficiently, a defined framework should be followed[35]. Knowledge Discovery in Databases (KDD), sample Explore Modify Model and Assess (SEMMA), and the CRISP-DM are all data mining process models. These process models outline the procedures that should be taken while developing a data mining technique. CRISP-DM is an industry standard that specifies a series of procedures for applying data mining techniques. It is used widely in the literature. As a result, the international revenue sharing fraud system is implemented using the CRISP-DM process model in this thesis.

The CRISP-DM methodology employs six steps. The phases used in the process model are as follows:

- Business understanding: choosing objectives and business goals.

- Data understanding: consists of considering the data requirements and initial data collection.
- Data preparation: activities that go into producing the final target dataset from the raw data.
- Modeling entails choosing and applying suitable modeling techniques.
- Evaluation: corresponds to the evaluation of the model experiment results.
- Deployment: final and implementation phase.

3

EXPERIMENTAL DESIGN

The main objective of the study is to compare and assess the classification performance of three algorithms over a one-hour aggregate period to detect international revenue share fraud. The cross-industry standard procedure for the data mining process model is used for data understanding, data preparation, model training and testing, and evaluation.

The classifiers' performance is evaluated using 10-fold cross-validation and the percentage split training option. Many measurement parameters are used to evaluate the performance of models. In this study, the detection performance of model results of classification techniques is measured using accuracy, f-measure, time evaluation (built and inference time), and the receiver operating characteristic curve. The whole methodology is depicted as follows.

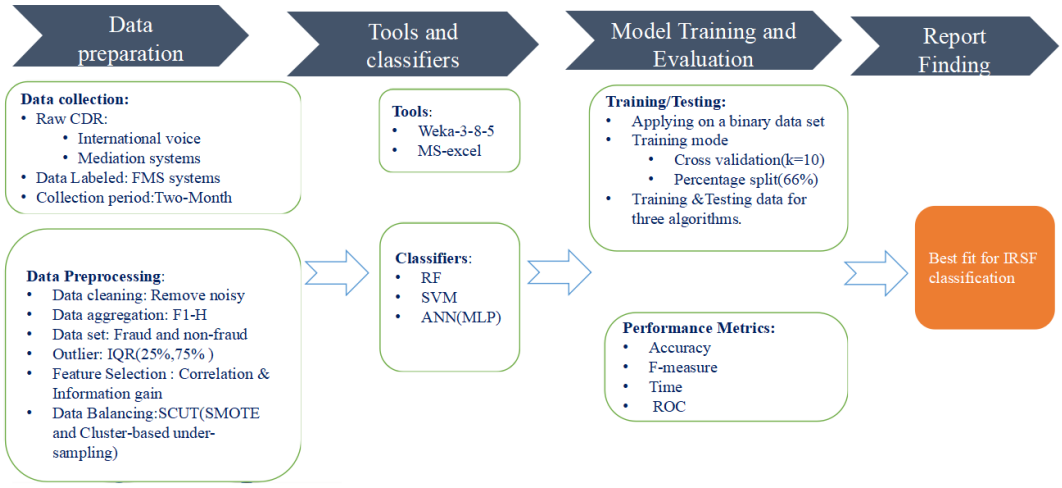

Figure 3.1: System methodology used

3.1 UNDERSTANDING THE DATA

The first step in extracting knowledge from target data is to understand the row data. This phase includes identifying fields, analyzing the values they contain, and assessing their importance for this study. A thorough analysis of the data and its structure is required as part of this process. Data-mining techniques, as well as the relationships between the data and the problem at hand, are evaluated.

The total collected data, as shown in Section A.1, consists of a total of thirty fields. Identifying valuable fields, valuing the values they contain, and evaluating their importance are the main goals of this research. The data set contains information related to international calls, such as the number initiating a call, receiving the call, time of event initiation, duration of the call, and service type.

3.2 DATA SELECTION

The focus of this study is to detect fraud through the use of missed call fraud scheme behavior. To do this, data was collected over two months, from March 1 to April 30, 2021. The 197,674 records are generated by randomly selecting raw call detail records with thirty attributes from normal international 92726 and 3921 fraud numbers. Oracle databases are used for the data extraction process. Aside from that, international voice data is also utilized in this research. Storing all the raw data for that period is challenging due to its bulkiness and resource limits. In addition, there is an insufficient number of fraud records, affecting the performance of the classifiers.

The following subsections go through data selection in-depth, including field selection and sampling.

3.2.1 *Field Selection*

We focused on the fields that represent and contribute to the study's objectives being achieved before we started selecting features. The performance of algorithms is directly determined by the quality of selected input data. For two months, international voice data from mediation was collected. After collecting data, we discovered relevant features that may help to determine fraudulent behavior.

In this study, the behavior of IRSF fraud was used as an input to the data selection process. Aside from that, for IRSF feature selection, domain expert opinion and literature review[4], played a crucial role in the selection process for minimizing unrelated attributes that help to decrease training time, minimize complexity, and increase algorithm performance.

On the collected call detail records, some fields are empty, while others are unrelated to the thesis study. Original fields before derived features are chosen from the thirty fields considered the most important for the study. In Section A.2, the fields that are chosen are listed. Furthermore, IRSF fraud definitions in this case study are based mainly on voice service types, which are the only ones taken into account.

3.2.2 *Sampling*

Data sampling is an important step in data mining, and it is regularly used to address problems with large amounts of data. It is the systematic selection of some sample elements from a larger data collection to estimate or learn something from it at a low cost. Sampling is used to reduce the data to be considered because mining a massive data database takes time. Furthermore, the fraudulent numbers are much lower than normal, resulting

in an imbalanced data set proportionality between the normal and fraud classes.

The international numbers are chosen from a mediation database by using simple random sampling for fraudulent and non-fraudulent classes as the basis for constructing classification rules to detect future cases of fraud. Each international number has an equal chance of being chosen. Simple random sampling is a technique for selecting a smaller sample size from a larger dataset to conduct research and make generalizations about the group[37, 38]. The ease of usage and representation of the larger data source are two benefits of a simple random sample. Based on this, a total of 96,647 international numbers were used in the study.

Table 3.1: Incoming and outgoing international voice sample dataset

CDR Type	**Number**	**Percentage (Number)**	**Record**	**Percentage (Record)**	**Class Label**
Normal Calls	92726	95.9%	172,956	87%	N
Fraudulent Calls	3921	4.1%	24718	13%	F
Entire data set	96,647	100%	197,674	100%	

3.3 DATA PREPROCESSING

Data collection is normally an uncontrolled procedure that results in out-of-range values, missing values, and other issues [39]. Analyzing data that hasn't been thoroughly checked for such issues can lead to inaccurate conclusions. As a result, before executing an analysis, the representation and quality of data must come first. Knowledge discovery is more difficult to conduct when there is a lot of irrelevant and redundant information or noisy and inaccurate data.

Data preprocessing includes tasks like data cleaning, aggregation, and formatting, as well as feature selection. A final data set, which can be considered correct and useful for further data mining algorithms[21]. In the model-building process, data preprocessing is crucial. The raw call detail record data is not used directly for data analysis in this research to develop a classification model to detect IRSF fraud. The call detail records are aggregated into a single user's behavior. The summarized data for each user was created as a column and row data set[38].

3.3.1 *Data Cleaning*

The practice of identifying and correcting (or removing) records from a data set is known as data cleaning[38, 39]. The data for this study contains errors such as incomplete values (a combination of hexadecimal and decimal integers), missing values, duplicate records, and so on. Records having any error values in any of their fields are not included in the target dataset.

3.3.2 *Data Aggregation*

The volume and quality of data used determine the data analysis observations obtained. It is essential to collect large amounts of high-quality data in order to gain relevant results. The procedure of collecting data and presenting it in a summary format is referred to as data aggregation. Based on international voice calls, we selected relevant attributes for aggregation. Such attributes are the frequency of calls made from source to destination, incoming calls, calling duration, called duration (in seconds), incoming trunk id, and outgoing trunk id. We also used derived attributes obtained from the original features, such as average calling duration, average called duration,

distinct ratio of calling, and distinct ratio of called. Based on this, a total of ten attributes are used for the experimentation.

The aggregated features of voice call usage at a single international number level is depicted in Table 3.2

Table 3.2: Description of selected attributes

No.	Field Name	Description
1	CALLING_TIMES	Frequency of calling made from source to destination
2	CALLED_TIMES	Frequency of called number
3	CALLING_DURATION	Time take to conduct outgoing call
4	CALLED_DURATION	Time take to conduct incoming call
5	AVG_CALLING_DU	Average calling duration
6	AVG_CALLED_DU	Average called duration
7	IN_TRUNK	Incoming number trunk id
8	DIST_RATIO_CALLING	The ratio of distinct calling over total calling
9	DIST_RATIO_CALLED	The ratio of distinct called over total called
10	OUT_TRUNK	Outgoing number trunk id
11	CLASS	Class to be classified (Normal or Fraud)

3.3.3 *Feature Selection*

Various attributes are collected throughout the data collection procedure in many cases, even if they are not important. Moreover, there are many irrelevant features present, so knowledge discovery is more difficult to conduct. Hence, the primary objective is to maximize classification accuracy[33]. The feature selection method requires the selection of important parameters to improve the performance of the learning model. The most significant part of feature selection is deciding which best distinguishes between classes. For different types of data, several feature selection techniques may be more

appropriate[38]. The method employs a variety of algorithms. There are different ways of evaluating attributes: subset and individual attribute evaluation. For this research, we selected an information gain and a correlation evaluation approach from individual feature selection for cross-checking and validation purposes, both of which use ranker-based search methods to select relevant parameters for specified data mining techniques. The ranking technique applies weights to the given features based on the value of the attributes against classes. The features are then prioritized based on their weights, and those that are most suitable for use in the learning method are filtered. Based on this, both techniques are used to rank the importance of features, from most important to least important.

We processed the Weka train set and generated the results using these attribute approaches. The complete list of outcome variables is shown in both Figure 3.2, and Figure 3.3, and the attributes are ranked according to the results of the correlation attribute and information gain selection process.

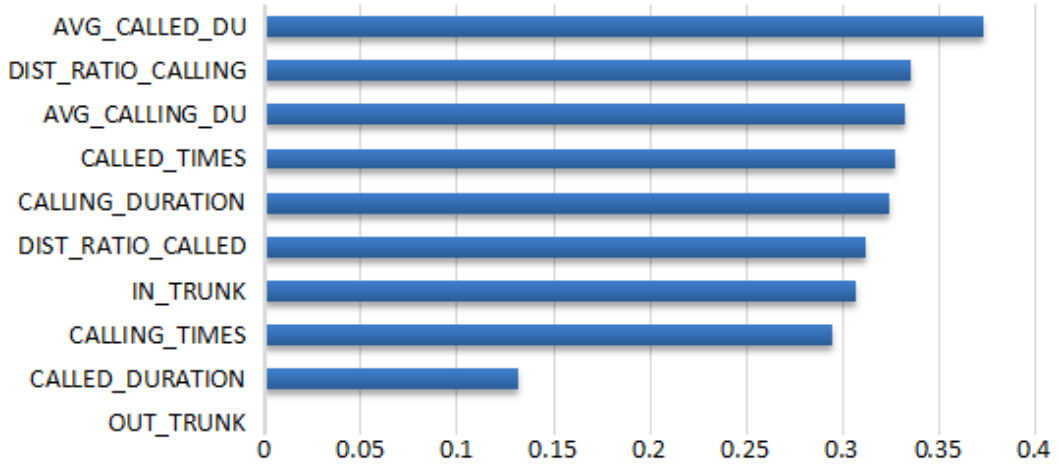

Figure 3.2: Correlation ranking attribute selection result

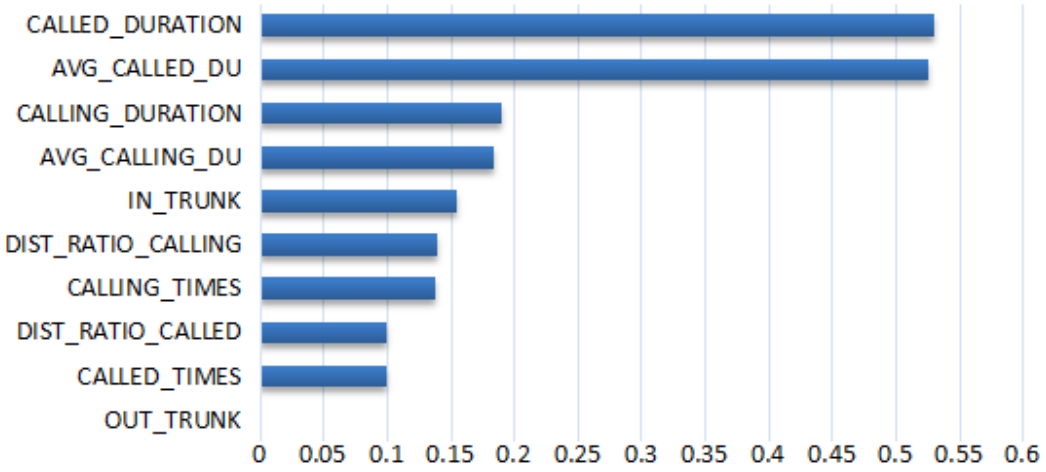

Figure 3.3: Information gain ranking attribute selection result

We determined that one feature out of ten, out_trunk feature, gives no information and only plays a minor role in detecting international revenue sharing fraud, based on the outcomes of the assessments. As a result, this feature was deleted from the data set.

3.3.4 *Data Set Formatting*

The data format is important for understanding data, representation of data, storage space requirements, data input-output during data processing, intermediate processing results, in-memory data analysis, and overall processing time[40]. Different data mining techniques require different types and formats of input data. For any type of data storage, a variety of data formats are accessible. Comma-separated values, Tab-separated values, attribute-relation file format, sequence file format, and zip file format are some examples of data formats.

The attribute-relation file format contains an American standard code for information interchange(ASCII) coding scheme that illustrates multiple cases

that share common features. It has two different units: header and data information. The name of the association, as well as an array of features and types, appears in the header portion of these files.

We used Weka to analyze the data format. It recommends loading data in the attribute-relation file format or comma-separated values formats.

3.3.5 *Outliers Detection and Removal*

An outlier is a value of typical data points that is far away from the mean or median, having variable values that are significantly different from the majority of data points. It is a variable that emerges at the extremes of its range. It is difficult to determine whether an observation is an outlier; the selection varies for each application because it is dependent on the context and model to be created.

Outlier detection mainly are based on distance, density, or distributions. In data mining, distance-based techniques are popular, where an outlier is recognized based on the distance between observations. Range quartiles, interquartile range, box plots, variance, and standard deviation are the most frequent data dispersion measurements.

The interquartile range shows whether the values of cases are outliers or extreme [22]. Both values have meanings depending on the difference between the 25^{th} and 75^{th} quartiles of an attribute's values, and they are marked as extreme if they exceed the 75^{th} quartiles or fall below the25^{th} quartiles. Outliers are values that are not extreme but exceed the 75^{th} quartiles or fall below the 25^{th} quartiles by the product of the outlier factor and the interquartile range. A data set's quartiles divide the data set into four parts, each containing 25% of the data[41].

- The first quartile (Q1) corresponds to the 25th percentile.
- The second quartile (Q2) is the median or 50th percentile.
- The 75th percentile is in the third quartile (Q3).

The interquartile range (IQR) is the spread of the middle 50% of the data and it is difference between Q1 and Q3 as shown in Equation 3.1

$$IQR = Q3–Q1 \tag{3.1}$$

A data value x is an outlier if either is depicted in Equation 3.2, Equation 3.3 respectively.

$$x \leqslant Q1 - 1.5(IQR) \tag{3.2}$$

$$x \geqslant Q3 + 1.5(IQR) \tag{3.3}$$

Outliers are detected and removed from the aggregated data once the process is completed. Table 3.3 displays the number of detected and removed outliers.

Table 3.3: Summary of aggregated data set results with outliers

Aggregation Mode	**Normal Calls**	**Fraudulent Calls**	**Total Record**	**Extreme and Outliers Values**
Fixed-one hour	114440	6494	120,934	29,162

3.3.6 *Data Balancing*

Data imbalance occurs when the sample size from a class is very small or larger than the another class. The performance of classified models is greatly affected when the dataset is highly imbalanced and the sample size

increases. Overall, imbalanced training data has a major negative impact on performance. It makes classification of noisy data and dataset deviation[42]. Class imbalance is a challenge for most learning algorithms [38], which are biased toward recognition of the dominant class and assume an essentially balanced class distribution. As a result, instances involving minority groups are frequently falsely labeled.

In this research, processed data before balanced data set distribution is shown in Table 3.3. From this data set, it is necessary to balance data sets to solve the problem of imbalanced data sets. Data sets are available for the normal (negative) class of 95.9% and the fraud (positive) class of 4.1%.

To achieve a balanced data set, various re-sampling techniques are available. Under-sampling is a method of reducing the number of instances in a majority of class data set to make it comparable to minor classes, while oversampling generates additional data within the minority class. The Synthetic Minority Over-sampling Technique (SMOTE), which generates new synthetic minority instances by interpolating between minority class samples that are close together, is one of the most well-known of these techniques.

Due to the limited number of minority classes as a consequence of identifying the right balancing data set for both legitimate and fraud classes, both techniques, such as SMOTE and Clustered Based Under Sampling Technique[43], are used in this study.

The SCUT (SMOTE and Clustered Based Under Sampling Technique) algorithm works in the following manner:

- The numbers of instances of each class's mean (m) is calculated.
- Oversampling: SMOTE is for minor classes, computed so that after oversampling, the number of instances in the class the equals mean (m).

Table 3.4: Resample a data set by applying the SCUT techniques

Class label	**Number**	**Percentage (Number)**	Record	**Percentage (Record)**	**F1-H**	**After IQR**	**Sampling Percentage**	**Final Dataset (1:1)**
N (Majority)	92726	95.9%	172956	87%	114440	89428	EM (51.311%)	45886
F (Minority)	3921	4.1%	24718	13%	6494	2344	SMOTE (1957.59%)	45886
Total	96,647	100%	197,674	100%	120,934	91,772		

- Under-sampling: Expectation-Maximization (EM), through cross-validation, to create a cluster. Instances are randomly selected for each cluster inside the major class so that the total number of instances from all clusters is equal to the mean (m).
- Finally, merge the two sets of data to create the final balanced data set.

3.4 PERFORMANCE MEASURES

A confusion matrix is a summary of classification issue prediction outcomes and a performance measuring technique. The four classification outcomes are true negative, false negative, true positive, and false positive. It can be used to calculate a variety of popular metrics[44, 45]. The main diagonals (true negative, true positive) carry the most weight, indicating that they are correctly classified; anything off the main diagonal indicates an error. The goal of fraud detection, in general, is to maximize accurate classification while decreasing incorrect classifications. By reducing undetected fraud and false alarms, a high correct diagnostic probability can be obtained[46].

Table 3.5: Confusion matrix result for IRSF detection model

		Classified as	
		Positive(Fraud)	**Negative(Non-fraud)**
Actual class	Positive(Fraud)	True Positive(TP): Correctly classified as fraud	False Negative(FN): Incorrectly flagged as normal
	Negative(Non-fraud)	False Positive(FP): Incorrectly flagged as fraud	True Negative(TN): Correctly classified as normal

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.4}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.5}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.6}$$

$$F1 = \frac{2 * (Precision * Recall)}{Precision + Recall} \tag{3.7}$$

Building time and inference time: Speed refers to the computational cost of creating and using the classifier.

3.5 MODEL BUILDING AND VALIDATION TECHNIQUES

Following the completion of data preprocessing, feature selection, and aggregation, the next step is to train and build a classification model using the algorithm chosen. Random forest, artificial neural network, and support vector machine are the three classification learning methods that are

used to create the models. In chapter 3, the learning classification technique was discussed in depth.

The classifiers are tested using two different training modes: K-fold cross-validation and percentage split (66%/34%), which are experimentally recommended and result in a model calculated with low bias and a modest variance[47]. As a result, data collection takes one hour, and in order to check what difference will be seen in one hour's data set, both training modes are used. The K-fold cross-validation method divides the original data set into K subsets and iterates on each subset for K times. (K-1) subsets are utilized as training sets, and a single subset is used as a validation set in each iteration. The set for each of the K subsets is used once in the process, and the results are averaged over K iterations to produce the final result. 10-fold cross-validation is a popular recommendation, and it is used in this research. The use of k-fold cross-validation has the advantage of ensuring that each record appears exactly once in the test set. Nevertheless, it has the disadvantage of making the required validation process more complicated.

The entire data set is divided into a learning set for training and a test set for testing the classifiers in the percent split method(66%) of the data is utilize as the training set and the remaining is used as the test in this study, and a total of six models are developed to detect IRSF fraud. The number of building models that can be created by combining all three is shown in Table 3.6. With this, each model building is described in detail during the experiment.

Table 3.6: Total of built models in each classifier algorithm

Classifier	**Two Training Mode**	**Performance Metrics**	**Total Models**
RF, MLP, and SVM	Cross-validation	Accuracy, f-measure, time, and receiver operating characteristic curve	Six
	Percentage split		

3.6 TECHNIQUE TOOLS

During the classification process, popular data mining tools, specifically weka, are used to assess the effectiveness of classifiers[36]. It is free data-mining software that uses different techniques like classification, clustering, association, and data visualization. The weka-3-8-5 version and its experimental user interface are utilized to classify the three algorithms in this research.

4

RESULT AND DISCUSSION

In the following sections, the model-building process used by the proposed algorithms and the discussion of the results is presented in this chapter. Developed models with the random forest model are presented in detail in 4.1.1, models with artificial neural networks in 4.1.2, and models with the support vector machine algorithm in 4.1.3. For each section, some of the best classification models were chosen, and specific performance measures were presented.

4.1 MODEL BUILDING

To build a fraud detection model, this study analyzes the results of three data mining algorithms with international voice usage. After the experimentation setup was completed, we used three data mining techniques on the aggregated data set to generate all the models in hourly manner. Classifiers include support vector machines, random forests, and multilayer perceptron. Accuracy, f-measure, time evaluation (build time and inference time), and the receiver operating characteristic curve are used as performance evaluation metrics, and as well as two training modes: cross-validation and percentage split.

4.1.1 *Random Forests Model Building*

For classification tasks, random forests are a form of ensemble learning technique. They are made up of decision tree ensembles, with the final classification determined by a vote among the trees. Table 4.1, displays the detailed performance data for the models, which were all built using 10-fold cross-validation and percentage split to test the parameters during the model building process.

The greatest classification accuracy in this experiment was 97.0%, which was achieved using the cross-validation training mode. The accuracy, f-measure, and ROC curve performance of both training mode were also almost similar. The time it takes to build and evaluate using the percentage training mode is less than the time it takes to evaluate a model developed using the 10-fold cross-validation method, which is 1.51 and 48.76 seconds respectively.

Table 4.1: The outcomes of random forest performance metrics

Algorithm	**Training Mode:**	**Accuracy**	**F-Measure**	**ROC Area**	**Build Time**	**Inference Time**
RF	10-fold	97.0%	97.0%	99.1%	319	48.76
	%Split	96.8%	96.8%	99.0%	47.69	1.51

From a confusion matrix for a random forest classifier that compares expected and actual results in Table 4.2. There are 44093 true negatives and 44913 true positives in the 10-fold cross-validation training mode, indicating that the model is correct for 89006 out of 91772 samples. In other words, the overall error rate is 3.0% or 2766 out of 91772 samples. There are 14837 true negatives and 15358 true positives when implementing a percentage split mode on a data set, meaning that the model is correct for 30195 of the 31202 observations. To put it another way, the overall error rate is 1007 out of 31202, or 3.2%.

In general, the developed model that included both validation modes was shown to be preferable in terms of the overall evaluation. Detailed run information is depicted as Section A.3.

Table 4.2: Confusion matrix for random forest classifier results

Classifier	Training Mode	Correctly Classified Instances		Incorrectly Classified Instances	
		TP	TN	FP	FN
RF	10-Fold Cross-validation	44913	44093	1793	973
	Percentage Split	15358	14837	638	369

4.1.2 *MultiLayer Perceptrons Model Building*

The experiment's highest classification accuracy was 92.9%, which was achieved using the percentage split training model. The f-measure, time to build, inference time, and is also outperformed as demonstrated in Table 4.3. On the contrary, ROC takes a little less than the cross-validation training mode.

Table 4.3: Results of performance metrics for multilayer neural networks

Algorithm	Training Mode	Accuracy	F-Measure	ROC Area	Build Time	Inference Time
MLP	10-fold	92.8%	92.8%	96.6%	256	52.71
	%Split	92.9%	92.9%	96.2%	49.21	0.15

In Table 4.4, the confusion matrix displays the anticipated actual results. For the multilayer perceptrons with 10-fold cross-validation, there are 43766 true positives and 41382 true negatives, so the model is correct for 85148 observations out of 91772. That is, the overall error rate is 6624 out of 91772, or 7.2%. While applying the percentage split, there are 15022 true positives and 13958 true negatives, so the model is correct for 28980 observations out of 31202. That is, the overall error rate is 2222 out of 31202, or 7.1%

Even so, in terms of the overall evaluation, the percentage split training mode outperforms. Both training modes are consistent and have almost similar results in terms of classifying the provided problem.

Table 4.4: Confusion matrix for multilayer perceptrons classifier results

Classifier	Training Mode	Correctly Classified Instances		Incorrectly Classified Instances	
		TP	TN	FP	FN
MLP	10-Fold Cross-validation	43766	41382	4504	2120
	Percentage Split	15022	13958	1517	705

4.1.3 *Support Vector Machines Model Building*

The experiment's highest classification accuracy was 92.9%, which was achieved using the 10-fold cross-validation method. In addition, this model also outperforms the percentage split training model in terms of f-measure and ROC performance, as shown in Table 4.5.

Table 4.5: Results of performance metrics for support vector machines

Algorithm	training mode	Accuracy	F-Measure	ROC Area	Build Time	Inference Time
SVM	10-fold	92.9%	92.9%	92.9%	11829	1245.43
	%Split	92.7%	92.7%	92.7%	1166.5	514.52

The performance evaluation measurement, classified, and actual results are all shown in the confusion matrix Table 4.6. For the support vector machine applied to the data set cross-validation training mode, the diagonal entries are 43186 true positives and 42105 true negatives, so the model is correct for 85291 observations out of 91772. That is, the overall error rate is 6481 out of 91772, or 7.1%. While applied to the data set percentage split, the model is correct for 28927 out of 31202 observations, with 14705 true positives and

14222 true negatives. That means the overall error rate is 7.3 %, or 2275 out of 31202.

Table 4.6: Confusion matrix for support vector machine classifier results

Classifier	Training Mode	Correctly Classified Instances		Incorrectly Classified Instances	
		TP	TN	FP	FN
SVM	10-Fold Cross-validation	43186	42105	3781	2700
	Percentage Split	14705	14222	1253	1022

The 10-fold cross-validation model outperforms in terms of the overall evaluation. In the end, both training modes produce consistent and nearly equal results when it comes to classifying the given problem. But in the case of time evaluation, the 10-fold cross-validation method, for example, takes a long time to complete its classification process; the inference time is 1245.43 seconds. This is incompatible with our goal of having a fraud detection classifier that is as close to real-time as possible. In this case, the classifier is unreliable.

4.2 MODEL EVALUATION

In the 10-fold cross-validation method, as shown in Figure 4.1, each algorithm achieves a high level of accuracy, precision, recall, f-measure, and receiver operating characteristic area. A high-precision value means that all the algorithms have a low positive rate. Similarly, a high recall value indicates a low risk of false-negative rate. Random Forest scored accuracy, precision, recall, f-measure, and receiver operating characteristic curve higher than other two classifiers.

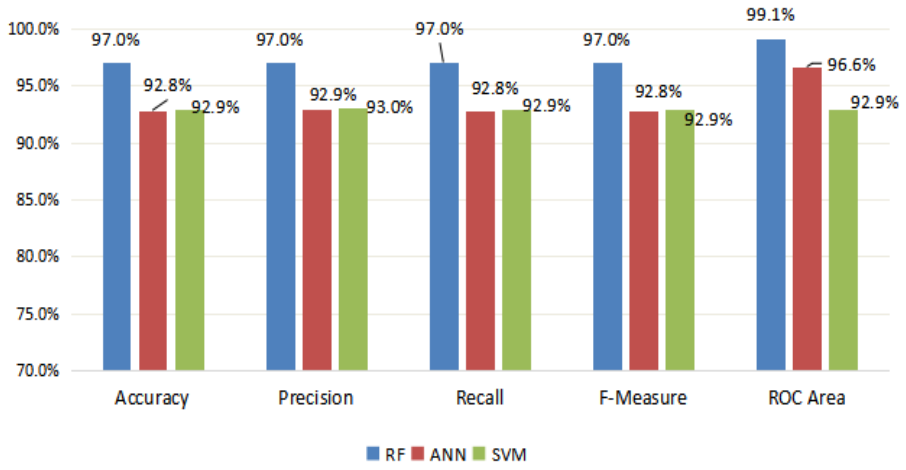

Figure 4.1: Model comparison using the 10-fold cross-validation method

All algorithms are tested using a percentage of split tests with the same performance metrics as stated in Figure 4.2. The highest score outcomes are 96.8%, and 99.00% when the RF classifier is employed with the percentage of split mode, evaluation metrics such as accuracy, f-measure, and receiver operating characteristic area. The RF outperforms SVM and MLP. This illustrates that the learning model taught with the RF algorithm correctly recognizes a low miss rate. On the contrary, the remaining two classifiers are consistent and produce similar results in both training modes, except for ROC and time to build as well as inference time.

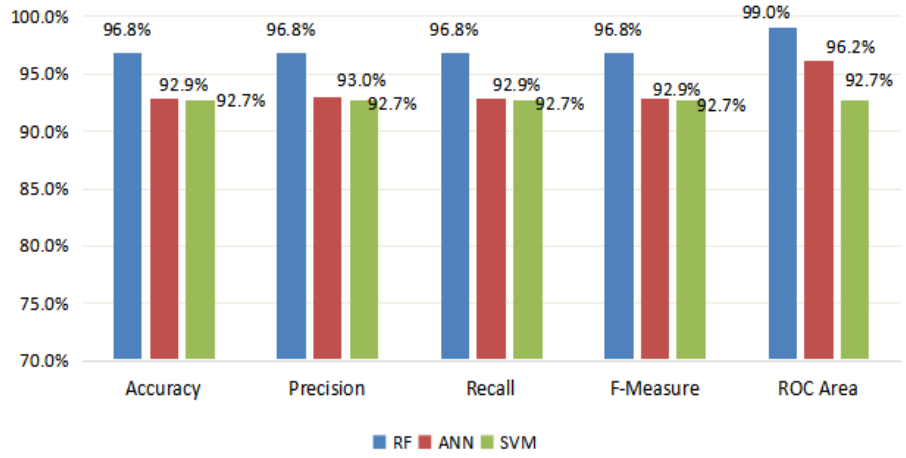

Figure 4.2: Model comparison using the percentage split method

Similarly, we evaluated the performance of all methods using the receiver operating characteristic curve in the same way in Figure 4.3. The ROC curve for the three classifiers shows the results of both training modes. The RF model exceeds the other two classifiers in both training modes. It covers a wider area of the false positive rate against the true positive rate curve.

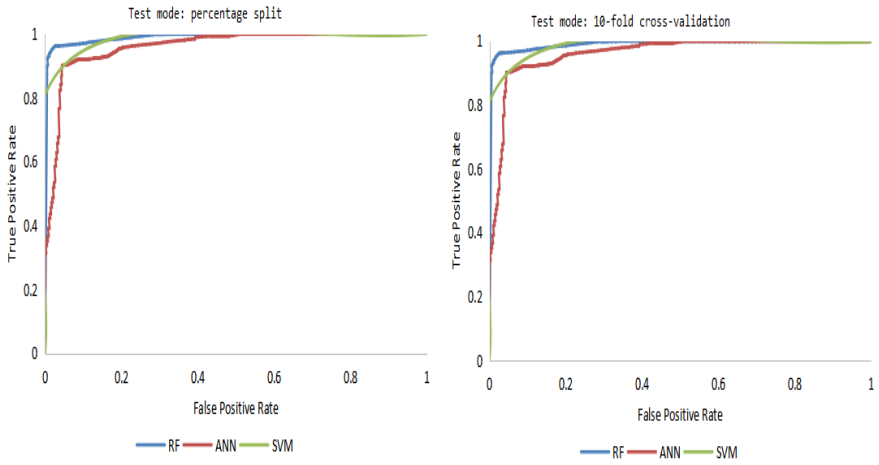

Figure 4.3: ROC Curve Comparison of three classifiers

In Figure 4.4, we evaluated the performance of all algorithms, using the same metrics such as accuracy, f-measure, and ROC score. In both test validation, the RF algorithm outperforms the MLP and SVM classifiers.

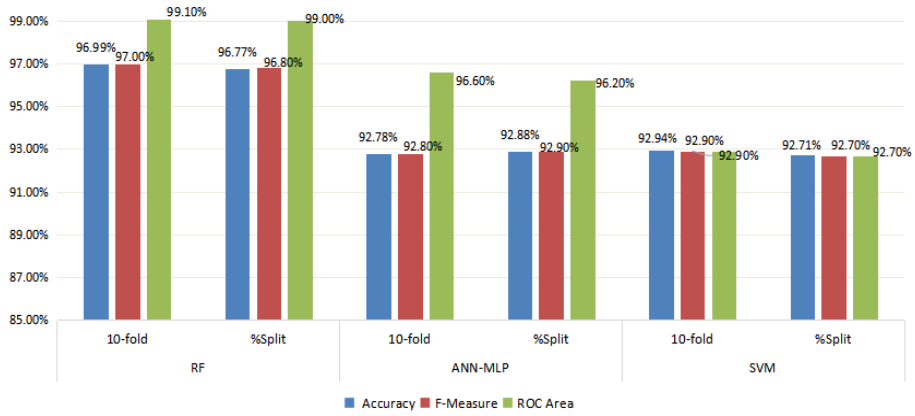

Figure 4.4: Model comparison performance result using both training modes

As compared to RF models, the MLP and SVM models generally achieve similar performance. On the other hand, the time to build as well as inference time, the performance of the SVM classifier algorithm model is high when compared to the other two classifiers.

Table 4.7: Summarized evaluation metrics of all the classifiers

Metrics	**Training mode: 10-Cross**			**Training mode: Percentage**		
	Classifiers			**Classifiers**		
	RF	**MLP**	**SVM**	**RF**	**MLP**	**SVM**
Accuracy	97.0%	92.8%	92.9%	96.8%	92.9%	92.71%
Precision	97.0%	92.9%	93.0%	96.8%	93.0%	92.70%
Recall	97.0%	92.8%	92.9%	96.8%	92.9%	92.70%
F-Measure	97.0%	92.8%	92.9%	96.8%	92.9%	92.70%
ROC Area	99.1%	96.6%	92.9%	99.0%	96.2%	92.70%
Build Time	319	256	11829	47.69	49.21	1166.50
Inference Time	48.76	52.71	1245.43	1.51	0.15	514.52

The random forest classifier, with both test validation, outperforms the other two classifiers in terms of accuracy, f-measure, building and evaluation time, and receiver operating characteristic area as demonstrated in Table 4.7. Similarly, the confusion matrix also shows in Table 4.8 that the RF model has a lower false positive (for example, in percentage split, 638) than the SVM and MLP classifiers, which are 1253 and 1517 respectively. Since fewer false positives (incorrectly classified) are preferred in fraud detection because it reduces the possibility of blocking normal users, and the same is true for false negatives (incorrectly failed to classify), they are undetected fraud, which are 369, 1022, and 705 respectively.

Instances that were correctly classified alongside instances that were incorrectly classified by all three classifiers are as follows.

Table 4.8: Instances classified as correctly and incorrectly

		Classifiers					
		RF		**MLP**		**SVM**	
Actual class		**Classified as**					
		F	**N**	**F**	**N**	**F**	**N**
Cross-Validation	F	44913	973	43766	2120	43186	2700
	N	1793	44093	4504	41382	3781	42105
Percentage Split	F	15358	369	15022	705	14705	1022
	N	638	14837	1517	13958	1253	14222

4.2.1 *Summary*

The scope of this study is to develop a method to classify IRSF frauds on an hourly basis, as well as to determine which algorithm is the most successful given the data set and attributes. For training and testing, we used random forest algorithms with 10-fold cross-validation and percentage of split techniques in the first experiment. The technique consists of numerous decision trees, and when developing each tree, it employs bagging and feature randomness to produce results through voting.

In the second experiment, two test methods are used, including cross-validation and percentage split, as well as a multilayer perceptron. The input layer, the output layer, and the hidden layer are the three layers that make up the system. The input layer receives the data to be processed, while the output layer handles task classification.

The support vector machine is used in the third experiment, along with the test method (cross-validation and percentage of split technique). Algorithms transform input data sets using a kernel trick methodology and then calculate an optimal boundary.

Three classifier algorithms namely, RF, SVM, and MLP, were evaluated using classification performance metrics such as accuracy, f-measure, model evaluation time (build and inference time), and receiver operating characteristic curve. The experimentation was done on a personal laptop with an Intel (R) i7-3540M CPU, 3.00GHz speed, 3001 MHz, 2 Core (s), 4 Logical Processor (s), 4 GB RAM, and a fixed hard disk. Weka, a data mining tool, was used for preprocessing and creating the models.

The CDR was collected for two months. Then, 172,956 normal call records and 24,718 fraudulent call records were selected, a fixed one-hour aggregation was performed, and before removing outliers and extremes, the international normal number to fraudulent number data collection ratio was 95.9% to 4.1%. We employed two sampling techniques such as cluster-based under-sampling and oversampling techniques to reduce bias between the two classes. As a result, the final data set composition was 50% to 50%.

For model building and testing, we used two training modes for each classifier. These training modes are 10-fold cross-validation and percent split. Besides, 66% of the data set was used to train the model, while the remaining 34% was utilized to test the model.

There were thirty fields in the original CDR. Furthermore, with information gain and correlation evaluation approach, nine specific (original and derived) features that are important to this work were chosen and used.

Selected data features such as called duration, distinct ratio calling, distinct ratio called, called times, out trunk, calling duration, average calling duration, average called duration, in trunk, and calling times were useful for

detecting and improving the efficiency of the fraud model. The remaining features, except for out_trunk, are sufficient for categorizing calls as either fraudulent or normal.

A combination of three different data mining techniques and two training validation techniques were used in the training and testing experiment. Using a fixed one-hour aggregation time, a total of six models was generated. The models of each method are then compared, and the best classifier is chosen based on performance metrics. The 10-fold cross-validation and percentage split data validation training modes were utilized in the experiment. Aside from that, default values for classifier parameter settings were used. The model's results are analyzed and compared.

The RF classifier approach with both validations outperforms the MLP and SVM models in terms of performance matrices. In terms of accuracy, f-measure, building and inference time, and receiver operating characteristic curve, the RF classifier surpasses the other two classifiers with both test validation. When we look at the precision of the RF model, we can see that the results are significantly better. The fewer false positives there are the better and the lower the risk of barring legitimate users. As a result, our goal is to maximize recall and precision while minimizing false positives and false negatives.

In terms of training mode, for each classifier, there is such a small difference between them in both percentage split test and 10- cross-validations that it is hard to distinguish which one is higher to the other. As a result, both training modes are consistent and have almost similar results in terms of classifying the provided problem.

As seen in Table 4.7, the RF and MLP models outperform the SVM in terms of time to build and inference time. The needed inference time classifications are 48.76 and 52.71 seconds respectively. The SVM classifier, on

the other hand, takes significantly longer to complete its classification (for example, 1245.43 seconds). One disadvantage is that we have low performance, so the data set is very large, and SVM since it takes more training time to classify. It is unreliable as a result of these results, which goes against our goal of having a fraud detection classifier that is as close to real-time as possible.

When we compare the results to our goals, we see that valuable features for international revenue sharing fraud detection are identified, with a period level of one fixed hour, which allows for near-real-time fraud detection in an offline data set. The performance of a fixed one-hour data granularity level is acceptable. Moreover, it detects with comparable accuracy to the state-of-the-art.

5

CONCLUSION AND FUTURE WORKS

5.1 CONCLUSION

International revenue sharing fraud costs the operators a lot of money and affects their reputation. Having a near-real-time detection model is helps to reduce revenue losses while also providing high value. The key purpose, from a business point of view, is to identify fraudulent calls to minimize revenue loss as much as possible. We presented fixed one-hour data set with a balanced ratio using three classifiers: random forest, multilayer perceptron, and support vector machine, based on the fraud features with two training modes, such as percentage split and 10-cross-validation. By combining these, six models are made. Based on the findings, the models with the best detection performance are chosen.

The study's major focus is to create an IRSF model that uses data mining techniques to detect fraud by utilizing data from the ethio telecom mediation system's international voice call records. The labeled data was also collected by the fraud management section. This study began with a review of the literature on telecom fraud detection, which is required to identify the extent and limitations of existing fraud detection systems.

Overall, the experimentation generates promising results for detecting IRSF fraud and offers opportunities for the field of fraud detection in the telecom sector. The random forest algorithm achieves 97.0% accuracy, whereas the other algorithms, both support vector machine and multilayer percep-

tron, achieve 92.9% accuracy. According to the evaluation results, the model with the random forest algorithm also outperforms the others in terms of f-measure, receiver operating characteristic curve, and time to develop and evaluate (inference time) the models in both training modes. Furthermore, it also identifies with comparable accuracy to the state-of-the-art.

We identified nine key features that differentiate fraudulent data from legitimate data sets in our study. The study's successful testing provided a comprehensive understanding of the fraud type's behavior. The random forest methods are more effective and outperformed all other models in detecting potential IRSF fraud promptly.

Due to the study's near-real-time nature, the evaluation time for the algorithms chosen is assessed as one key comparison. Support vector machines take much longer than the other two classifiers in both training mode circumstances. It takes more than twenty minutes (1245.43 seconds) and eight minutes (514.52 seconds) in the classification process. As a result, in our case, it is not a good choice for near-real-time analysis.

5.2 FUTURE WORKS

International voice data is collected and aggregated hourly in the proposed international revenue sharing fraud detection model. It can detect fraudulent calls in less than an hour if the data is collected in less than an hour, for example, half an hour, bringing it closer to real-time. Fee-based features are another possible area for future research. Although deployment is beyond the scope of this study, it can be proposed for future work.

BIBLIOGRAPHY

[1] Luis Cortesão et al. "Fraud management systems in telecommunications: a practical approach." In: *Proceeding of ICT*. 2005.

[2] Aisha Abdallah, Mohd Aizaini Maarof, and Anazida Zainal. "Fraud detection system: A survey." In: *Journal of Network and Computer Applications* 68 (2016), pp. 90–113.

[3] Standardization Sector and O Itu. "Series E: overall network operation telephone service service operation and human factors quality of telecommunication services: concepts models objectives and dependability planning-Use of quality of service objectives for planning of telecommunication networks." In: *ITU* 16 (2008), p. 17.

[4] Olivier FESTOR. "Understanding Telephony Fraud as an Essential Step to Better Fight It." PhD thesis. TELECOM ParisTech, 2017.

[5] Paul Carroll. "Investigation into VoIP Communications fraud and TDoS attacks and solutions required for the corporate environment." PhD thesis. University of Dublin, 2018.

[6] Global Telecom et al. "Communications Fraud Control Association Announces Results of 2019 Global Telecom Fraud Survey." In: (2019).

[7] M. Sahin et al. "SoK: Fraud in Telephony Networks." In: *2017 IEEE European Symposium on Security and Privacy (EuroS P)*. 2017, pp. 235–250. DOI: 10.1109/EuroSP.2017.40.

[8] Anne Kouam, Aline Carneiro Viana, and Alain Tchana. "SIMBox bypass frauds in cellular networks: Strategies, evolution, and detection survey." In: (2021).

[9] Merve Sahin and Aurélien Francillon. "Understanding and Detecting International Revenue Share Fraud." In: (2021).

[10] Appavu Siva Prakasam et al. "Increased Smart Device Penetration Brings Malware Vulnerability: Methods for Detecting Malware in a Large Cellular Network." In: ().

[11] Constantinos S Hilas. "Designing an expert system for fraud detection in private telecommunications networks." In: *Expert Systems with applications* 36.9 (2009), pp. 11559–11569.

[12] Rupesh K Gopal and Saroj K Meher. "A rule-based approach for anomaly detection in subscriber usage pattern." In: *International Journal of Mathematical, Physical and Engineering Sciences* 1.3 (2007).

[13] Nan Jiang et al. "Isolating and analyzing fraud activities in a large cellular network via voice call graph analysis." In: *Proceedings of the 10th international conference on Mobile systems, applications, and services*. 2012, pp. 253–266.

[14] Mais Arafat, Abdallah Qusef, and George Sammour. "Detection of wangiri telecommunication fraud using ensemble learning." In: *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*. IEEE. 2019, pp. 330–335.

[15] Aliye Mehadi. "Performance Evaluation of Unsuper- vised Learning Techniques for Enter- prise Toll Fraud Detection." MA thesis. 2018.

[16] Yoram J Meijaard et al. "Predictive Analytics to Prevent Voice over IP International Revenue Sharing Fraud." In: *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer. 2020, pp. 241–260.

[17] Merve Sahin and Sophia Antipolis. "IRSF : a Billion $ Fraud Abusing International Premium Rate Numbers." In: (), pp. 16–17.

[18] *Exploring fraud in telephony networks*. (Online;accessed: 02/15/2021). URL: https://media.ccc.de/v/35c3-9852-exploring_fraud_in_telephony_networks.

[19] *Ireland's Call , Careful Now, by Cathal McDaid*. [Online; accessed 03/09/2021]. 20th Oct 2017. URL: https://www.adaptivemobile.com/blog/irelands-call-careful-now.

[20] *PRISM Report on IPRN trend 2020: An Analysis of the Destinations Fraudsters Use in IRSF and Wangiri Attacks*. (Online; accessed 01/23/2021). URL: https://bswan.org/prism_report_on_iprn_trend_2020.asp.html.

[21] Jiawei Han, Micheline Kamber, and Jian Pei. "Data mining concepts and techniques third edition." In: *The Morgan Kaufmann Series in Data Management Systems* 5.4 (2011), pp. 83–124.

[22] Ian H Witten and Eibe Frank. "Data mining: practical machine learning tools and techniques with Java implementations." In: *Acm Sigmod Record* 31.1 (2002), pp. 76–77.

[23] Mousa Albashrawi. "Detecting financial fraud using data mining techniques: A decade review from 2004 to 2015." In: *Journal of Data Science* 14.3 (2016), pp. 553–569.

[24] H Ali Ata and Ibrahim H Seyrek. "The use of data mining techniques in detecting fraudulent financial statements: an application on manufacturing firms." In: *Suleyman Demirel University Journal of Faculty of Economics & Administrative Sciences* 14.2 (2009).

[25] Kamal Eldahshan and Hoda A Abdel Hafez. "Decision Problems Taxonomy and the Choice of Data Mining Techniques." In: 2.1 (2009).

[26] Shujun Huang et al. "Applications of support vector machine (SVM) learning in cancer genomics." In: *Cancer Genomics-Proteomics* 15.1 (2018), pp. 41–51.

[27] Sotiris B Kotsiantis, I Zaharakis, P Pintelas, et al. "Supervised machine learning: A review of classification techniques." In: *Emerging artificial intelligence applications in computer engineering* 160.1 (2007), pp. 3–24.

[28] Milad Malekipirbazari and Vural Aksakalli. "Risk assessment in social lending via random forests." In: *Expert Systems with Applications* 42.10 (2015), pp. 4621–4631.

[29] Lior Rokach. "Ensemble-based classifiers." In: *Artificial intelligence review* 33.1 (2010), pp. 1–39.

[30] Leo Breiman. "Random forests." In: *Machine learning* 45.1 (2001), pp. 5–32.

[31] Thanh-Tung Nguyen, Joshua Zhexue Huang, and Thuy Thi Nguyen. "Unbiased feature selection in learning random forests for high-dimensional data." In: *The Scientific World Journal* 2015 (2015).

[32] Hyontai Sug. "The effect of training set size for the performance of neural networks of classification." In: *WSEAS Trans Comput* 9 (2010), pp. 1297–306.

[33] Alaa F Sheta and Amneh Alamleh. "A professional comparison of c4. 5, MLP, SVM for network intrusion detection based feature analysis." In: *The International Congress for global Science and Technology*. Vol. 47. 2015, p. 15.

[34] Mohssen Mohammed, Muhammad Badruddin Khan, and Eihab Bashier Mohammed Bashier. *Machine learning: algorithms and applications*. Crc Press, 2016.

[35] Antonia Dåderman and Sara Rosander. *Evaluating frameworks for implementing machine learning in signal processing: A comparative study of CRISP-DM, semma and kdd*. 2018.

[36] Begüm Çığşar and Deniz Ünal. "Comparison of data mining classification algorithms determining the default risk." In: *Scientific Programming* 2019 (2019).

[37] Imelda T Coyne. "Sampling in qualitative research. Purposeful and theoretical sampling; merging or clear boundaries?" In: *Journal of advanced nursing* 26.3 (1997), pp. 623–630.

[38] Salvador Garc, Julin Luengo, and Francisco Herrera. *Data preproceesing in data mining*. Vol. 72. Springer, 2015.

[39] Zhiqiang Ge et al. "Data mining and analytics in the process industry: The role of machine learning." In: *Ieee Access* 5 (2017), pp. 20590–20616.

[40] Shahzad Ahmed et al. "Modern data formats for big bioinformatics data analytics." In: *arXiv preprint arXiv:1707.05364* (2017).

[41] Daniel T Larose and Chantal D Larose. *Discovering knowledge in data: an introduction to data mining*. Vol. 4. John Wiley & Sons, 2014.

[42] Pradeep Kumar et al. "Classification of Imbalanced Data: Review of Methods and Applications." In: *IOP Conference Series: Materials Science and Engineering*. Vol. 1099. 1. IOP Publishing. 2021, p. 012077.

[43] Astha Agrawal, Herna L Viktor, and Eric Paquet. "SCUT: Multi-class imbalanced data classification using SMOTE and cluster-based undersampling." In: *2015 7Th international joint conference on knowledge discovery, knowledge engineering and knowledge management (IC3k)*. Vol. 1. IEEE. 2015, pp. 226–234.

[44] Yufeng Kou et al. "Survey of fraud detection techniques." In: *IEEE International Conference on Networking, Sensing and Control, 2004*. Vol. 2. IEEE. 2004, pp. 749–754.

[45] Max Bramer. *Principles of Data Mining Third Edition*. 2016.

[46] Rekha Bhowmik. "Data mining techniques in fraud detection." In: *Journal of Digital Forensics, Security and Law* 3.2 (2008), p. 3.

[47] George A Marcoulides. *Discovering knowledge in data: an introduction to data mining*. 2005.

A

APPENDIX

A.1 CDR FIELDS DESCRIPTION

Field Name	Description
EVENT_INST_ID	CDR sequence number
RE_ID	Service type(voice, SMS, and data)
BILLING_NBR	Billing number
BILLING_IMSI	Billing side IMSI
CALLING_NBR	Calling number
CALLED_NBR	Called number
TRUNK_OUT	Outgoing truck id
DURATION	Call duration
STATE_DATE	Status time
STATE	Status
CELL_B	Called cell
LAC_A	Calling located area
CELL_A	Calling cell

LAC_B	Called located area
START_TIME	Call start time
THIRD_NBR	The third party number
CHARGE1	Charge amount
CHARGE2	Charge amount
CHARGE3	Charge amount
CHARGE4	Charge amount
PRICE_ID1	Rate ID1
PRICE_ID2	Rate ID2
PRICE_ID3	Rate ID3
PRICE_ID4	Rate ID4
ACCT_ITEM	Account item
CALLING_IMEI	Originating user IMEI
CALLED_IMEI	Terminating user IMEI
BYTE_UP	Uploaded traffic
BYTE_DOWN	Downloaded traffic
TRUNK_IN	Incoming truck id

Table A.1: Call detail record fields description

A.2 SELECTED ORIGINAL FIELDS DESCRIPTION

No.	Field Name	Description
1	CALLING_NUM	The number that initiated the call
2	CALLED_NUM	The number that receiving the call
3	RE_ID	CDR type id for voice, SMS and data
4	DURATION	Call duration in seconds
5	IN_TRUNK	Incoming trunk id
6	OUT_TRUNK	Outgoing trunk id

Table A.2: Selected original fields description

A.3 RUN INFORMATION FOR RF SAMPLE TRAINING

=== Run information ===

Scheme: weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Relation: DataSet

Instances: 91772

Attributes: 10

CALLING_TIMES, CALLING_DURATION, DIST_RATIO_CALLING

IN_TRUNK, AVG_CALLING_DU, CALLED_TIMES,

CALLED_DURATION, DIST_RATIO_CALLED, AVG_CALLED_DU

CLASS

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 48.76 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 89006 96.986%

Incorrectly Classified Instances 2766 3.014%

Kappa statistic 0.9397

Mean absolute error 0.0482

Root mean squared error 0.1593

Relative absolute error 9.6309 %

Root relative squared error 31.8512 %

Total Number of Instances 91772

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.961 0.021 0.978 0.961 0.970 0.940 0.991 0.990 N

0.979 0.039 0.962 0.979 0.970 0.940 0.991 0.991 F

Weighted Avg. 0.970 0.030 0.970 0.970 0.970 0.940 0.991 0.990

=== Confusion Matrix ===

a b <– classified as

44093 1793 | a = N

973 44913 | b = F

A.4 IEEE CONFERENCE PAPER

International Revenue Sharing Fraud (IRSF) Detection Using Data Mining Techniques: The Case of ethio telecom

1st Alamarew Digafe
School of Electrical and Computer Engineering
Addis Ababa Institute of Technology(AAiT)
Addis Ababa University
Addis Ababa, Ethiopia
adtsim16@gmail.com

2nd Dr. Ephrem Teshale
School of Electrical and Computer Engineering
Addis Ababa Institute of Technology(AAiT)
Addis Ababa University
Addis Ababa, Ethiopia
ephrem.teshale@aait.edu.et

***Abstract*—Telecommunications operators lose a large part of their annual revenue due to international revenue sharing fraud. International revenue sharing fraudsters generate traffic for their services without paying the originating network for the calls. Existing detection methods include monitoring call patterns and blocking international high-risk range list numbers. However, these approaches have issues that make them ineffective, as they are frequently bypassed by fraudsters. This research aims to develop an international revenue sharing fraud model that uses classification algorithms to detect fraudulent traffic generation schemes from missed calls. To do so, a data set is collected and aggregated over one hour, data preprocessing is carried out, and key attributes are identified. The models are also trained in two separate training modes: percentage split and 10-cross-validation, utilizing the Waikato Environment for Knowledge Analysis machine training tool. Random forest techniques outperform artificial neural networks and support vector machines in terms of accuracy, f-measure, receiver operating characteristic, and time required to train a model, as well as inference time, in tackling the classification problem. It also achieved comparable accuracy to the state-of-the-art. The support vector machine and neural network classification algorithms are found to be the next best after the random forest in overall performance metrics. On the contrary, in both training modes, support vector machine techniques exceed the acceptable detection delay in the classification process.**

***Index Terms*—International premium rate number, Fraud detection, Artificial neural network, Support vector machine, Random forest, International revenue sharing fraud, Classification.**

I. Introduction

In the telecommunications business, international revenue sharing fraud is the most common revenue diversion fraud. The carrier's business interconnection agreement, known as carrier trust, is bypassed by such fraudulent activities[5]. High destination termination rates provide a strong motivation for fraudsters to commit fraud. Many telecom operators have serious concerns about number-resource misuse, a practice whereby calls never reach the destination designated by the international country code. Such fraudulent activity is carried out as a consequence of such misuse. They are terminated early through a carrier and content provider agreement for revenue-generating content services [1, 17].

One of the most common types of interconnect fraud involves calls to certain destinations being hijacked by fraudulent operators and diverted to the so-called international premium rate services. The fraudster partners with a local operator that charges high rates for call termination and has an agreement to share the revenue generated by the fraudster [6]. International revenue sharing fraud is one of the five main types of fraud and the highest fraud loss contributor, as per the communications fraud control association (CFCA) global fraud loss survey, 2019. Fraudsters take advantage of the telecom operator's infrastructure to fraudulently increase traffic to high-risk foreign destinations[4, 18].

Fraudsters are increasingly exploiting the data to generate traffic calls without paying for them. International revenue sharing fraud typically includes various aspects (for example, a fraudulent carrier in collaboration with a premium rate service provider) that collect and share the revenue from calls[6]. An international call initiated from a cellular or IP network travels over transit operators before reaching its destination/termination country. Each of these transit operators gets a share of the call revenue for passing over the call, and the local operator in the terminating country receives a termination fee [11].

The interconnection business is mostly for international and long-distance calls. There are also exchange platforms where operators can buy and sell minutes directly or anonymously [6, 9]. Fraudsters may use a variety of methods to generate call traffic, including fraudulently obtained SIM cards, hacking private branch exchanges, callback schemes (missed calls), and mobile malware (malicious mobile applications).

International revenue shared fraud is a critical and growing issue in the telecommunications industry. To avoid financial losses, an effective near-real-time fraud detection system should be developed[6, 16]. Data mining techniques have become increasingly popular in telecommunications fraud and other fields. International revenue share fraud can occur on

both local (home) and roaming networks. This study will focus on local networks' international voice calls as well as a missed call fraud generation scheme. We used ethio telecom call detail record data as a case study.

The goal of this research is to find an answer to the following questions: what data features or attributes can be used to identify international revenue sharing fraud and which classification technique is more appropriate for detecting potential international revenue sharing fraud quickly.

The rest of the contents are organized in the following manner. Section II discusses related works, Section III describes the experimental design in this study, Section IV analyzes the classifiers and shows which produced the best results, Section V summarizes our findings, and Section VI discusses future research possibilities.

II. Literature Review

Several scientific research works and studies on detecting telecom fraud related to international revenue sharing fraud detection using data mining techniques have been published. The below literature is closely related to the title of the study.

The authors[15], provide a framework for better understanding and assessing telecom fraud. The goal is to clarify inconsistencies in current fraud terminology and increase awareness. A fraud taxonomy was used to identify root causes, vulnerabilities, exploitation methods, fraud types, and, finally, how fraught benefits fraudsters. In [6] researchers proposed an international revenue sharing fraud detection feature. They analyze internet international premium rate service resellers and their test portals, which are frequently utilized by fraudsters. The naive algorithms can detect fraudulent calls with a 98% accuracy rate and a 0.28% false-positive rate. In [8], the authors developed a rule-based approach to detect fraudulent subscriber usage. Sampled estimations from historical (study) and current (test) call detail record data were collected. The study period holds call records of customers' non-anomalous behavior. Any deviation beyond a threshold is used to raise an alarm. The authors[10], used a clustering-based approach to detect fraud activities on voice calls. They looked at 6-month international voice calls collected in the UMTS network's MSCs by large mobile operators. The approach is successful in detecting the destination numbers for 78% of the international revenue sharing fraud calls in the dataset. The authors' aims [14], to identify international revenue sharing fraud activities on large-scale cellular networks. The model was evaluated and tested on a one-year dataset. Apriori mining is used to identify correlated foreign phone numbers and then associate them with billing information. The authors in [3] utilized ensemble learning data mining algorithms, such as random forests, AdaBoost, and boosting methods,their detection. The extreme gradient boosting algorithm was the best among the other two algorithms for detecting fraudulent calls from unknown subscribers. Researchers have collected data on 517,319 unique international premium rate service test numbers from ten websites advertising different international premium rate numbers as well as commercial numbering plans. They found a wide variety of number ranges that could potentially be used for both fixed and mobile numbers by the international revenue sharing fraud numbers [6]. The researcher[12] aim to detect toll fraud committed through enterprise private branch exchange hacking. The k-means algorithm performs 97.96% in 0.09 seconds, which is higher than the Expectation-Maximization algorithm and rule-based. The methodology included 4998 private branch exchange users by analyzing data traffic, and additional manually generated synthetic data was also used in the research. The authors aim [13], to discover fraudulent calls at the moment of their establishment, thereby preventing international revenue sharing fraud from happening. The model they used is an isolation forest, which is an unsupervised data mining approach. The result was a pre-call detection success rate of at least 45% up to 87% false-positive rate of 2% to 5% respectively. But using such an approach does not allow for enforcing which features should be considered. Beside, the model has high false-positive rates.

III. Experimental Design

Three algorithms' classification performances are compared and evaluated in this study. The performance of classifiers is evaluated using 10-fold cross-validation and percentage split data. The accuracy, f-measure, time evaluation (built and inference time), and receiver operating characteristic curve were used to evaluate the detection performance of classification algorithms. Experimentation was carried out using the cross-industry standard process for the data mining process model. All raw data was collected for two months, after which it was preprocessed,, aggregated for a fixed one-hour, and relevant features were selected and the performance of the classification algorithm assessed.

A. Understanding the Data

Data-mining techniques, as well as the relationships between the data and the problem at hand, were evaluated. The total collected call detail record contains a total of thirty fields. The data set contains information related to international calls, such as the number initiating a call, receiving the call, time of event initiation, and duration of the call in seconds.

B. Data Preparation

The goal of this study is to use fraudulent behavior generation schemes of missed calls to detect international revenue sharing fraud. The raw call detail records with thirty attributes were selected at random. The data set was generated from normal international 92,726 and 3,921 fraud numbers. The behavior of international revenue sharing fraud was used as an input to the data selection process. For two months, international voice data from mediation was collected. Irrelevant data from the collected data was removed for the data mining algorithms to learn relevant information. The performance of classification algorithms is directly determined by the quality of the selected input data. The data set was collected for two

TABLE I
INCOMING AND OUTGOING INTERNATIONAL VOICE SAMPLE DATASET

CDR Type	Number	Percentage (Number)	Record	Percentage (Record)
Normal	92,726	95.9%	172,956	87%
Fraud	3,921	4.1%	24,718	13%
Entire data	96,647	100%	197,674	100%

months. 172,956 normal call records and 24,718 fraudulent call records were selected.

The call detail record data was used to develop a classification model to detect international revenue sharing fraud. The key data for each user was created as a column and row data set. Data preprocessing includes tasks like data cleaning, aggregation, and formatting, as well as feature selection.

1) Data Cleaning: The practice of identifying and correcting (or removing) incorrect records from a data set is known as data cleaning [7]. The data for this study contains errors such as incomplete values (a combination of hexadecimal and decimal integers), missing values, duplicate records, and so on. Similarly, records having any error values in any of their fields are not included in the target dataset.

2) Data Aggregation: The volume and quality of data used determine the data analysis observations obtained. It is essential to collect large amounts of high-quality data in order to obtain relevant results. The method of collecting data and presenting it in a summary format is referred to as data aggregation. Based on international voice calls, we selected relevant attributes for aggregation. The aggregated features of voice call usage at a single international number level is depicted in Table II

TABLE II
DESCRIPTION OF SELECTED ATTRIBUTES

Field Name	Description
Calling_times	Frequency of calling numbers
Called_times	Frequency of called numbers
Calling_duration	Time take to conduct outgoing call
Called_duration	Time take to conduct incoming call
Avg_calling_du	Average calling duration
Avg_called_du	Average called duration
In_trunk	Incoming number trunk id total count
Dist_ratio_calling	The ratio of distinct calling over total calling
Dist_ratio_called	The ratio of distinct called over total called
Out_trunk	Outgoing number trunk id total count
Class	Normal or Fraud

3) Feature Selection: Various attributes are collected throughout the data collection procedure, even if they are not important. The most significant part of feature selection is deciding which best distinguishes between classes. The data mining learning method employs a variety of algorithms.

The ranking technique applies weights to the given features based on the model's evaluation criteria and techniques. For cross-checking and validation purposes, the attributes are ranked based on the results of the correlation attribute and information gain selection processes. We discovered that one feature out of ten, out_trunk, no information given against classes, plays a minor role in detecting international revenue sharing fraud based on the results of the evaluation methods. As a result, the target data set's feature was removed.

4) Data Set Formatting: The data format is important for understanding data, representation of data, storage space requirements, and processing time. Different data mining techniques require different types and formats of input data. The attribute-relation file format contains an ASCII coding scheme that is used in our case.

5) Outliers Detection and Removal: An outlier is a variable that emerges at the extremes of its range. It is difficult to determine whether an observation is an outlier. Its detection techniques are mostly based on distance, density, or distribution. Distance-based techniques are popular, and in this research, interquartile rages are used to remove outliers from the data set.

C. Random Resampling Imbalanced Dataset

Imbalanced training data has a major negative impact on performance. Various re-sampling techniques are available to achieve a balanced data set. In this research, data sets are available for the normal (negative) class of 95.9% and the fraud (positive) class of 4.1%. So, both techniques, such as the Synthetic Minority Over-sampling Technique (SMOTE) and the clustered-based under-sampling technique, are used in this study due to the limited number of minority classes as a result of finding the correct new balance data set (50:50) ratio for both classes[2]. Under sampling is used to get a number of instances equal to the mean for a negative class with a number of instances greater than the mean, while oversampling is used to get a number of instances equal to the mean for a positive class with a number of instances less than the mean. Finally, as a result, the number of instances in both classes equals the mean, which is 45886, for a total of 91,772 records.

D. Performance Assessment Measures

A confusion matrix is a summary of the classification issue prediction outcomes and a performance measuring technique. It can be used to calculate a variety of popular metrics. The main diagonals (true negative, true positive) carry the most weight, indicating that they are correctly classified.

The following metrics are used to evaluate the international revenue sharing fraud detection models performance: accuracy, f-measure, time, and ROC.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (2)$$

Training time and inference time: Speed refers to the computational cost of creating (building time) and using the classifier (inference time).

E. Validation Techniques

After data preprocessing, feature selection, and aggregation, the next step is to build a classification model. Random forest, artificial neural network, and support vector machine are the three classification learning methods used. The classifiers were tested using two different training modes: K-fold cross-validation and percentage split. As a result, data collection takes one hour, and in order to check what difference will be seen in one hour's data set, both training modes are used. And 10-fold cross-validation and percentage split (66%/34%) are popular recommendations, and they were used in this study.

IV. Results and Discussion

The classification models used in this study were developed using random forest, artificial neural networks, multilayer perceptrons, and support vector machine algorithms. For each section of the study, some of the best classification models are chosen, and performance measures are presented. The discussion of the results is presented in the following sections.

For training and testing, we used random forest algorithms with 10-fold cross-validation and percentage of split techniques. The classifier consists of numerous decision trees, and when developing each tree, it employs bagging and feature randomness to produce results through voting. The support vector machine algorithm is used in the second experiment, along with both test methods. Support vector machine algorithms for higher dimensional transform input data sets using a kernel trick methodology and then calculate an optimal boundary. In the third, the input layer, the output layer, and the hidden layer are the three layers that make up the system. The input layer receives the data to be processed, while the output layer handles the task classification. The detailed classifier results are shown in Table III below.

TABLE III
SUMMARIZED EVALUATION METRICS OF ALL THE CLASSIFIERS

Metrics	Cross-Validation			Percentage Split		
	RF	MLP	SVM	RF	MLP	SVM
Accuracy	97.0%	92.8%	92.9%	96.8%	92.9%	92.71%
Precision	97.0%	92.9%	93.0%	96.8%	93.0%	92.70%
Recall	97.0%	92.8%	92.9%	96.8%	92.9%	92.70%
F-measure	97.0%	92.8%	92.9%	96.8%	92.9%	92.70%
ROC	99.1%	96.6%	92.9%	99.0%	96.2%	92.70%
Train time	319	256	11829	47.69	49.21	1166
Test time	48.76	52.7	1245	1.51	0.15	514.52

A combination of three different data mining techniques and two training validation techniques were used in the training and testing experiment. Using a fixed one-hour aggregation time, a total of six models were generated. The models of each method are then compared, and the best classifier is chosen based on performance metrics.

The 10-fold cross-validation and percentage split data validation training modes were utilized in the experiment. As a result, data collection takes one hour, and in order to check what difference will be seen in one hour's data set, both training modes are used. Aside from that, default values for classifier parameter settings were used. The model's results are analyzed and compared. In both training modes, each algorithm achieves a high level of accuracy, precision, recall, f-measure, and receiver operating characteristic values. A high precision value means that all the algorithms have a low positive rate. Similarly, a high recall value indicates a low risk of false-negative rate.

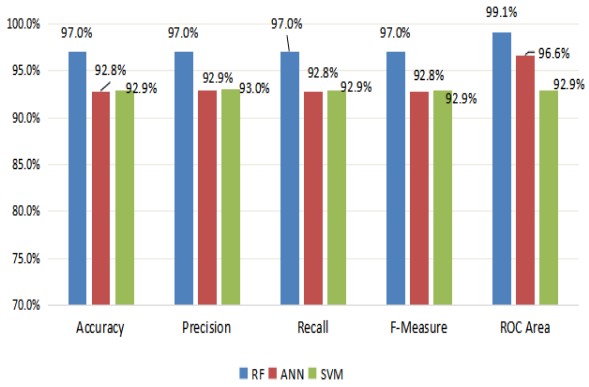

Fig. 1. Model comparison using the 10-fold cross-validation method

All algorithms were tested using both training modes with the same performance metrics as stated in Fig.1 and Fig.2, the random forest classifier approach outperforms the multilayer perceptron and support vector machine models in terms of performance metrics. On the contrary, the remaining two classifiers are consistent and produce almost similar results in both training mode techniques. Similarly, we evaluated the performance of all methods using the receiver operating characteristic curve in the same way as shown in Fig. 3. The random forest model exceeds the other two classifiers, and it covers a wider area of the false positive rate against the true positive rate curve.

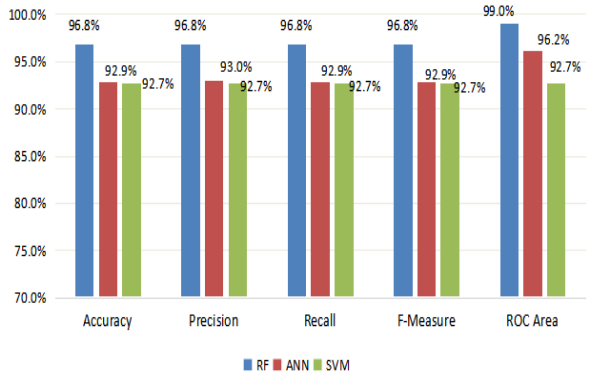

Fig. 2. Model comparison using the percentage split method

We evaluated the performance of all algorithms using the same metrics such as accuracy, f-measure, and receiver operating characteristic curve. In both test validations, the random

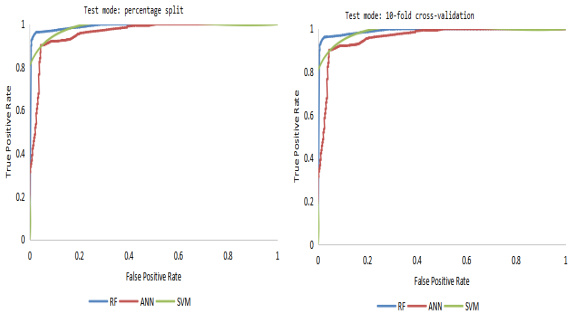

Fig. 3. The ROC curve for the three classifiers show the results of both training modes

forest algorithm outperforms the multilayer perceptron and support vector machine classifiers as shown in Fig. 4

As compared to random forest models, the multilayer perceptron and support vector machine models generally achieve similar performance. On the other hand, the time to build as well as the inference time, the performance of the support vector machine classifier algorithm model is high when compared to the other two classifiers.

In terms of training mode, for each classifier, there is such a small difference between them in both percentage split test and 10-cross-validations that it is difficult to tell which one is superior to the other. As a result, both training modes are consistent and have almost similar results in terms of classifying the provided problem.

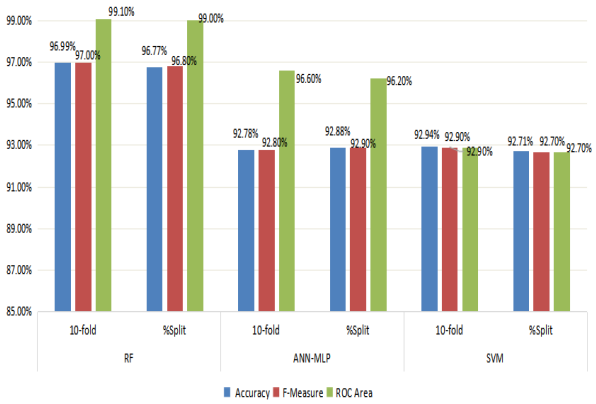

Fig. 4. Model comparison using both training modes

In terms of time to build and inference time, the random forest and multilayer perceptron models outperform the support vector machine classifier, as seen in Table III. The needed inference time classification is 48.76 and 52.71 seconds respectively. The support vector machine classifier, on the other hand, takes significantly longer to complete its classification (for example, 1245 seconds). It is unreliable as a result of these results, which goes against our goal of having a fraud detection classifier that is as close to real-time as possible.

V. Conclusion

International revenue share fraud costs the operators a lot of money and affects their reputation. It has also had an impact on ethio telecom. Having a near-real-time detection model for international revenue sharing fraud helps to reduce these losses while also providing high value. From a business point of view, the key purpose is to identify those fraudulent calls and minimize revenue loss as much as possible. We presented a fixed one-hour data set with a balanced ratio using three classifiers: Random Forest, multilayer perceptron, and support vector machine, based on the fraud features with two training modes, such as percentage split and 10-cross-validation. By combining these, six models were made. Based on the findings, the models with the best detection performance were chosen.

The primary purpose of the research is to develop an international revenue sharing fraud model that employs data mining techniques to detect fraud using data from the ethio telecom system's international voice call records. Labeled data was also collected by the fraud management section. This research began with a survey of the literature on IRSF fraud detection, which was necessary to determine the scope and limitations of existing fraud detection systems.

Overall, the experimentation generates promising results for detecting international revenue sharing fraud and offers opportunities for the field of fraud detection in the telecom sector. The random forest algorithm achieves 97.0% accuracy, whereas the other algorithms, both support vector machine and multilayer perceptron, achieve 92.9% accuracy. According to the evaluation results, the model with the random forest classification algorithm also outperforms the others in terms of f-measure, receiver operating characteristic curve, and time to develop and evaluate (inference time) the models in both training modes.

In our study, we identified nine key features that differentiate international revenue sharing fraudulent data sets from legitimate data sets. The study's successful testing provided a comprehensive understanding of the fraud type's behavior. The random forest classification methods are more effective in detecting potential IRSF fraud promptly.

Due to the study's near-real-time nature, the evaluation time for the algorithms selected is seen as an important comparison. In both cases (building and evaluating), the support vector machine takes significantly longer than the other two techniques. It takes more than twenty minutes (1245.43 seconds) and eight minutes (514.52 seconds) in the classification process. One disadvantage is that SVM has poor performance because of the large size of the data set. As a result, a support vector machine is not a good choice for near-real-time analysis in our case.

When we compare the results to our goals, we see that valuable features for international revenue sharing fraud detection are identified, with a period level of one fixed hour, achieving comparable accuracy in the state-of-the-arts, which allows for

near-real-time fraud detection in offline data. The performance of a fixed one-hour data granularity level is acceptable.

VI. Future works

International voice data is collected and aggregated hourly in the proposed international revenue sharing fraud detection model. It can detect fraudulent calls in less than an hour if the data is collected in less than an hour, for example, half an hour, bringing it closer to real-time. Fee-based features are another promising area for future research. Although deployment is beyond the scope of this study, it can be proposed for future work.

References

[1] Aisha Abdallah, Mohd Aizaini Maarof, and Anazida Zainal. "Fraud detection system: A survey". In: *Journal of Network and Computer Applications* 68 (2016), pp. 90–113.

[2] Astha Agrawal, Herna L Viktor, and Eric Paquet. "SCUT: Multi-class imbalanced data classification using SMOTE and cluster-based undersampling". In: *2015 7Th international joint conference on knowledge discovery, knowledge engineering and knowledge management (IC3k)*. Vol. 1. IEEE. 2015, pp. 226–234.

[3] Mais Arafat, Abdallah Qusef, and George Sammour. "Detection of wangiri telecommunication fraud using ensemble learning". In: *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*. IEEE. 2019, pp. 330–335.

[4] Paul Carroll. "Investigation into VoIP Communications fraud and TDoS attacks and solutions required for the corporate environment". PhD thesis. University of Dublin, 2018.

[5] Luis Cortesão et al. "Fraud management systems in telecommunications: a practical approach". In: *Proceeding of ICT*. 2005.

[6] Olivier FESTOR. "Understanding Telephony Fraud as an Essential Step to Better Fight It". PhD thesis. TELECOM ParisTech, 2017.

[7] Zhiqiang Ge et al. "Data mining and analytics in the process industry: The role of machine learning". In: *Ieee Access* 5 (2017), pp. 20590–20616.

[8] Rupesh K Gopal and Saroj K Meher. "A rule-based approach for anomaly detection in subscriber usage pattern". In: *International Journal of Mathematical, Physical and Engineering Sciences* 1.3 (2007).

[9] Constantinos S Hilas. "Designing an expert system for fraud detection in private telecommunications networks". In: *Expert Systems with applications* 36.9 (2009), pp. 11559–11569.

[10] Nan Jiang et al. "Isolating and analyzing fraud activities in a large cellular network via voice call graph analysis". In: *Proceedings of the 10th international conference on Mobile systems, applications, and services*. 2012, pp. 253–266.

[11] Anne Kouam, Aline Carneiro Viana, and Alain Tchana. "SIMBox bypass frauds in cellular networks: Strategies, evolution, and detection survey". In: (2021).

[12] Aliye Mehadi. "Performance Evaluation of Unsupervised Learning Techniques for Enter- prise Toll Fraud Detection". MA thesis. 2018.

[13] Yoram J Meijaard et al. "Predictive Analytics to Prevent Voice over IP International Revenue Sharing Fraud". In: *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer. 2020, pp. 241–260.

[14] Appavu Siva Prakasam et al. "Increased Smart Device Penetration Brings Malware Vulnerability: Methods for Detecting Malware in a Large Cellular Network". In: ().

[15] M. Sahin et al. "SoK: Fraud in Telephony Networks". In: *2017 IEEE European Symposium on Security and Privacy (EuroS P)*. 2017, pp. 235–250. DOI: 10.1109/EuroSP.2017.40.

[16] Merve Sahin and Aurélien Francillon. "Understanding and Detecting International Revenue Share Fraud". In: (2021).

[17] Standardization Sector and O Itu. "Series E: overall network operation telephone service service operation and human factors quality of telecommunication services: concepts models objectives and dependability planning-Use of quality of service objectives for planning of telecommunication networks". In: *ITU* 16 (2008), p. 17.

[18] Global Telecom et al. "Communications Fraud Control Association Announces Results of 2019 Global Telecom Fraud Survey". In: (2019).