American Journal of Intelligent Systems 2022, 12(2): 43-50
DOI: 10.5923/j.ajis.20221202.01

# Smart Mobile Telecommunication Network Fraud Detection System Using Call Traffic Pattern Analysis and Artificial Neural Network

**John C. Daka*, Mayumbo Nyirenda**

Department of Computer Science, School of Natural Science, University of Zambia, Lusaka, Zambia

**Abstract** Fraud is a major challenge facing telecommunication industry. A huge amount of revenues is lost to fraudsters who have developed different techniques and strategies to defraud the Mobile Telecommunication service providers. Mobile Network Fraud Detection systems used in the telecom industry by Mobile Network Operators (MNO) in Zambia has not explored most out of the Artificial Neural network techniques, analysing the Call Detail Records (CDRs), Call Traffic Pattern Analysis and Machine Learning classification algorithms. This study presents a Sim Box fraud detection model that identifies fraud patterns, it uses Artificial Neural Networks as an enabling tool to classify calls as either fraudulent or legitimate based on the attributes of the call. Actual Call Detail Records (CDR) collected from a Telecommunication company in Zambia was used. Out of a total of 13,398 CDR records, a total of 7,006 unique call records were obtained after sampling using a technique known as sampling without replacement where each sample unit of the call activity had only one chance to be selected in the sample for the study. The data set contained call records from both fraudulent and legitimate callers. Feature selection was performed on the data in order to eliminate redundant variables and select the attributes that would best describe fraudulent behaviour. The data set was partitioned as follows: 70% of the data set was set aside for training, 15% for validation and the remaining 15% was the testing set. The implementation of the Artificial Neural Network was based on the Matlab Neural Network toolbox known as the Neural network pattern recognition tool using Matlab R2018a. The trained network achieved 100% classification performance on the test data set as a result of having a balanced data set for the test. The study established that Artificial Neural Network are a successful technology that can be applied in Sim Box fraud detection since it was able to detect abrupt changes in established calling patterns which may be as a consequence of fraud. The implementation of the fraud detection tool will be a big step towards detection and mitigation of Sim Box fraud for mobile telecommunication companies in Zambia.

**Keywords** Call Detail Record (CDR), Artificial Neural Network, Matlab R2018a, Fraud, Sim Box Fraud, etc.

## 1. Introduction

According to the cyber-telecom crime report 2019, "The annual cost of telecommunications subscription fraud is estimated by some to reach up to more than US$12 billion, while others foresee the actual losses to be far greater, estimating it to be between 3 percent and 10 percent of the operators' gross revenues" (Cyber-Telecom Crime Report 2019 p.6). Within Zambia, telecommunication companies have broaden ways of carrying out operations from getting into the banking sector (Fintech Industry through mobile Money) and manufacturing, into real estate and find themselves in an unfamiliar territory of unprecedented threats from consumers. In a society with increasing high cost of living, many consumers of the telecom service tend to disregard the traditional values and tend to be under pressure in committing fraud. PwC 2018 Global Economic Crime and Fraud Survey: Zambia Report estimates fraud Committed by the Consumer is the second most prevalent type of economic crime experienced by Zambian organisations after Asset Misappropriation at an incident rate of 39%, This is 10% more than the reported prevalence globally'' [1]. These figures are a clear indication that fraud is a major problem, which requires serious study by scholars to minimise illegal activities.

Telecommunications or telecom has been part of the evolution of modern society. Telecommunication is paramount in the operation of businesses and has become a major industry itself and with the onset of the Covid 19 pandemic and working from home culture, telecommunication has been one of the pivotal tool in this era and a very useful tool to accomplish communication thereby company goals at various aspect levels of the

* Corresponding author:
dakajohnc2@gmail.com (John C. Daka)
Received: May 14, 2022; Accepted: Jun. 10, 2022; Published: Jun. 23, 2022
Published online at http://journal.sapub.org/ajis

organisations. Deeply integrated even in day-to-day activities, it is an aspect of modern technology that is treated as a constant. However, the reality of its own threats and vulnerabilities exists. Given how critical telecom is, its threat landscape should be explored and understood as telecom technology continues to thrive.

Telecommunication fraud can be defined as the misuse of the telecom infrastructure; this includes voice as well as data networks. The fraudsters intentions could be to avoid the services charges that would be charged or reduce that charge to a minimal charge, thereby not fully been charged the actual cost of the service. The intention could also be deeper than that and the fraudster's aim might be to gain profit by misusing the network of the provider [1]. Losses due to fraud in telecom industry are highly significant. Even though telecommunication industry suffers major losses due to fraud there is no comprehensive published research on this area mainly due to lack of publicly available data to perform experiments on. The data to be used for the experiments contains confidential information of customers and in most cases law and enforcement authorities prohibit exposing the confidential information of customers [2]. On the other hand, any broad research published publicly about fraud detection methods will be utilized by fraudsters to evade from detection [1], [3].

Consequently, huge amount of revenues is lost to fraudsters who have developed different techniques and strategies to defraud the Mobile Telecommunication service providers. Telecommunication Company worldwide suffers from customers who use the provided services without paying. Even though this is a small percentage comparing to the Telecom Operators' revenue, it is still a significant loss. For any service provider to remain in the industry, the expected loss from the activities of these fraudsters should be highly minimized if not eliminated completely. But due to the nature of huge data and millions of subscribers involved, it becomes very difficult to detect this group of people [4].

For this purpose, there is a need for a Smart Mobile Telecommunication Network Fraud Detection System model that can capture both the present and past history of the subscribers and classify them accordingly with a very accurate deep learning algorithm and near real time processing system.

Toll Bypass fraud is the unlicensed insertion of traffic onto another carrier's network. In many countries and in Zambia in particular, toll bypass for international call termination is criminal fraud. This scenario requires that the fraudsters obtain network access which makes international calls appear to be cheaper, domestic calls, effectively "bypassing" the normal payment system for international calling. One common technique for perpetrating this Interconnect fraud is GSM Gateway fraud, or SIM Boxing.

Neural Networks are promising solutions to this type of problem as they can learn complex patterns and trends within a noisy data. Neural networks have particularly shown better performance results than other techniques in the domain of telecom fraud. Therefore, supervised learning method was applied using Multi-layer perceptron (MLP) as a classifier. The dataset that was used for this study was obtained from a real mobile communication network in Zambia and contains subscriber's/SIM cards that have been tested and approved by the operator to be Toll bypass fraud and Schemes conducted over the telephone cards as well as normal SIM cards.

## 2. Related Work

Today, telecommunication market all over the world is facing a severe loss of revenue due to fraudsters [5]. To overcome such business hazards and to retain the market, operators are forced to look for alternative ways of using artificial neural networks techniques and statistical tools to identify the cause in advance and to take immediate actions in response. This can be possible if the past history of the subscribers were analysed systematically. Fortunately, telecom industries generate and maintain a large volume of data such as Call detail data and Network data [6]. One reason for the non-utilization of this potential is the insufficient knowledge of the Artificial Neural Networks and algorithms to be used on such data.

Despite being a relatively new technology that has not fully matured, there are a number of industries such as telecoms, hospitals, schools, banks, insurance companies and retail store that are already using artificial neural networks to drive business insight on the customers and therefore recommend or improve the customers experience.

A general neural network model for estimating telecommunications network reliability, Studies on the design of communications networks, reliability has been defined in a number of ways. In this study, a probabilistic measure, all-terminal reliability, was considered (this is sometimes termed overall network reliability). All-terminal reliability is the probability that a set of operational edges provides communication paths between every pair of nodes. A communications network is typically modelled as a graph with nodes, and edges; nodes represent sites (computers), and edges represent communication links. Each node, and each edge has an associated probability of failure, and the reliability of the network is the probability that the network is operational. The definition of reliability thus depends on which components are operational. F. Altiparmak, et al proposed a new method, based on an artificial neural network (ANN), to estimate the reliability of networks with identical link reliability [7].

Y. Harkouss et al in the use of artificial neural networks in nonlinear microwave devices and circuits modelling: An Application to Telecommunication System Design. They investigated in detail possible application of neural networks to modelling of large-signal hard-nonlinear behaviour of power transistors for circuit design purpose, and modelling of nonlinear circuits such as power amplifiers for system design purpose. The problem of finding a good model is

discussed through solutions offered by neural networks, with particular interest in wavelet networks trained by BFGS algorithms [8].

To survive in the fierce competition of telecommunication industry and to retain the existing loyal customers, prediction of potential churn customer has become a crucial task for practitioners and academicians through predictive modelling techniques. The identification of loyal customers can be done through efficient predictive models. By allocation of dedicated resources to the retention of these customers would control the flow of dissatisfied consumers thinking to leave the company. Y. Khan et al proposed artificial neural network approach for prediction of customers intending to switch over to other operators. This model works on multiple attributes like demographic data, billing information and usage patterns from telecom companies' data set. In contrast with other prediction techniques, the results from Artificial Neural Networks (ANN) based approach can predict the telecom churn with accuracy of 79% in Pakistan. The results from artificial neural network are clearly indicating the churn factors, hence necessary steps can be taken to eliminate the reasons of churn [9]. Much work and research in fraud mitigation has not been around the use of ANN in analysing or finding a pattern in the CDRs due to the fact that CDRs are not readily made available as in most countries this is a violation against the privacy of the subscribers of the service, Hence Therefore, this study intends to study the set of attributes that can be used to detect Toll bypass fraud and Schemes conducted over the telephone fraud as these are the most prevailing type of telecommunication frauds in a developing country like Zambia.

# 3. Study Design

This research study used a quantitative based approach. Therefore, the choice of this research method depended upon the underlying philosophy of research, data which was numerical and presentation of results [10]. Our research study did not involve any participatory worldview or participants as the data used was secondary data, thus, no questionnaires, interviews or surveys were formulated. Quantitative based approach was chosen because it emphasizes objective measurements and numerical analysis of data collected by manipulating pre-existing data using computational techniques.

This project employed the CRISP-DM process model in the implementation of the fraud detection system particularly the Sim box type of fraud. The CRISP-DM model has been recommended as the best model by various data miners as it encourages best practices and offers organizations the structure needed to realize better, faster results from data mining [11].

The CRISP-DM methodology employs six steps as illustrated below. Each of the steps is described in more detail in the section:

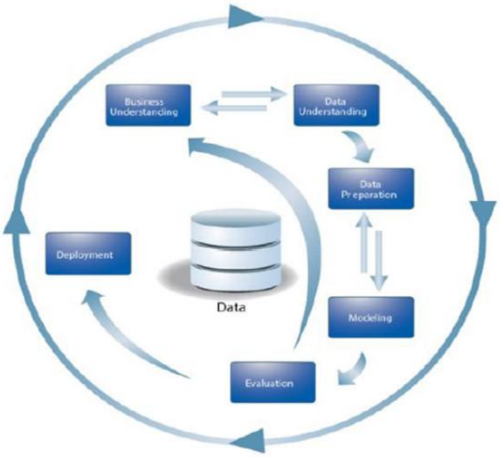


**Figure 3.1.** The CRISP-DM Process Model [11]

CRISP-DM as an approach to knowledge discovery in databases or data mining has been defined as a standard by the consortium of companies that formed it [12]. This standard contains detailed descriptions of all the tasks that are undertaken under each phase and the corresponding outputs from each phase task.

### 3.1. Data Sources and Collection Procedures

The sources of data used in this study is secondary data. The data collection procedure was conducted through direct data extractions from the database system at Telecom X and this did not involve any human subjects or participants. The identified data source was a database system at Telecom X and discussing the details of the data source system such as system name, version number and operating system platform is out of scope for this study. Historical call records data was extracted covering December 2020 and April 2021 for a specific site and sector under the experiment.

Successful data mining projects usually employ the use of large collections of data normally subject oriented, historical and time variant. The goal for successful data mining is always to ensure the data covers all possible phases of change in the subject being investigated. In order to meet this requirement, we aim to use data collected over a specific period of time containing all possible scenarios.

For this study, actual Call Detail Records (CDR) collected over five months from a telecommunication company in Zambia the provides telecommunication services to the Zambian local and foreigners customers. The five months' sampled randomly data provided a total of 13,398 CDR records for a particular site and sector that was under experiment for this project. CDR is a data record that contains information related to a telephone call, such as the origination and destination addresses of the call, the time the call started and ended, the duration of the call, the time of day the call was made and any toll charges that were added through the network or charges for operator services, among other details of the call.

The data set contains time series of call records from both fraudulent and legitimate callers and will be used in developing the neural network model. For privacy reasons the results will be presented with masked call data.

The data set was partitioned into three parts: training set, validation set and a test set. The training set was used in training the network. The Validation set was used to fine-tune model. Finally, the test set was used in testing the accuracy of the model.

Figure below shows a section of the raw CDR data. The first row is a description of the attributes. The actual data begins on the second row.

**Table 3.1.** CDR File

| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EVENT_INST | RE | BILLING_NBR | BILLING_IMSI | CALLING_NBR | CALLED_N | THIRD_N | START_TI | DURA | TRUNI | TRUNI | STATE | STATE_D | SUBS | BILLIN | EVENT | FILE_II | RECOR | BYTE_ | BYTE_ | LAC_A | CELL_ |
| 4450148382 | 959 | 260954902062 | 645030153229909 | 260954902062 | 971603523 | | 01-Mar-21 | 34 | | | A | 01-Mar-21 | 9.1E+07 | 945 | -1 | -1 | 63 | 0 | 0 | | 6.5E+14 |
| 4449852709 | 960 | 260955743781 | 645030143482684 | 260955743781 | 969516613 | | 01-Mar-21 | 97 | | | A | 01-Mar-21 | 9832120 | 945 | -1 | -1 | 19 | 0 | 0 | | 6.5E+14 |
| 4450430078 | 959 | 260954902062 | 645030153229909 | 260954902062 | 971603523 | | 01-Mar-21 | 43 | | | A | 01-Mar-21 | 9.1E+07 | 945 | -1 | -1 | 101 | 0 | 0 | | 6.5E+14 |
| 4450710359 | 959 | 260954902062 | 645030153229909 | 260954902062 | 977624882 | | 01-Mar-21 | 827 | | | A | 01-Mar-21 | 9.1E+07 | 945 | -1 | -1 | 115 | 0 | 0 | | 6.5E+14 |
| 4450886582 | 957 | 260957586281 | 645030147923585 | 260957586281 | 957586256 | | 01-Mar-21 | 94 | | | A | 01-Mar-21 | 6.3E+07 | 945 | -1 | -1 | 151 | 0 | 0 | | 6.5E+14 |
| 4450946122 | 957 | 260957586281 | 645030147923585 | 260957586281 | 957586256 | | 01-Mar-21 | 280 | | | A | 01-Mar-21 | 6.3E+07 | 945 | -1 | -1 | 1123 | 0 | 0 | | 6.5E+14 |
| 4451038231 | 960 | 260954902062 | 645030153229909 | 260954902062 | 966116070 | | 01-Mar-21 | 27 | | | A | 01-Mar-21 | 9.1E+07 | 945 | -1 | -1 | 279 | 0 | 0 | | 6.5E+14 |
| 4451356263 | 959 | 260954902062 | 645030153229909 | 260954902062 | 972116124 | | 01-Mar-21 | 19 | | | A | 01-Mar-21 | 9.1E+07 | 945 | -1 | -1 | 799 | 0 | 0 | | 6.5E+14 |
| 4451384684 | 960 | 260954902062 | 645030153229909 | 260954902062 | 968239391 | | 01-Mar-21 | 48 | | | A | 01-Mar-21 | 9.1E+07 | 945 | -1 | -1 | 544 | 0 | 0 | | 6.5E+14 |
| 4451380229 | 960 | 260954902062 | 645030153229909 | 260954902062 | 968914956 | | 01-Mar-21 | 7 | | | A | 01-Mar-21 | 9.1E+07 | 945 | -1 | -1 | 385 | 0 | 0 | | 6.5E+14 |
| 4451282451 | 959 | 260954902062 | 645030153229909 | 260954902062 | 977507871 | | 01-Mar-21 | 184 | | | A | 01-Mar-21 | 9.1E+07 | 945 | -1 | -1 | 871 | 0 | 0 | | 6.5E+14 |
| 4451502715 | 960 | 260954902062 | 645030153229909 | 260954902062 | 960258668 | | 01-Mar-21 | 80 | | | A | 01-Mar-21 | 9.1E+07 | 945 | -1 | -1 | 443 | 0 | 0 | | 6.5E+14 |
| 4451722122 | 959 | 260954902062 | 645030153229909 | 260954902062 | 971603475 | | 01-Mar-21 | 144 | | | A | 01-Mar-21 | 9.1E+07 | 945 | -1 | -1 | 440 | 0 | 0 | | 6.5E+14 |
| 4451700387 | 957 | 260954954702 | 645030141919409 | 260954954702 | 950033301 | | 01-Mar-21 | 557 | | | A | 01-Mar-21 | 3.4E+07 | 945 | -1 | -1 | 771 | 0 | 0 | | 6.5E+14 |
| 4451804352 | 960 | 260954902062 | 645030153229909 | 260954902062 | 969950688 | | 01-Mar-21 | 43 | | | A | 01-Mar-21 | 9.1E+07 | 945 | -1 | -1 | 254 | 0 | 0 | | 6.5E+14 |
| 4451926001 | 957 | 260954954702 | 645030141919409 | 260954954702 | 950033301 | | 01-Mar-21 | 7 | | | A | 01-Mar-21 | 3.4E+07 | 945 | -1 | -1 | 1348 | 0 | 0 | | 6.5E+14 |
| 4451866016 | 959 | 260954902062 | 645030153229909 | 260954902062 | 979072044 | | 01-Mar-21 | 43 | | | A | 01-Mar-21 | 9.1E+07 | 945 | -1 | -1 | 978 | 0 | 0 | | 6.5E+14 |
| 4451910315 | 959 | 260954902062 | 645030153229909 | 260954902062 | 977309430 | | 01-Mar-21 | 38 | | | A | 01-Mar-21 | 9.1E+07 | 945 | -1 | -1 | 248 | 0 | 0 | | 6.5E+14 |
| 4451972484 | 957 | 260954954702 | 645030141919409 | 260954954702 | 950033301 | | 01-Mar-21 | 224 | | | A | 01-Mar-21 | 3.4E+07 | 945 | -1 | -1 | 506 | 0 | 0 | | 6.5E+14 |
| 4452088978 | 959 | 260954902062 | 645030153229909 | 260954902062 | 973001205 | | 01-Mar-21 | 110 | | | A | 01-Mar-21 | 9.1E+07 | 945 | -1 | -1 | 328 | 0 | 0 | | 6.5E+14 |
| 4452068017 | 959 | 260954902062 | 645030153229909 | 260954902062 | 977725855 | | 01-Mar-21 | 123 | | | A | 01-Mar-21 | 9.1E+07 | 945 | -1 | -1 | 929 | 0 | 0 | | 6.5E+14 |
| 4452242859 | 959 | 260954902062 | 645030153229909 | 260954902062 | 974338122 | | 01-Mar-21 | 281 | | | A | 01-Mar-21 | 9.1E+07 | 945 | -1 | -1 | 432 | 0 | 0 | | 6.5E+14 |
| 4452356131 | 959 | 260954902062 | 645030153229909 | 260954902062 | 977908018 | | 01-Mar-21 | 99 | | | A | 01-Mar-21 | 9.1E+07 | 945 | -1 | -1 | 1110 | 0 | 0 | | 6.5E+14 |
| 4452792287 | 959 | 260954902062 | 645030153229909 | 260954902062 | 977575777 | | 01-Mar-21 | 23 | | | A | 01-Mar-21 | 9.1E+07 | 945 | -1 | -1 | 266 | 0 | 0 | | 6.5E+14 |
| 4452702439 | 960 | 260954902062 | 645030153229909 | 260954902062 | 762073717 | | 01-Mar-21 | 48 | | | A | 01-Mar-21 | 9.1E+07 | 945 | -1 | -1 | 1092 | 0 | 0 | | 6.5E+14 |
| 4452682217 | 960 | 260954902062 | 645030153229909 | 260954902062 | 968946307 | | 01-Mar-21 | 24 | | | A | 01-Mar-21 | 9.1E+07 | 945 | -1 | -1 | 1095 | 0 | 0 | | 6.5E+14 |
| 4452938350 | 959 | 260954902062 | 645030153229909 | 260954902062 | 971644347 | | 01-Mar-21 | 109 | | | A | 01-Mar-21 | 9.1E+07 | 945 | -1 | -1 | 357 | 0 | 0 | | 6.5E+14 |
| 4453076775 | 960 | 260954902062 | 645030153229909 | 260954902062 | 976252555 | | 01-Mar-21 | 130 | | | A | 01-Mar-21 | 9.1E+07 | 945 | -1 | -1 | 884 | 0 | 0 | | 6.5E+14 |

**Data Pre-processing**

After identifying the data source, data selection and data extraction followed. Before modelling, the data needed to be processed and presented in the correct format. Data pre-processing is the manipulation of data into a form suitable for further analysis and processing. The main activity during data pre-processing was to select the features to be used for building the network pattern recognition neural network model. Each of the call records has a set of call attributes as described earlier. These attributes need to be converted to a form that is compatible with the model. Data pre-processing involved the following:

i. Removing unwanted fields-these are the attributes that were not important in building the model.
ii. Replacing abbreviations with understandable values.
iii. Normalizing values between 0 and 1.
iv. Sorting the data to include only call data with sufficient number calls for modelling.

**Feature Selection**

The selection of features strongly influences the subsequent analysis. The goal of feature selection was to capture the desired attributes from the raw CDR data.

The provided CDR records consisted of individual CSV records containing all 91 attributes describing the call made through the system. Some attributes were noted to be of a zero value and were therefore were discarded in the analysis. On further analysis of the CDR records for the different call scenarios (Transferred calls, Conference calls, abandoned calls, Intercom Calls etc.) and from literature reviewed together with domain knowledge, the following set of attributes in table were selected and are very important in the fraud detection of sim box [13]:

The data obtained from CDR are not directly used for data mining since it may contain unreliable and noisy data or irrelevant and redundant data. Before the development of the model, the data must undergo the pre-processing process which such as feature extraction, integrating data, handling missing data and also identifying and removing outliners. Then, all the numerical variables have been normalized and compressed to a scale of 0 to 1 to prevent one attribute overly impact the algorithm's processing power simply because it contains large numbers.

Real-world data is often lacking attribute values (incomplete), lacking certain attributes of interest, containing only aggregate data, lacking in certain behaviours or trends, containing discrepancies in codes or names (inconsistent), and likely to contain many errors or outliers (noisy). A Python data pre-processing script was written to read into the raw CDR data and select the above features described above then label the records as either genuine or fraudulent based on the calling patterns.

**Table 3.2.** Feature selection

| Attribute | Description | Data type |
|---|---|---|
| Call sub | This is the Subscriber Identity Module (SIM) number which will be used as the identity field | Continuous |
| Total Calls | This feature is derived from counting the Total Calls made by each subscriber on a single day | Continuous |
| Total Numbers Called | This feature is the total different unique subscribers called by the customer (subscriber) on a single day | Continuous |
| Total Minutes | Total duration of all calls made by the subscriber in minutes on a single day | Continuous |
| Total Numbers Called in a day | The total different unique subscribers called during on a single day | Continuous |
| Total Incoming | Total number of calls received by the subscriber on a single day | Continuous |
| Called Numbers to Total Calls ratio | This is the ratio of the Total Numbers Called/Total calls | Continuous |
| Average Minutes | The is the average call duration of each subscriber | Continuous |

### Sampling Final Data Set

The pre-processed files produced for the three months used in this study were subjected to random sampling in order to produce a final data set of 7,006 records for building the network model. 7,006 records (3,330 frauds and 3,676 genuine) were pre-processed to produce the total of 7,006 records efforts were made to have a balanced dataset so as not to have the miss classification by the machine learning algorithm.

In the final data set 47.5% (3,330) of the records were fraudulent while 52.3% (3,676) were genuine. An almost equal percentage was used in order to train the network with an equal number of records of both classes.

### Encoding in Matlab Format

The sampled final data set of 7,006 records was encoded in a binary 1-of-N encoding format for use by the neural network built using the MATLAB neural network toolbox. Normalization is a technique often applied as part of data preparation for machine learning.

The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information. Normalization is also required for some algorithms to model the data correctly. For example, an input dataset contains one column with values ranging from 0 to 1, and another column with values ranging from 10,000 to 100,000.

The great difference in the scale of the numbers could cause problems when you attempt to combine the values as features during modelling. Normalization avoids these problems by creating new values that maintain the general distribution and ratios in the source data, while keeping values within a scale applied across all numeric columns used in the model. This module offers several options for transforming numeric data: You can change all values to a 0-1 scale, or transform the values by representing them as percentile ranks rather than absolute values. You can apply normalization to a single column, or to multiple columns in the same dataset. If you need to repeat the experiment, or apply the same normalization steps to other data, you can save the steps as a normalization transform, and apply it to other datasets that have the same schema [14].

A section of the python script written to perform this normalization is shown in figure below. Machine learning algorithms and deep learning neural networks require that input and output variables are numbers.

The encoding script produced a excel file with 4 fields of data attributes and 2 fields of class labels (fraud and genuine).

The Python script also produced a MATLAB data file from the normalised excel file. This was the data set used by the MATLAB neural network toolbox. The MATLAB data file contained two matrices:

i) The 7,006 X 4 inputs matrix contained 4 data fields for the 7,006 sample records
ii) The 7,006 X 2 targets matrix contained 2 class label fields for the 7,006 sample records

## 4. Performance Modeling & Results

Network Training Results

The neural network training ran for 73 epochs (iterations on the training data set) as shown in Figure 4-1 below;

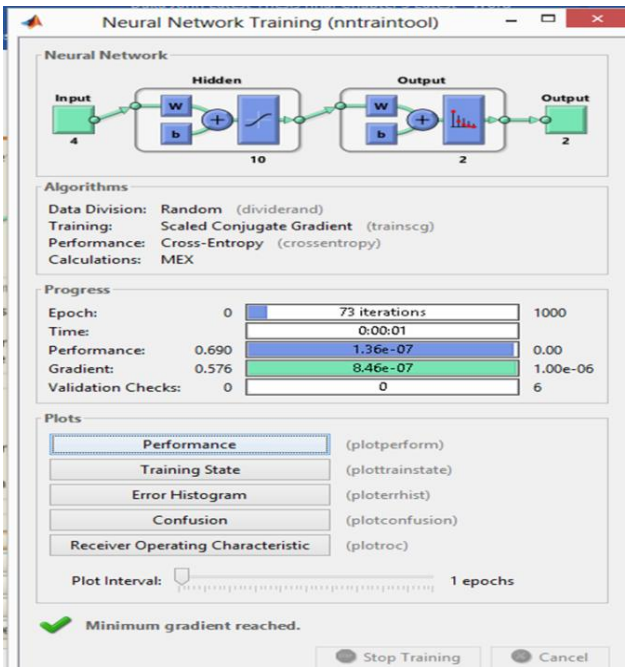


**Figure 4.1.** Neural Network training

The performance measure used was the Mean Squared Error (MSE). This is the average squared difference between

the class labels assigned by the network (outputs) and the actual class labels (targets). Lower MSE values are better and an MSE value of zero means no error.

In addition, the percentage Error was also used as a performance measure alongside the MSE. Which indicates the fraction of samples which are misclassified. A value of 0 means no misclassifications, 100 indicates maximum misclassifications.

The table 4-1 below shows the MSE and Percent Error results for the training, validation and testing data sets. It shows that the lowest MSE value was recorded on the validation data set while the highest MSE value was recorded on the testing data set. This is because the testing data set samples are not used during the training of the network.

**Table 4.1.** Training Results

Results

| | Samples | CE | %E |
|---|---|---|---|
| Training: | 4904 | 5.21522e-0 | 0 |
| Validation: | 1051 | 14.77366e-0 | 0 |
| Testing: | 1051 | 14.76764e-0 | 0 |

Plot Confusion Plot ROC

All the data sets had very small MSE values. This indicates that the network training produced a network capable of high accuracy classification ability. In additions, the three data sets had a Percent Error of 0 meaning that there were no misclassifications in any of the data sets.

The confusion matrices shown in Figure 4-2 below further elaborate these results. In the diagrams, 1 represents Fraud cases and 2 represents genuine cases. The Output Class is what is predicted by the network while the Target Class is the correct classification.

Crucially, the confusion matrices show that for the training, validation and testing data sets, all the samples were classified correctly. This is indicated by the value of 100% for correct classifications and 0% for incorrect classifications. The network outputs are very accurate, as seen by the maximum number of correct responses in the green squares and the minimum number of incorrect responses in the red squares. The lower right blue squares illustrate the overall accuracies.

The neural network achieved the high classification accuracy because it was able to learn quite well the patterns of genuine calls that were used to label the sample data set. It was then able to extend this pattern to unseen instances of call records.

These results indicate that the features selected for the training were very relevant. The calls are well defined by the call frequency, call duration and the time and day of the calls. These features are adequate for classifying the nature of calls as either being fraudulent or genuine.

Figure 4-3 below shows how the MSE reduces to a minimum value of 5.288e-07 at epoch 73 of training on the validation data set. The validation data set achieved the minimum MSE at epoch 73. This is what informed the training process to stop at that point since the latter epochs had higher MSE values.

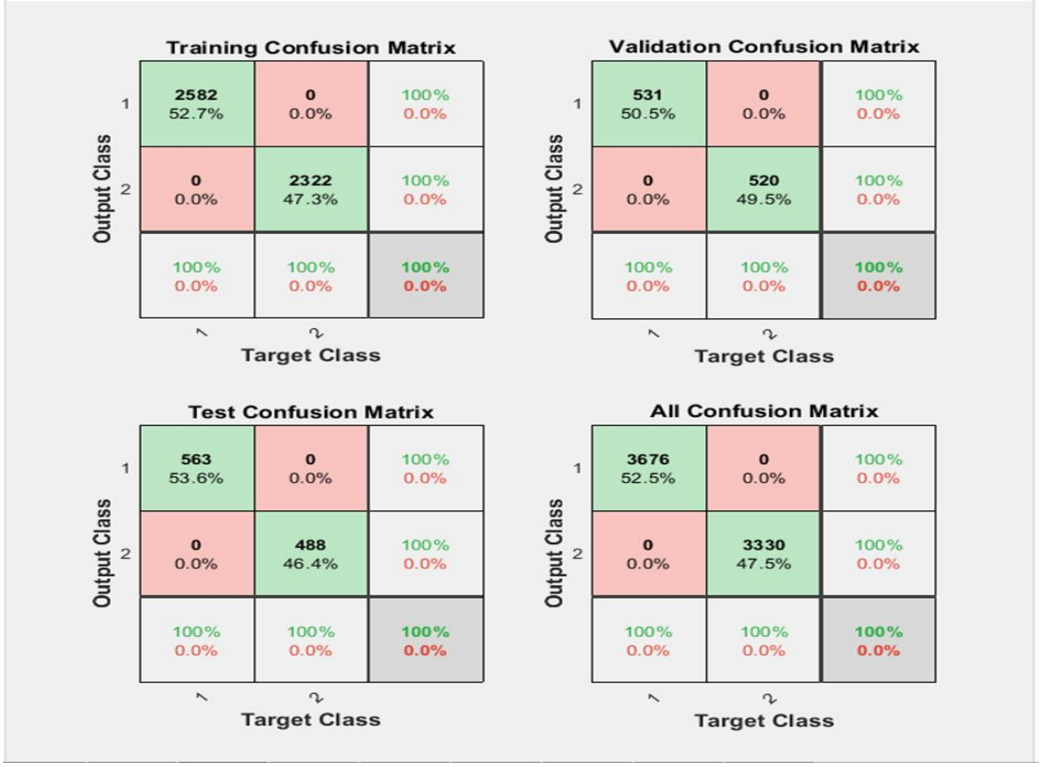


**Figure 4.2.** Confusion Matrix

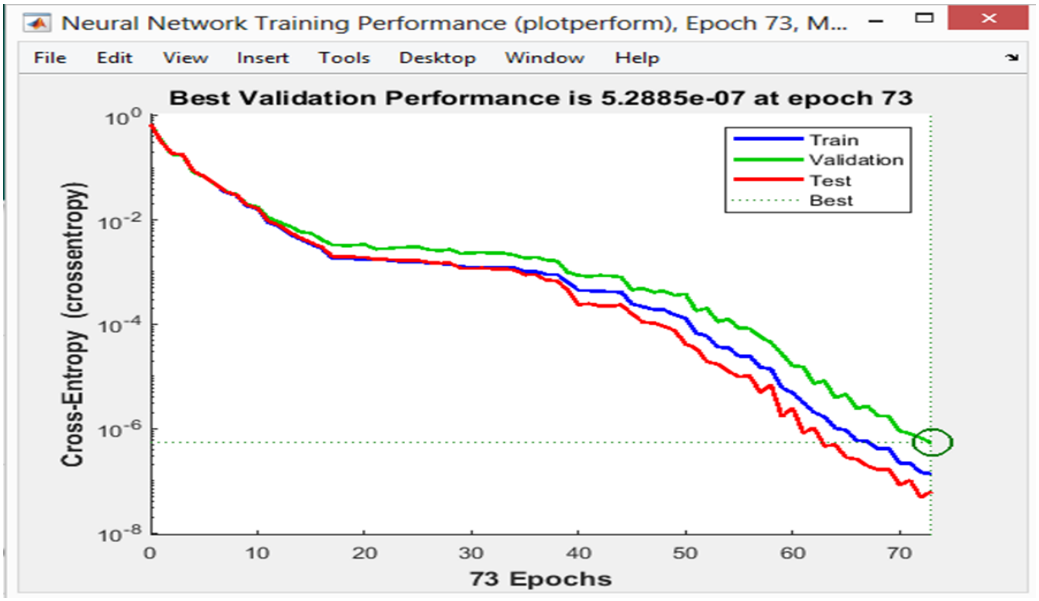


**Figure 4.3.** Neural Network Training performance

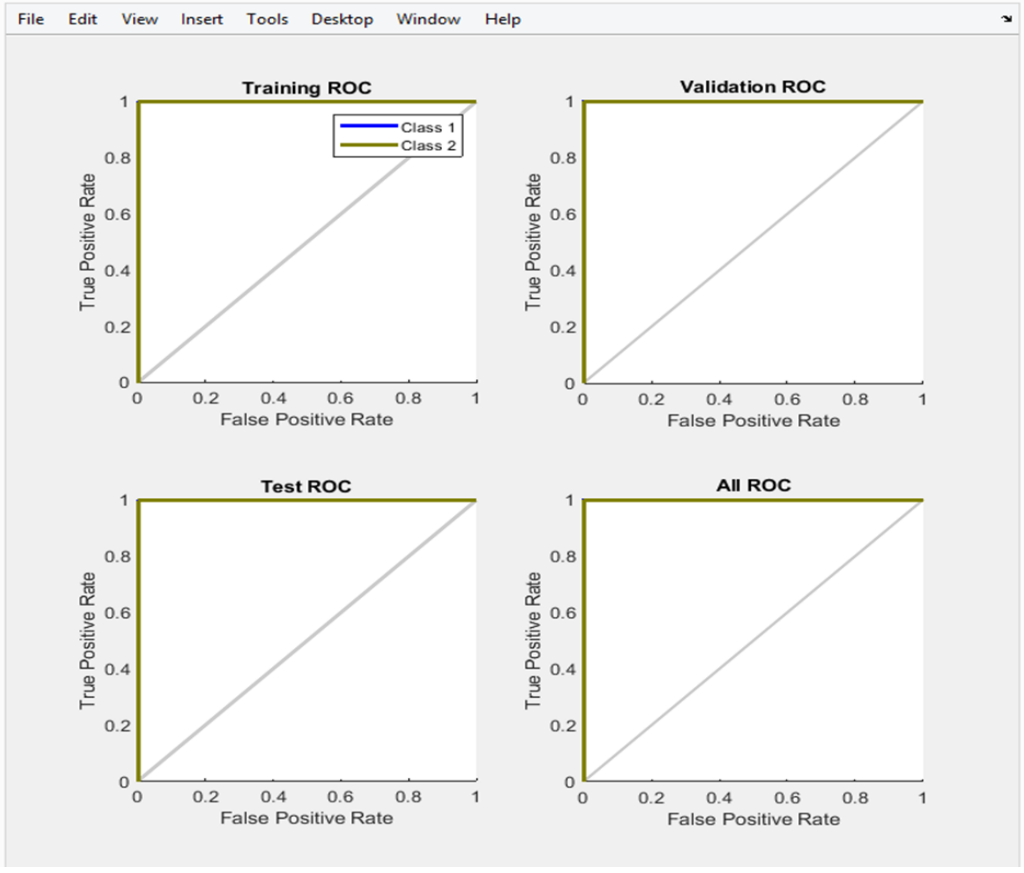


**Figure 4.4.** Receiver Operating Characteristic (ROC) Curves

**Receiver Operating Characteristic (ROC) Curves**

The blue colored lines in each of the axes of Figure 4-4 represent the Receiver Operating Characteristic (ROC) curves. The ROC curves show the relationship between the False Positive Rate (1 - specificity) and True Positive Rate (sensitivity) for the three data sets. A perfect test would show points in the upper-left corner, with 100% sensitivity and 100% specificity.

The ROC curves here all show that the False Positive Rate remains at a minimum while the True Positive Rate rises to a maximum due to the 100% correct classification performance of the network. For this problem, the neural network performs very well.

## 5. Future Scope and Conclusions

Fraud detection is one of the major step in mitigating revenue loss by telecom companies that is incurred annually, future work need to be focused on certain months of the year as fraudsters have focused on scamming callers in certain months of the year compared to other months. We also recommend that future research in this area combines different data for data mining and even use more than one model each with different algorithm to mine data. Comparing the results obtained by each model might increase confidence in the results.

## ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr. Mayumbo Nyirenda, Head of Department - Department of Computer Science, School of Natural Sciences at the University of Zambia, Special thanks also go to my wife Hilda, my sons-Taonga and Takondwa Bukata, Mum, Dad in memory, my siblings, and my friends. I love you guys. May God Continue Blessing you!

## REFERENCES

[1] M. Taniguchi, M. Haft, J. Hollmén, and V. Tresp, "Fraud detection in communication networks using neural and probabilistic methods," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, 1998, vol. 2, pp. 1241–1244.

[2] C. S. Hilas and P. A. Mastorocostas, "An application of supervised and unsupervised learning approaches to telecommunications fraud detection," *Knowledge-Based Systems*, vol. 21, no. 7, pp. 721–726, 2008.

[3] N. Laleh and M. A. Azgomi, "A taxonomy of frauds and fraud detection techniques," in *International Conference on Information Systems, Technology and Management*, 2009, pp. 256–267.

[4] R. A. Becker, C. Volinsky, and A. R. Wilks, "Fraud detection in telecommunications: History and lessons learned," *Technometrics*, vol. 52, no. 1, pp. 20–33, 2010.

[5] A. B. Desai and R. Deshmukh, "Data mining techniques for Fraud Detection," *International Journal of Computer Science and Information Technologies (IJCSIT)*, vol. 4, no. 1, pp. 1–4, 2013.

[6] D. M. Balasubramanian and M. Selvarani, "Churn Prediction in Mobile Telecom System Using Data Mining Techniques," *International Journal of Scientific and Research Publications*, vol. 4, no. 4, pp. 1–5, 2014.

[7] F. Altiparmak, B. Dengiz, and A. E. Smith, "A general neural network model for estimating telecommunications network reliability," *IEEE transactions on reliability*, vol. 58, no. 1, pp. 2–9, 2009.

[8] Y. Harkouss, J. Rousset, H. Chehade, E. Ngoya, D. Barataud, and J.-P. Teyssier, "The use of artificial neural networks in nonlinear microwave devices and circuits modeling: An application to telecommunication system design (invited article)," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 9, no. 3, pp. 198–215, 1999.

[9] Y. Khan, S. Shafiq, A. Naeem, S. Hussain, S. Ahmed, and N. Safwan, "Customers churn prediction using artificial neural networks (ANN) in telecom industry," *Editorial Preface From the Desk of Managing Editor*, vol. 10, no. 9, 2019.

[10] J. Gregar, "Research Design (Qualitative, Quantitative and Mixed Methods Approaches)," *Book published by SAGE Publications*, vol. 228, 1994.

[11] K. D. Shearer, "Experimental design, statistical analysis and modelling of dietary nutrient requirement studies for fish: a critical review," *Aquaculture nutrition*, vol. 6, no. 2, pp. 91–102, 2000.

[12] G. A. Chapman, K. Moores, D. Harrison, C. A. Campbell, B. R. Stewart, and P. J. Strijbos, "Fractalkine cleavage from neuronal membranes represents an acute event in the inflammatory response to excitotoxic brain damage," *Journal of neuroscience*, vol. 20, no. 15, pp. RC87–RC87, 2000.

[13] R. Sallehuddin, S. Ibrahim, and A. Hussein Elmi, "Classification of sim box fraud detection using support vector machine and artificial neural network," *International Journal of Innovative Computing*, vol. 4, no. 2, 2014.

[14] Xiaoharper, "ML Studio (classic): Normalize Data - Azure." https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/normalize-data (accessed May 08, 2021).