JIMMA UNIVERSITY

JIMMA INISTITUTION OF TECHNOLOGY

SCHOOL OF GRADUATE STUDIES

FACULTY OF COMPUTING AND INFORMATICS

COMPUTER NETWORKING

SUBSCRIPTION FRAUD DETECTION USING CONVOLUTIONAL AUTO-ENCODER AND DNN APPROACH IN CASE OF ETHIO TELECOM

A thesis submitted to the school of graduate studies, jimma university, jimma institute of technology, faculty of computing and informatics in partial fulfillment of the requirement for the degree of masters of science in computer networking

by

Rediet Wuhib

December, 2021

Jimma, Ethiopia

JIMMA UNIVERSITY

JIMMA INISTITUTION OF TECHNOLOGY

SCHOOL OF GRADUATE STUDIES

FACULTY OF COMPUTING AND INFORMATICS

COMPUTER NETWORKING

SUBSCRIPTION FRAUD DETECTION USING CONVOLUTIONAL AUTO-ENCODER AND DEEP NEURALNETWORK APPROACH IN CASE OF ETHIO TELECOM

BY: Rediet wuhib

Principal Advisor: Dr. Girum kettema (Ph.D.)

Co-Advisor: Mr. Gemechu Birhanu (M.Sc.)

Approved by Board of Examiners:

Dr. Melkamu Deressa sign: ____________________ Date: ____________________

External Examiner:

Mr. Temesgen Derejaw Sign: ____________________ Date: ____________________

Internal Examiner:

Mr. Faiz Akram Sign: ____________________ Date: ____________________

Chairperson:

DECLARATION

I declared that this research entitled “Subscription Fraud Detection using convolutional Auto-encoder and DNN Approach in case of Ethiotelecom” is my own original work and has not been submitted as a requirement for the award of any degree in Jimma University or elsewhere.

Declared by: Rediet Wuhib: Sign: ___________________

As research Advisor, I hereby certify that I have read and evaluated this thesis paper prepared under my guidance, by Rediet Wuhib entitled “**subscription fraud detection using convolutional auto-encoder and DNN approach in case of ethiotelecom**” and recommend and would be accepted as a fulfilling requirement for the degree Master of Science in Computer networking.

Confirmed By:

Principal Advisor

Dr Girum Kettema Sign: _______________ Date: ________________

Co-advisor

Mr. Gemechu Birhanu: Sign: _______________ Date:_________________

ABSTRACT

Nowadays, telecom services are becoming essential communication and business facilitators. However, the development of telecom services motivates fraudsters for illegal use. Hence, Telecom fraud becomes a serious challenge in the telecommunication sector which leads telecom companies to lose yearly profits and to deliver unfortunate quality of services for their Subscribers. Subscription fraud is a common and significant type of telecom fraud in today's business. The primary goal of the fraudsters is to make money illegally or to obtain telecom services with the intent of not paying for the service they used. Ethiotelecom is one of the oldest service providers in Africa and the sole service provider of Ethiopia which offers telecommunication services and products for the enhancement and development of the nation. In this paper, we use a predictive model using deep learning techniques to detect subscription fraud. To build a predictive deep learning model, we used convolutional auto-encoder and deep neural network (DNN) techniques. The fraud detection experimental result showed that the performance evaluation of the experimental result is 99.14% in terms of accuracy measurement using DNN technique on 1048576 voice datasets.

Keywords: Ethiotelecom, Subscription Fraud Detection, Deep Learning, Autoencoder, DNN

ACKNOWLEDGEMENTS

Above all, I would like to thank the almighty God and his mother, Saint Marry, for giving me strength endurance throughout my journey. My special thanks to my advisors Dr. Girum Kettema and Mr. Gemechu Birhanu for their kindly support.

Lastly, I would like to thank my family and friends, who without them would have not been possible. I also appreciate all the support I received from the Ethiotelecom employees especially for those employees who work in fraud class and my co-workers.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Acronym, Abbreviation and Terminology

AE:	Auto-Encoder
ANN:	Artificial Neural Network
DNN:	Deep Neural Network
CDR:	Call Details Record
CFCA:	Communications Fraud Control Association
CRM:	Customer Relation Management
FMS:	Fraud Management System
ML:	Machine Learning
QoS:	Quality of Service
RBM:	Restricted Boltzmann Machine
TSP:	Telecom Service Provider
VOIP:	Voice over Internet Protocol
ETC:	Ethiopian Telecommunications Corporation
CDMA:	Code Division Multiple Access
ADSL:	Asymmetric Digital Subscriber Line
VSAT:	very small aperture termina
EVDO:	Evolution-Data Optimized
VAS:	Value_Added Services
ccTLD:	Country code top-level domain
DNS:	Domain name system
ANFIS	Adaptive Neuro Fuzzy Inference

CHAPTER ONE

INTRODUCTION

1. Background

The telecommunications sector is fast evolving around the world, due to innovation replacing another innovation within a short period [1]. This sector is transforming the world into a global village at breakneck speed. As a result of the increasing number of mobile phone subscribers for virtual switching traditional network and voice over IP(VOIP) services, it becomes more users reachable in both urban and rural areas. Despite the significant increase in revenue and technological capabilities, this firm has suffered significant harm as a result of its subscribers' fraudulent activities [2]. Telecommunication fraud is defined as unauthorized access to a mobile operator's network to illegally profit from the network's operators and/or subscribers [3]. Yearly, fraudsters cost telecoms businesses across the world billions of dollars. Furthermore, due to the rising sophistication of fraud schemes, detecting fraud is becoming increasingly difficult. [4]. The result of the fraud problem is a significant loss of money, which has an impact on the credibility and performance of the operators.

Telecommunication fraud is classified into several categories based on its nature and characteristics. According to Kang and Yang [5], It is categorized into two as subscription and superimposed categories whereas Becker Volinsky and Wilks [6] classified telecommunication fraud into seven groups namely superimposed, Subscription, Technical, Internal Fraud, Fraud based on loopholes in technology, Social Engineering, and Fraud based on new technology. One of the common types of fraud is subscription fraud which we focus on it to detect and put problem-solving direction.

Subscription fraud costs Internet service providers a lot of money, according to the Communications Fraud Control Association (CFCA). Recent(2019) CFCA reports the subscription fraud has shown big impact on telecommunication companies still[4].

Data mining and artificial neural networks (ANN) have been utilized for forecasting, prediction, and categorization since network operators lose a significant amount of their revenue to fraud [7]. Following the Data mining approach, models have been built using naïve byes, tree, a Random

Forest and other algorithms to predict frauds. The classifier model performs with an accuracy of 95.43% [8]. A deep neural network (DNN) was used but the problem still needs to conduct researched on fraud detection because the performance result scored by data mining is very general whereas the performance gained by DNN shows less detection of fraud. So, we researched subscription fraud detection mechanisms using convolutional autoencoder and DNN deep learning techniques.

1.1. Motivation

Currently, prepaid and postpaid payment analysis in ethiotelecom is done using standard simple statistical approaches. However, the ethiotelecom call management system employs basic statistical approaches to analyze large amounts of data. It is a time-consuming, error-prone, tedious, routine, and non-dynamic activity. This traditional management system cannot handle the new behavioral change of fraudsters' activity because of the non-predictive nature of the system.

This research aims to design a predictive model that can identify subscribers who use telecommunication services without payment from ethiotelecom infrastructure. To overcome the drawback of the simple statistical method, we proposed a deep learning technique to detect subscription frauds. Fraud is a major problem in telecommunication industries which results in money loss in the finance component of ethiotelecom. This problem motivates us to conduct research that requires a solution to maximize profit and increase customer satisfaction.

1.2. Statement of the Problem

Globally, telecom fraud is on the rise, and it is now one of the major sources of revenue loss for operators [9]. According to a report by CFCA, revenue loss caused by subscription fraud alone is billions of dollars [10]. In Africa in 2015 mobile network operators lost more than $38 billion due to telecom fraud. Fraudsters are always looking for technological flaws in telecom firms to profit. Similarly, Telecom fraud has been a common business in Ethiopia. It is Ethiopia's sole telecommunications provider that offers a variety of services to its clients. Depending on the customer's needs, this organization offers both prepaid and postpaid payment options. Depending on the payment method, services might be prepaid or postpaid. Ethiotelecom registers its clients on the Customer Relationship Management (CRM) system, which is a database including subscriber information such as name, address, service plan, contact details, credit score, and payment history. In June 2016, Ethiopia's Information Network Security Agency (INSA) reported that Ethio telecom during nine months had lost over one billion birr (about $35.5 million at current exchange rates) due to telecom fraud. Ethio Telecom claimed a loss of $52 million per year due to telecom fraud in March 2017 [11]. In October 2018, the business revealed that it had lost 2.5 billion birrs (about $89 million at the current exchange rate) due to telecom fraud in a year [12]. Fraud detection in the telecommunication industry has been a major challenge still. Various fraud management systems are being used in the industry to detect and prevent increasingly sophisticated fraud activities even though they are not still secure the subscription fraud. The existing Fraud Management System (FMS) of ethiotelecom is rigid and non-dynamic in detecting and mitigating frauds, resulting in increased revenue losses for the organization as well as harm to subscribers' trust relationships. Because of CRM's rigid and non-dynamic nature, it has no way of knowing which subscribers are fake or not at the moment of service request or application.

Even though authors [9] [13] tried to identify frauds of ethiotelecom using a data mining approach. But Data mining is a more manual process that relies on human intervention whereas deep learning is a more dynamic approach. As new fraud behaviors have happened, ethiotelecom requires a dynamic way of fraud detection mechanism. Therefore, we propose subscriptions fraud detection model using convolutional autoencoder and deep neural network approaches which are suitable to build a detection model to identify frauds. This proposed model overcomes the fraud by detecting early potential frauds while operators are performing their tasks. We extract patterns from subscribers Call Detail Record (CDR) data for detecting fraudulent and non-fraudulent subscriptions. Therefore, we used dynamic and adaptive fraud detection and analysis approaches. Deep learning techniques are promising to effectively address the stated problem.

The Research shall answer the following questions

1. What kind of data features can be used to identify Subscription fraud using auto-encoder and DNN?
2. By how much accuracy Deep learning detect subscription fraud for ethiotelecom using auto-encoder and deep neural network (DNN)?
3. Which Epoch Becomes Most Optimal for Subscription Fraud Detection Purpose?
4. Which Method (Algorithm) performs better when comparison is made based on performance in term of performance accuracy and running time?

1.3. Objectives

1.3.1. General Objective

The main objective of the research is to develop a subscription fraud detection model for Ethiotelecom using deep learning approach.

1.3.2. Specific Objectives

To achieve the general objective of the study, the following specific objectives have been accomplished throughout the study:

- To Explore the best deep learning algorithm in the process of Subscription fraud detection.
- To investigate appropriate deep learning tools, algorithms, and techniques for Subscription fraud detection.
- To develop a prototype model that indicates the feasibility and applicability of the model for detecting and predicting.
- To design models with convolutional auto-encoder and DNN algorithms for the detection of Subscription fraud.
- Evaluate the performance of the models.

1.4. Methodology

1.4.1. Literature Review

The primary method for this study was a literature review. Problem identification through the review was the very first step towards achieving the final goal of the study. The study was first to review various researches on subscription fraud detection mechanisms in the telecommunication industry. Based on the understanding, Deep learning methods were used for the detection of subscription fraud in the case of Ethiotelecom.

1.4.2. Data Collection

In the second stage of this study Datasets were collected from Ethiotelecom real-world customers, attribute selection, preprocessing and creates the dataset for both train and testing of the developed fraud detection model.

1.4.3. Design and Development

We used experimental research design because we evaluate the performance of the developed model how much accuracy could have performs. The experimental research method attempts to deduce numerical figure from big data. We developed a model of DNN and convolutional Auto-encoder that can detect fraudsters from CDR data normal transactions to find anomalies from normal patterns. The developed deep learning based on convolutional auto-encoder (AE) was un supervised learning algorithm that applies backpropagation by setting the inputs equal to the outputs. In this research, we will use the TensorFlow library from Google to implement convolutional AE, and DNN, we also used Keras[14] as a high-level neural network API implemented for parallel processing to get AUC and confusion matrix.

1.5 Evaluation

The developed subscription fraud detection prototype model has evaluated its performance fairly and logically. We measured the performance of the developed predictive model through accuracy.

1.6. Scope and Limitation of the Study

To develop a fraud detection model, we collected call detailed data from ethiotelecom, analyses the data, extract features, design the architecture of the study, develop the model, and experimented and interpret results. Even though fraud can be grouped into many fraud types in the telecommunication sector, our work focused only on subscription fraud detection in the case of ethiotelecom. Hence, the paper is limited to subscription fraud in line with prepaid and postpaid payment approaches. In addition to this, subscription fraud detection is the concern of the study on call data from VOIP applications or directs from mobile, we do not deal with how to mitigate the detected subscription fraud.

1.7. Significance of the Study

This research work has the following significances

- The findings of the study can be used by departments such as Fraud Management and Revenue Assurance
- Describe the company's traditional fraud management system and its shortcomings.
- This work may also serve as a catalyst for future research on comparable and/or related themes.

- Minimizes or avoids the revenue loss of ethiotelecom if ethiotelecom applies following our fraud detection model.

1.8. Thesis organization

The thesis is organized as follows, Chapter two discussed different literature reviews concerning telecommunication, fraud detection, deep learning. Chapter three discusses the related works in detail. The selected works of literature are summarized in a table. Chapter four focuses on the methodology which comprises data collection, workflow, system architectures, evaluation mechanism, and tools used. Following this, Chapter Five explains the experimentation and evaluation, results, and its interpretations. Finally, Chapter Six concludes and recommends future works.

CHAPTER TWO

LITERATURE REVIEW

2.1 Overview

In this chapter, different points that are related to the current work are reviewed from literatures. Telecommunication and subscription Fraud, Ethiotelecom, Prepaid and postpaid Mobile Services, Artificial Neural Network, deep learning, auto encoders discussed in the context of our experimentation.

2.2 Telecommunication and Back ground of Telecommunication in Ethiopia

2.2.1 Telecommunication

A telecommunications system is a set of hardware and software that allows data to be sent from one location to another. The telecommunications system provides telephone, television, radio, and other services [15]. Text, data, pictures, speech, documents, and video information can all be transmitted using these platforms [3]. The telecoms sector creates and stores huge amounts of data when delivering telecom services. These data include call detail data, which describes the calls that go across telecommunication networks, network data, which defines the state of the network's hardware and software components, customer data, which identifies telecom customers, and prepaid end post-payment data. This vast amount of data must be managed properly for several purposes, including fraud detection, network performance analysis, call drop analysis, customer churn prediction, reporting to higher officials, network planning and optimization, and decision support [8]. A Call Detail Record is the data created by a telecommunications operator's switch for a subscriber (CDR). It is made comprised of data related to a certain phone call or other communication transaction handled by that switch. The network switch's capabilities dictate the degree of detail and information carried in the CDR. The decoder files or functional instruction manuals of the network switch generally show how to extract the information in CDR. CDRs are intended for invoicing purposes. It is also used for troubleshooting, determining Quality of Service (QoS), detecting fraud, gathering Business Intelligence (BI), and performing forensic investigations [18]. Telecommunications companies throughout the world provide a variety of services to their subscribers to compete in the market and win the hearts of their customers. In today's mobile global sector, subscribers utilize their Subscriber Identity Module (SIM) card numbers to subscribe to numerous mobile telecom services for personal use.

2.2.2 Background of Telecommunication in Ethiopia Ethiotelecom

Ethiotelecom Ethiopia, located in the Horn of Africa, is one of the world's oldest countries. Emperor Menelik II initiated a telecommunications service in Ethiopia in 1894 when the construction of a telephone line from Harar to Addis Ababa began [16]. Many of the Empire's main cities were linked by lines, allowing for long-distance communication with assistants or operators at intermediate stations, who often served as a spoken human repeater between the parties on the other end of the line. Between 1894 and 1952, the telco was renamed and rebuilt numerous times, finally taking over the whole country's telephone, telegraph, and radio communications.

Under the Federal Democratic Republic of Ethiopia, the telecommunications sector was restructured and two separate independent entities namely the Ethiopian Telecommunications Authority (ETA) and the Ethiopian Telecommunications Corporation (ETC) were established by Proclamation (Proclamation No. 49/1996: Art.3)[16].

The national operator, ethiotelecom, provides fixed, mobile, Internet and value-added services. It also offers dial-up Internet, CDMA 2000 wireless Internet, ADSL, and wireless Internet via ethiotelecom, the national operator, offers fixed, mobile, Internet, and value-added services. According to the corporation, it also provides dial-up Internet, CDMA 2000 wireless Internet, ADSL, and wireless Internet through AIRONET, VSAT, and EVDO. It delivers services to several government networks such as WorldNet, SchoolNet, and AgriNet, as well as non-government groups, using VSAT technology. In addition, ethiotelecom offers various Value-Added Services (VAS) such as Domain Name registration and maintenance for the.et country code top-level domain (ccTLD), the Domain Name System (DNS), Web hosting, and Internet Protocol Address service. Cyber cafés in the private sector are permitted to market airtime vouchers and Internet services[16]. The mobile network now covers a larger region than any other network in Ethiopia.

Despite its efforts to provide service to clients via wired and wireless services, Ethiotelecom investigations revealed that the company has experienced subscription fraud in a frustrating position. Subscription fraud has become one of the company's most critical problems. Using a bogus identity repeater between the distant calling parties, a subscription fraudster obtains the service from the operator without intending to pay for it. The telecom was renamed and restructured several times between 1894 and 1952, eventually taking over the entire country's telephone, telegraph, and radio communications[17]. The amount of harm caused by this form of fraud is determined by the fraudster's purpose to utilize the service for personal gain until he or she is discovered; on a more sophisticated level, the fraudster can exploit the service to benefit from its use. Bypass fraud deprives the terminating operator of incoming international call interconnect termination fees. To avoid international calls, this is commonly accomplished through the use of VoIP technology.

2.3 Prepaid and Postpaid Mobile Services

Prepaid and postpaid services are the two main mobile services supplied by TSPs the majority of carriers offer their consumers the choice of a postpaid or prepaid connection.

2.3.1 Prepaid Mobile Services

As the name implies, all transactions in this service are pay-as-you-go, which means that we cannot phone, send SMS, or use the Internet unless we have money in our account. Users can add credit to their accounts at any moment using several of payment options. When calls and messages are made, as well as data is consumed, the prepaid balance is deducted until no funds are left (at which time services stop functioning). A user can prevent service disruptions by making payments to increase the remaining amount. Nowadays, scammers use these sorts of services since there is no requirement for a guarantee to the TSP throughout the service request time. A single subscriber may have many SIM numbers with falsified Credentials due to a lack of sufficient subscriber identification. Subscription and SIM-Box fraud are the most common types of fraud that exploit this service[4]. With these and other difficulties, subscription fraud poses a global challenge to today's telecommunication service provider.

2.3.2 Postpaid Mobile Services

It is the most conservative service offered by telecom firms. The phrase 'post-paid' implies that credit is awarded for services used over a specific period, usually between one and six months.

A postpaid mobile phone is a payment method (also known as a mobile contract) in which a user enters a long-term contract (lasting 12, 18, or 24 months) or a short-term contract with a mobile phone operator (also known as a rolling contract or a 30-day contract). After each month, the user in this case gets billed after the fact for their use of mobile services. In most cases, the customer's contract stipulates the amount of "allowed" minutes, text messages, and other services. Any consumption equal to or less than that amount will be reimbursed to the client at a predefined rate. Any usage over that amount will be charged extra. A user in this position theoretically enjoys limitless access to mobile services and, as a result, unrestricted credit. This service is best suited to people who have a stable source of income. To make the 'post-usage' model possible, a postpaid service mobile phone normally requires two important components.

1. Contractual responsibilities and credit history This is the basis upon which the service provider may rely on the client to pay their bill on time and seek legal remedy if they fail to do so.
2. Service period to commit to utilizing the service, customers must sign long-term (1–3 year) contracts with most postpaid carriers. If the consumer fails to complete the term, he or she will be charged early termination costs. Credit history and contractual responsibilities give the foundation for the service provider to trust the client to pay their bill on time and have legal recourse if they do not. The bill is an important aspect of the service since it acts as a representative of the service provider and, in some cases, as verification of the service. The bill must be clear, intelligible, and aesthetically appealing in order for the subscriber to be engaged enough to notice information other than the bill amount.

2.4 Telecommunication Fraud

Telecom fraud is defined as the transmission of data or voice across a telecommunications infrastructure with the goal of not paying or paying at a very low rate for the service used[18]. Similarly, acquiring unbillable services and undeserved payments is characterized as fraud[19]. "Fraud is as ancient as humanity and may take on an infinite number of diverse shapes." However, in recent years, the advent of new technologies has made it simpler for users to interact and has helped enhance our spending power; however, this has also presented criminals with additional avenues to conduct fraud [7]. Traditional types of fraudulent activity include mobile phone fraud and computer penetration. The purpose of the subscriber plays an essential part in the depiction of fraud in most of the extant literature[7]. The fraudster views themselves as an entrepreneur, although one who employs unlawful tactics, but who is driven and led by the same challenges of cost, marketing, pricing, network design, and operations as any real network operator [3]. Fraud is appealing to fraudsters since the detection risk is low, no special equipment is required, and the goods in issue is readily converted to cash [15]. It is crucial to note that, while the term "fraud" has a specific legal meaning, it is commonly used to refer to abuse, dishonest purpose, or wrong behavior without indicating any legal ramifications.

2.4.1 Common types of telecommunication fraud

In today's world, there are several sorts of fraud in the telecommunications business, resulting in billions of dollars wasted each year. Different scholars classify different forms of fraud in various ways. According to Shawe-Taylor et al. [20], there are six types of fraud: subscription fraud, PABX fraud, handset theft, premium rate fraud, free phone call fraud, and roaming fraud. Hilas and Mastoro Costas, [21] classified fraud into four categories: technological fraud, contractual fraud, hacker fraud, and procedural fraud.

The third Scholars Kang and Yang [5]categories fraud types in to two subscription and superimposed frauds. in the following sub section, the researcher discussed most common types of frauds namely subscription, Superimposed, SIM box, Roaming, Sim cloning, Roaming.

Superimposed Fraud The most prevalent type of fraud on private networks is superimposed fraud. This is the scenario of an employee, the fraudster, who exploits the authorization number of another employee to access outbound trunks and pricey services[21]. In overlay fraud, a real account will be used. Mobile phone cloning and acquiring calling card authorization are two examples of this form of fraud. Fraudsters exploit valid accounts for unlawful purposes in a variety of ways. In such circumstances, anomalous usage is noted, which might be difficult to detect. Such scams are discovered when an authorized user complains about excessive billing. This scam may be avoided by constructing a strong system for verifying the validity of users. A SIM swapping attack works by convincing call center representatives working for a mobile phone provider to port a phone number to a new device. If they do that, they will innocently transfer control of the victim's phone number to the attacker. A SIM swap can be considerably easier when there is a collaborative insider to leverage. With someone working for the mobile carrier, an attacker doesn’t even need to carry out a social engineering ruse to gather the necessary information about the victim. It has become increasingly popular for cybercriminals to recruit mobile phone provider employees through social media accounts to scale their SIM swapping attacks. By posing as company hiring for open positions through these accounts, attackers have an opportunity to engage insiders through the promise of monetary gain[22].

Sim cloning: Sim cloning accomplishes the same aim as SIM changing, however it does not need contacting the mobile provider. Rather, it is a question of technological expertise. The cloning attack employs smart card copying software to duplicate the SIM card, allowing access to the victim's international mobile subscriber identity (IMSI) and master encryption key. Because the data is burned onto the SIM card, physical access to it is required. That entails removing the SIM card from the mobile device and inserting it into a card reader that may be connected to a computer that has the duplicating software installed. After the initial stealthy SIM replication takes place, the attacker inserts that SIM into a device they control. Next, the victim has to be contacted. The scam might start with an apparently benign text message to the victim requesting them to restart their phone within a certain amount of time. The attacker then begins their own phone before the victim resumes, initiating a successful clone followed by an account takeover. The attacker will have successfully taken over the victim's SIM and phone number once the victim resets their phone. The lawful phone user is subsequently charged for the calls made by the cloned phone. Cloning mobile phones is accomplished via cloning the SIM card located therein, rather than the phone's internal data [22].

SIM BOX: A SIM box scam is a setup in which fraudsters install SIM boxes with several low-cost prepaid SIM cards, most of which are subscribed with falsified credentials. The fraudster can then terminate international calls using local phone numbers in the relevant country to make the call appear to be a local call. This enables the box operator to circumvent international rates in order to fraudulently lower the prices charged by Mobile Network Operators (MNOs) and escape the government's tax. This statute prohibits the use of international phone calls by telecoms and the government. Aside from income loss, SIM Box operators degrade call quality, preventing them from reaching service level agreements. The fraudster will pay the network for a national call but bill the Wholesale operator for each minute he terminates; the Network Operator will lose the Interconnection fee [23].

Roaming: Maciá-Fernández [24] claims that Subscription fraud is one of the most popular ways of digital roaming fraud. Due to the delay in the home provider obtaining roamer call data, which can range from one to several days, thieves steal roamers' mobile phones, which are generally in holiday spots. It is the ability to utilize telecom services such as phone or data services outside of the home network without paying for them. In these circumstances, fraudsters take advantage of the lengthier timeframes required by the home network. When acquired SIM cards are transported to a foreign network, roaming fraud might begin as an internal or subscription fraud in the home network.

Premium Rate Service (PRS) The service charges rise when a service provider rents a premium rate number and a subscriber contacts the rented number in exchange for a service. Fraudsters rent out a large number of premium rate service numbers. Furthermore, they create new subscriptions that supply false or stolen identities (subscription fraud). The scammers then dial (PRS) numbers that they own using the subscriptions they produced, leaving those numbers with an unpaid debt. International revenue sharing fraud, like PRS fraud, involves the usage of rented premium lines while traveling outside of the country. The concept of international revenue share fraud is identical to that of PRS fraud, except that the fraudster calls the rented premium lines while roaming outside of the country. The ability to roam allows the fraudster to more easily make the alteration.

Subscription Fraud: Subscription fraud is defined as a fraudster utilizing his or her own, stolen, or created identities to get services with no intention of paying. Subscription fraud has been identified as the most harmful sort of non-technical fraud [4]. It is usually the pre cursor to other types of fraud such as Premium Rate Fraud, International Revenue Share fraud, SIM-Box fraud and Roaming fraud which are lethal in their own rights [25]. The true cost of this form of fraud is difficult to quantify since it does not end with income loss. The consequences can be disastrous in terms of growing complaints, poor customer experience, and support staff unhappiness[26]. Subscription fraud is a contractual fraud[19]. Subscription fraud is the most popular since a fraudster does not need to use a digital network's encryption or authentication methods when using a stolen or created identity. Their favored approaches are low techniques, which use the network below the FMS threshold level. This has a lower likelihood of being discovered. In relation to Koi-

Acroti et al. [5] Subscription fraud is now the most popular and fastest-growing sort of fraud. In a similar vein, considers subscription fraud to be the most major and frequent sort of global telecommunications fraud. Subscription fraud occurs when a fraudster acquires a subscription (perhaps using false identification) and begins a fraudulent activity with no intention of paying the bill. Subscription fraud is the most difficult type of telecommunications fraud to identify, and it may cost firms a lot of money [4].

Cryptography, data mining [27], machine learning [22], neural networks[20] and deep learning have all been used to investigate fraud detection and mitigation. Deep learning is also being used to uncover subscription fraud in worldwide telecommunications[18]. Deep learning has been used in many fields such as image recognition in Facebook, speech recognition in Apple or Siri, natural language processing in Google translator and in fraud detection areas[28].That is why the researcher like to work with the state of the art deep learning methodology for subscription fraud detection in case of ethiotelecom.

2.5 Artificial Neural Network

Another algorithm is the neural network. Neural networks are a set of algorithms that are based after the capacity of the human brain to recognize patterns. The systems are driven by distinct parts of neurons, known as "units" or simply "neurons." Each unit has a unique set of weighted input sources. These weighted information sources were linked together and then sent via an activation function to obtain the unit's output [27].

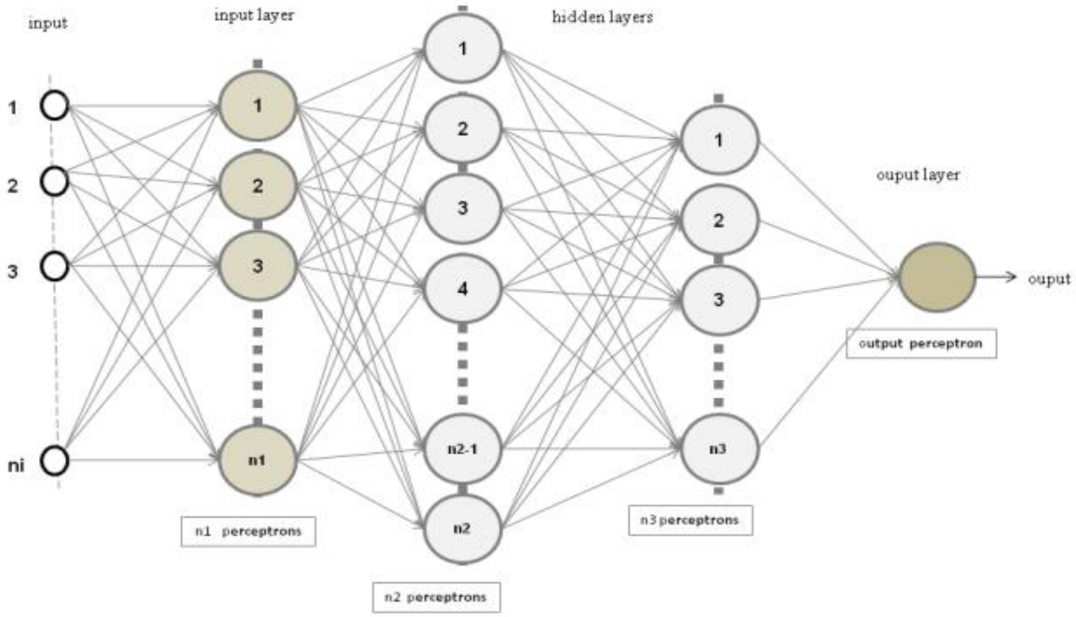


Figure 1: Multilayer Neural network [29]

2.6 Deep Learning

Machine learning is fantastic, but deep learning takes it to the next level. It is a branch of machine learning that aims to construct increasingly complicated hierarchical models that are more closely related to human cognitive processes than simple machine learning models. As a result, it advances beyond machine learning while being a subset of AI and a step above machine learning. Machine learning is becoming the norm, but deep learning offers the greatest promise for addressing real-world problems that go beyond entertainment [30].

Between the input and output in DNN, there is a multilayer perceptron or hidden layer. By moving through each layer, all of the layers are connected to the previous layers, and the network calculates the precise output based on the weights and activation function. We can model any complicated non-linear relationship using DNN. The feature of learning about the feature is the DNN's backbone that is most relevant to the targets[31]. By combining graph convolution neural network combination optimization and Bayesian neural network, the DNN bridges a research gap in model selection, training dynamics, and uncertainty assessment. DNN may be used for a wide range of applications, including computer vision, machine translation, social network filtering, playing boards, video games, and medical diagnostics. These layered representations are learnt using neural networks, which are deep learning models that are constructed in literal layers placed on top of one other. Deep learning techniques based on deep neural networks have risen in popularity as high-performance computer resources have been more widely available. Because of its ability to analyze a large number of characteristics, this approach provides more power and flexibility while working with unstructured data. This method transmits the input across numerous levels, each of which can extract characteristics gradually and transfer them to the next layer The first layers extract low-level information, while subsequent layers integrate them to generate a comprehensive representation[32]. The deep in deep learning does not refer to any form of greater knowledge attained by the technique, but rather to the concept of successive layers of representations. The number of layers that contribute to a data model is referred to as the model's depth. Other names for the field that may have been used were layered representations learning and hierarchical representations learning. Deep learning in the modern day frequently contains tens or even hundreds of consecutive layers of representations, all of which are learnt automatically by exposure to training data. " Other machine learning algorithms, on the other hand, tend to focus on learning only one or two layers of data representations[27]. As a result, they are also known as superficial learning. Deep learning models called neural networks are used to learn these layered representations, which are arranged in literal layers piled on top of each other. Deep learning is a catch-all name for multilayer neural networks. Deep learning techniques such as AE, deep convolutional network, support vector machine, and others are available for implementation. One disadvantage of using an algorithm to solve a problem is that the developer must understand the underlying problem and what each deep learning method performs. The three unsupervised learning algorithms in deep learning are RBM, AE, and the sparse coding model. Unsupervised learning extracts the important aspects of the data automatically, takes advantage of the availability of unlabeled data, and adds a data-dependent regularization for training. Deep learning has been applied in a variety of fields, including Facebook's image identification, Apple's Siri's speech recognition, and Google Translator's natural language processing. Yamini Pandey used deep learning with the H2O To understand complex patterns in the dataset. H2O is an open-source platform for Big Data predictive analytics. Predictive analytics underpin supervised learning. To discover credit card fraud trends, the author employed a multi-layered, feed forward neural network based on H2O’s performance[33].

2.7 Deep Learning Approaches

Deep learning shows good performance in supervised learning, unsupervised learning, Reinforcement learning, as well as hybrid learning.

Supervised Learning: In supervised learning, the input variables represented as X are mapped to output variables represented as Y by using an algorithm to learn the mapping function f.Y=f(X)(1)The aim of "To forecast the output (Y) for a new input (X), the learning algorithm approximates the mapping function[55]. The output can be corrected using the mistake from the predictions generated during training. When all of the inputs have been taught to produce the desired output, learning can be halted. Regression for solving regression problems, Support Vector machines used for classification, Random Forest for classification as well as regression problems.

Unsupervised Learning: In unsupervised learning, we have the input data only and no corresponding out-put to map. This learning aims to learn about data by modeling the distribution of data. Algorithms can be able to discover the exciting structure present in the data. Unsupervised learning is used to solve clustering and association problems. The unsupervised learning algorithms such as K-means algorithm is used in clustering problems, Apriority algorithm is used in association problems.

Reinforcement Learning: In Reinforcement Learning to train the algorithm, employs a reward and punishment scheme. In this case the algorithm or agent is learning from its surroundings. When the agent performs correctly it is rewarded and when it performs incorrectly it is penalized [34]. For example, consider the case of a self-driving car the agent gets a reward for driving safely to destination and penalty for going off-road. Similarly, in the case of a program for playing chess, the reward state may be winning the game and the penalty for being checkmated. The agent tries to maximize the reward and minimize the penalty. In reinforcement learning, the algorithm is not told how to perform the learning; however, it works through the problem on its own.

2.7.1 Keras Framework

Keras is a Python-based API that runs on top of TensorFlow. It is a model level library, providing high-level building blocks for developing deep learning models. It doesn’t handle low-level operations such as tensor manipulation and differentiation and fast experimentation. Instead, it relies on a specialized, well-optimized tensor library to do so, serving as the backend engine of Keras. Rather than choosing a single tensor library and tying the implementation of Keras to that library, Keras handles the problem in a modular way; thus, several different backend engines can be plugged seamlessly into Keras. Currently, the three existing backend implementations are the TensorFlow backend, the Theano backend, and the Microsoft Cognitive Toolkit (CNTK) backend[35]. It runs on CPUs and GPUs and supports both CNNs and RNNs. TensorFlow, a Google Brain project, supports languages including Python, C, and R. It enables us to deploy.

2.8 Deep Learning in Fraud Detection

Deep learning is the current trend in the field of machine learning where the multiple hidden layers are used to give a better outcome than other machine learning algorithms. This is evident from the literature review, where the systems that use deep learning have better advantages than machine learning [28]. Its hidden layers have a high computational capability and hence able to achieve a high accuracy. since Ethiotelecom CDR contain millions of subscribers, the data is considered to be big data. Ordinary machine learning algorithms will not be able to handle such amounts of data. Also, other algorithms reach a saturation point, where the performance peaks at a particular amount of data and does not increase even when the data size is increased. However, this is not the case with deep learning where the performance continues to increase when the data size is increased[34]. Hence, this huge number of subscribers can be learnt and classified into legitimate and fraudulent subscribers by the DNN. Deep learning constitutes different types based on the type of the data and the methodology used. Some of these are Deep Neural Network (DNN), Autoencoders, Convolution Neural Network (CNN), Generative Adversarial Network (GAN), and Recurrent Neural Network (RNN)[22]. Deep Neural Networks contain multiple hidden layers between input and output layers. Author [7] [9] have used deep neural networks to detect fraud in telecommunication operations. The proposed system is based upon the subscriber’s duration and call fee amount.

Autoencoders are a category of neural networks that learn efficient data encodings, and their reconstructions, and hence can be used as a classifier based on their reconstruction error. Energy based probabilistic models like Restricted Boltzmann Machines (RBMs) can be used to learn the distribution of data. Autoencoders and RBMs have been used to detect fraud in[33]. Author [36] have applied autoencoder for fraud detection in credit card transactions. They have utilized a deep autoencoder to perform feature selection and then classified the transactions with high accuracy and the low variance. The survey it can be concluded that although learning algorithms have been largely used for fraud detection. However, little attempt has been made to extract features from the transaction attributes. This extraction allows the learning model to learn the distributions of fraud more effective unencumbered by irrelevant features. The model proposed in this work uses convolutional autoencoder to extract important features from Ethiotelecom CDR data. Autoencoders have the ability to detect complex non-linear correlations among features of the data. This ensures that the encoded data thus obtained from the autoencoder is devoid of correlated and irrelevant features The compressed data thus obtained is used to train a classifier. Autoencoders[37] are a specific category of feed-forward neural networks that are used to learn efficient encodings of the training data. An autoencoder network has the same input and output dimensions it transforms the input to a hidden representation, having a different dimension than the input (and output) dimension, and then reconstruct the input from this hidden representation.

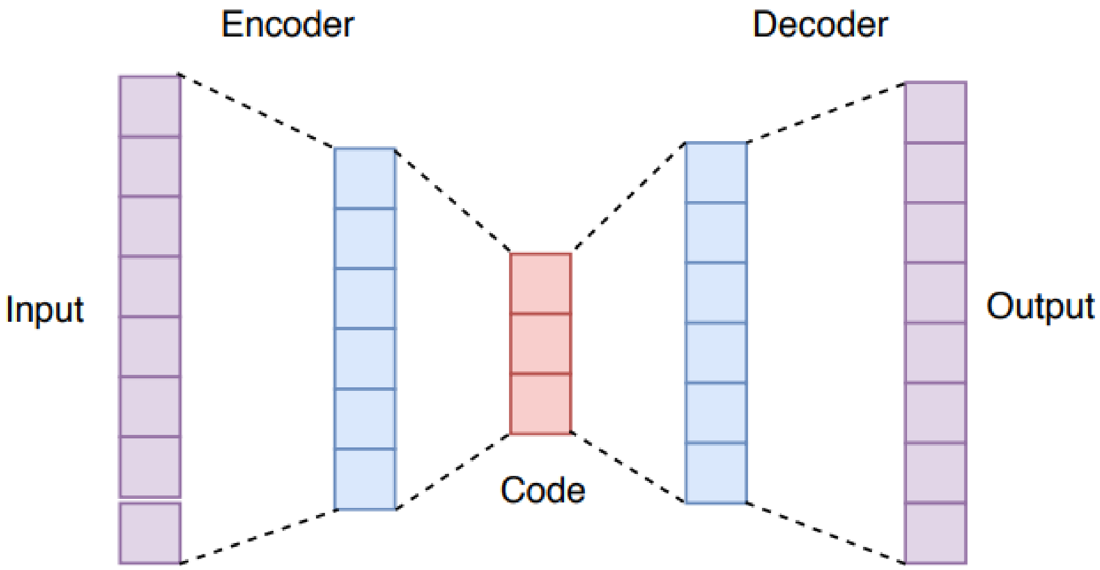


Figure 2: General structure of Auto-encoder [38]

Auto-encoder comprises of input, encoder, bottleneck, decoder, and reconstructed input. Encoder figures out how to decrease the input dimension and compress the input data into an encoded portrayal. Bottleneck contains the compress representation of input data. This is the least conceivable dimension of input data. Then again, the decoder figures out how to recreate the information from the encoded representation to be as near as could be expected under the circumstances.

CHAPTER THREE

RELATED WORK

3.1 Overview

In this section, different works which are related to machine learning-based fraud detection mechanisms for telecommunication sectors are presented. The papers presented fraud detection methods from customer subscription data. Relevance and similarities of the papers with this thesis work are presented. To have a detailed understanding of these research topics about fraud detection mechanisms particularly, subscription fraud detection, we review both international and local research works such as journals, articles, magazines are reviewed. Understanding the characteristics of fraudulent behavior of customers in the telecom industry is a critical task.

S.wu N. Kang Naidong and Yang, Liu [5] Analyze the common characteristics of fraudulent behavior of customers in telecom industry systematically. They used data mining approach to find outliers following clustering mechanism. Outlier finding requires an effective solution to forecast the customers who are maliciously in arrears beyond data mining implementation.

Mert Sanve, Adem Karahoca [28] In addition, using data mining (DM) techniques, they devised and executed a subscription fraud detection system. The Adaptive Neuro Fuzzy Inference (ANFIS) method is recommended as a way of effective fraud detection in their study, which assesses several data mining methodologies to determine the best practical option for detecting telecom fraud. ANFIS International Works properly categorized 98.33 percent of the cases in the test run. As a result, it is critical to deal with telecom fraud detection utilizing a cutting-edge fraud detection method rather than a standard data mining strategy. Deep learning is now attracting the fraud world by merging several methodologies such as Deep neural network (DNN), long short-term memory (LSTM), and auto encoders.

In these papers[19][21] Deep learning is accomplished by the application of state-of-the-art deep learning methods. Becomes a promising anomaly detection for the real application of predicting the consumers who are maliciously in arrears in the telecom business, but the article did not particularly exhibit anomaly identification on call detailed records of customers. Following their deep learning strategy, we present a special way to anticipate fraud calls from subscribers utilizing convolutional auto encoder and DNN on top of deep learning approaches.

Mhd Redwan Al Bougha compared four main methods, those algorithms are Boosted Trees Classifier, Support Vector Machines, Logistic Classifier, and Neural Networks. The findings show that Boosted Trees and Logistic Classifiers performed the best of the four algorithms, with a false-positive proportion of less than 1%. Support Vector Machines outperformed Boosted Trees and Logistic Classifiers, but with an 8% higher false-positive rate. The accuracy rate of neural networks was 60%, with a false positive ratio of 40%. Boosted Trees and Support Vector Machines classifiers are among the best algorithms for detecting SIM box fraud due to their high accuracy and low false-positive ratios[23].

Even though, the author [23] [9]Classification score becomes satisfactory, the selected algorithms are pure classification algorithms which are not specialize for anomaly detection's sake. Attack detection mechanisms such as autoencoder based algorithms were better for experimentation which becomes feasible for real telecommunication fraud detection task. So, we use auto encoder fraud detection mechanism on top of Keras framework of deep learning approach.

Priyanka Sharma, Santoshi Pote [33] performed study on Credit Card Fraud Detection on an Online Payment System Using Deep Learning Based on Neural Network and Autoencoder. The techniques used now a day detects the anomaly only after the fraud transaction takes place. The intruders have figured out how to exploit security flaws in the system. These frauds are not consistent in their actions, they constantly alter. Auto-Encoder (AE) is designed to remake high-dimensional information utilizing a neural system model with a narrow bottleneck layer at the center. AE has the input equivalent to the output in the output layer that has pretty much the sort of input units. This paper discussed the performance analysis and the comparative study of the two Deep Learning algorithms which include auto-encoder and the neural network. Fraud detection score of auto encoder was 99.48% accuracy. Having the promising method, algorithm and obtained performance result, we want to implement auto encoder for subscription fraud detection in case of ethiotelecom.

Apapan Pumsirirat, Liu Yan [36] also conducted Fraud Detection entitled as Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine. These paper focuses on fraud cases that cannot be detected based on previous history or supervised learning and create a model of deep Auto-encoder and restricted Boltzmann machine (RBM) that can reconstruct normal transactions to find anomalies from normal patterns. The proposed deep

learning based on auto-encoder (AE) is a supervised learning algorithm that applies backpropagation by setting the inputs equal to the outputs using the Tensor flow library from Google to implement AE, by using deep learning.
In this paper, they use the Keras framework [14] which is implemented using python, and we code in python on Jupiter notebook. The datasets need to be collected as prepared as per the German credit card data set or the Australian dataset feature or attribute formats for accuracy trustworthiness [33].

Mohammad Braei and Sebastian Wagner [39] wrote a survey on anomaly detection in univariate time series. Researchers used (deep) neural networks to try to improve these procedures. In the light of the increasing number of anomaly detection methods. Multi-Layer Perceptron (MLPs), Deep Neural Network (DNNs), and Long-Short Term Memory (LSTMs) deep learning branches have been implemented for anomaly identification. However, Time series-based anomaly detection is not still researched in the case of ethiotelecom.

Lulu[13] acquired sample data on three-time stamps such as 24-hour, 1-month, and 1-week day and night peak traffic hours, showing QoS KPIs from ethio telecom's live network The author employs the knowledge discovery process (KDD) and tests the dataset with the Weka tool. J48 decision tree, Nave Bayes, and Multilayer Perception classifiers were used as classification methods. The data set in this research is too little to generate knowledge from. Data mining is the process of extracting knowledge from large amounts of data rather than small amounts of data.

Yard[8] developed a predictive model that can determine mobile call drops from ethiotelecom mobile network data using data mining techniques from Fault Management System (FMS) data since the FMS data includes enough information about call drop reasons. In telecommunications, the dropped-call rate is the fraction of telephone calls due to technical reasons. Call drop is a situation where calls were cut off before the speaking parties had finished their conversational tone and before one of them had hung up. This fraction is usually measured as a percentage of all calls A call attempt initiates a call setup procedure, which results in a connected call if successful. A connected call may be terminated (disconnected) due to a technical reason before the parties making the call would wish to do so such calls are classified as dropped calls. The study was conducted using WEKA software version 3.8.1 and four classification techniques; namely, J48 and Random Forest, PART and JRIP algorithms. As a result, J48 decision tree algorithm scores

better performance with processing speed of 95.43% and 0.06 sec respectively. Even though both research works [8][9] showed good processing performance, data mining is a very general method which is not based on deep analysis of data as the state of the arts` deep learning approach.
Kahsu [23] conducted Telecommunication fraud as it drives telecom operators to lose a portion of their annual revenue for ethiotelecom. The author selected SIM-Box Fraud from many fraud types and detect it Using Data Mining techniques. SIM-Box is a Bypass fraud which is the most worrying fraud type in today's telecom business. With the introduction of new technology, fraudsters gained new tools for device bypass fraud. Subscriber Identity Module box (SIM box) fraud is a common sort of bypass fraud that has occurred as a result of the adoption of Voice over Internet Protocol (VoIP) technologies. Models have been built. to classify Call Detail Records (CDRs) to identify illegal subscriber data from legitimate subscribers one. Three classification algorithms were Random Forest (RF), Artificial Neural Network (ANN) and Support Vector Machine (SVM) used and RF performed better among the three algorithms with accuracy of 95.99%.

Derebe [22] conducted research to explore the potential application of supervised machine learning algorithms in fraud detection in the case of ethiotelecom call data. According to the researcher, the applications of supervised machine learning methods and tools to large quantities of data generated by the Call Detail Record (CDR) of telecommunication switch machine were expected to address the serious problems of telecommunication operators. The CDR consists of a vast volume of data set about each call made and it is a major resource of data for research works to find out hidden in addition to the traditional use for bill processing activities, customers' call habits.
They designed subscription fraud detection system using three supervised machine learning algorithms Artificial Neural Network (ANN), Support Vector Machine (SVM) and J48. As a result, J48 algorithm using Cross Validation (CV) options is found to be the best classifier algorithm by scoring 99.3% accuracy followed by the two algorithms highest scores of ANN (CV) and SVM(ST) with 97.51% and 96.0% respectively. This result happens because of J48's capable of learning disjunctive expressions in addition to it reduced error pruning. Pruning decreases the complexity in the final classifier, so that improves predictive accuracy from the decrease of over fitting.

All related papers presented have their own roles of detecting subscription fraud depending on the nature of the telecom service provider's customer usage behavior. So, we can see different types of methods and techniques of subscription fraud detection. What we clearly observe is that algorithm types, feature choices, and dataset size limits have the most impact on classification performance. In this study, by considering subscription fraud usage behaviors we use feature number of duration and call fee amount of the subscribers.

To sum up, in this research work, we propose convolutional auto encoder and DNN based fraud detection using deep learning models. Deep learning is a part of machine learning (ML) which makes fraud to be detected from subscribers' call data.

The following is the summary of the existing works on the given domain:

Table 1 Summary of related works

Author	**Dataset Used**	**Tool Used**	**Technique Used**	**Algorithm used**	**Accuracy/conclusion**
Derebe Tekeste	CDR	WEKA	Machine Learning	J48, SVM	96%
Yared Alibo	Fault Management data	WEKA	Data mining	J48	95.43%
Kahsu Hagos	CDR	WEKA	Data mining	RF, ANN and SVM	95.99%
Hailemeskel G/Tsadik	CDR	WEKA	Machine Learning	RF, ANN and SVM	**99.46%**
Lulu Deyu	KPIs	WEKA	Data mining	J48, the Naïve Bayes, ANN	84.5%
Priyanka Sharma and Santoshi Pote	European dataset	PyTorch	Deep Learning	Auto-encoder, neural network.	99.48%, and 99.94% respectively
ApapanPumsirirat, Liu Yan	German dataset Australian dataset European Dataset	Keras top of TensorFlow with H2o	Deep learning	Auto-Encoder and Restricted Boltzmann	43.76%,54.83% and 96.03% respectively

Sen Wu, Naidong and Kang Liu Yang	Kdd dataset	SPSS	Artificial Neural Network	Kohonen neural network algorithm	forecast the cheat in telecom industry

CHAPTER FOUR

METHODOLOGY

4.1. Overview

To achieve the objective of this study, we construct a methodology which includes data collection (Raw CDR Data), preprocessing and evaluation techniques. We discussed about the dataset, research design, corpus preparation and feature extraction, workflow and procedures, designing of the proposed system architecture and model emulation mechanisms.

4.2. Research design

In this Research work we used Experimental research designed. Experimental design manipulates data that works with original materials to build models. Models can be mathematical formula or algorithm which can be expressed in terms of quantified numbers. The type of design to use is determined by the nature of the challenges given by the study objectives. Accordingly, we used experimental design which is characterized by careful selection of 80% training set to be tested using 20 % test set. So, this research is experimental research because it is to get the performance of autoencoder and deep neural network numerically as accuracy and loss amount is quantified using training and predicting experiments. We conduct experiments several times with different parameters to quantify the performance of the proposed model.

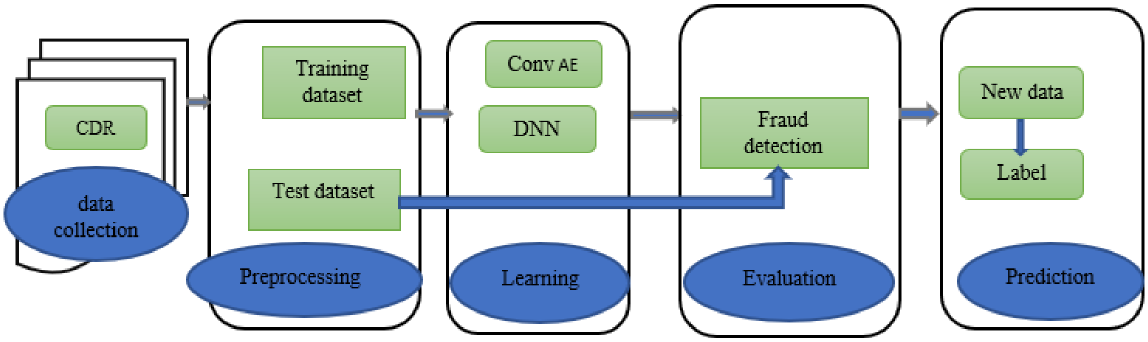


Figure 3: Proposed System Architecture (model)

4.3. Research methods/Approaches

We used CNN autoencoder and deep neural networks methods for building a fraud detection model. An autoencoder is a specific type of neural network, which is mainly designed to encode the input into a compressed and meaningful representation, and then decode it back such that the reconstructed input is similar as possible to the original one [40]. The autoencoder accepts filter inputs by using CNN potential. The autoencoder is then used to find how much loss is observed after compression from CDR input data given. Autoencoders are used for the reduction of the dimension of data, novelty detection problems, as well as in anomaly detection problems. The deep Neural network uses a feed-forward multilayer neural network concept under the Keras framework. Collecting data from Ethiotelecom has a challenging task. Features selection has been done based on specific domain knowledge. So, it is difficult to scale and prone to information loss which affects the performance of models. Hence, we have used CNN autoencoder to counterattack such aforementioned challenges. CNN helps to filter representative records even from small datasets and Autoencoder is chosen to detect how much loss is happening. Integrating CNN and autoencoder helps for fraud detection model to overcome challenges.

4.4. Data Collection

Data from Ethiotelecom is being collected to detect subscription fraud based on usage call patterns in CDR. The information is obtained during three months period. CDRs are created by telephone switches on a call-by-call basis and contain all of the information required to define the important components of a phone conversation or other telecommunication transaction, whether from or to a subscriber. It contains the information used by billing systems to rate and bill customers for particular phone calls. The CDR describes a voice data consumption transaction that begins and ends on a subscriber's device. A Field is a property or feature of a CDR instance. It might include information like the phone numbers involved in the conversation, the date and time of the call, the length of the call, the identification of the cell that relayed the call to the subscriber's phone, and other fields. To ensure the success of data preparation, it is required to go deeper into the meaning of the records to obtain a full picture of the data.

Table 2 Dataset Description

no	Name	Description
	CALLING_NUM	the subscriber who initiating the call
1)	CALLED_NUM	identify the destination number
2)	START_TIME	the calling start time
3)	END_TIME	Time of call terminated
4)	DURATION	duration of a call spent
5)	CALL_FEE	How much money is spent per call?

To meet the study's goal, we developed a methodology. As the CDR data in the table above is collected, and we understand the value of data, we choose important variables that might detect subscription fraud activities. Irrelevant data to this study purpose was eliminated from the gathered CDR data for the machine learning system to learn relevant information.

4.4.1 Data Selection

Data selection is a procedure that necessitates domain expertise to choose meaningful features that capture the variability and importance of the data for the target machine learning algorithm to properly learn patterns from the data. Furthermore, it plays an important role in lowering the complexity of the learning process and increasing the efficacy of fraud detection.

4.4.2 Attribute Selection

When creating a predictive model, attribute selection is the process of minimizing the number of input variables. It is preferable to limit the number of input variables to reduce modeling computational costs and, in certain situations, increase model performance. Some predictive modeling issues include a huge number of variables, which can impede model construction and training and necessitate a lot of system memory. The feature chosen to identify the behavior of customers who utilized the service without paying are call duration and call fee. The remaining qualities that are unrelated to the study's goal were removed. Some properties are duplicated or have the same calling number value.

4.5. Class labeling

class is labeled in a supervised learning manner based on call_ fee and call duration attribute values. We created a class column in the last column of our dataset. For the record with 0 value of call fee, we assign fraud class because the caller is not paying the expected payment whereas if the record with greater than 0 value of call fee attribute, we can infer that the caller is legal due to his payment value so, we can assign a normal class label for this case. We can conclude that subscription fraud is related directly to call fees. If the customers called for some duration of time without any payment, it is fraud So, it is labeled as fraud. we used an Imbalanced response variable distribution mechanism since it is a common occurrence for fraud detection where abnormal behavior is rarely observed the data under study often features disproportionate target class distribution.

4.6. Data Preprocessing

To have a data collection appropriate for analysis, data preparation is necessary. Data cleaning to reduce noise or handle missing values, relevance analysis to remove irrelevant or redundant attributes, and data transformation, such as generalizing the data to higher-level concepts or normalizing the data, can all be part of the preprocessing of the data in preparation for classification and prediction [51]. Some selected data may have different formats because the data is very huge, and they stored the data in damp files. Then, in ordered to use the data it needs to convert in to suitable format. The collected data has saved as a CSV file format, which is compatible with the deep learning python code.

Table 3 CDR Voice attributes

calling_num	called_num	start_time	end_time	Duration	call_fee

In theory, a dataset with more characteristics produces better outcomes throughout the learning process. However, in practice, this may not always be the case. There are several features involved with the learning process, some of which may be crucial while others may be irrelevant or redundant. The problem is deciding on a representative set of attributes from which to construct a classification model. As a result, in order to discover relevant characteristics, the dataset must be preprocessed. Furthermore, dimensionality reduction results in a more compact, readily interpretable representation of the goal notion, focusing attention on the most significant elements. [21].

Use training set: The classifier is assessed based on how well it predicts the class of the cases on which it was trained.

Supplied test set: The classifier's performance is measured by how effectively it guesses the class of a collection of cases imported from a file. In this given test experimental approach, the researcher divides the complete dataset into two datasets for training and testing.

Tools used

We installed Anaconda version 3 on Toshiba core i5 laptop computer. We also installed components that help us to implement the deep learning experimentation. Some of the components are numpy, pandas, matplotlib for the sake of handling array of data and visualize the experimental results graphically. Keras and TensorFlow modules were installed on the anaconda in order to obtain deep learning-based auto-encoder and DNN libraries.

4.7. Proposed architecture

In the telecom CDR datasets, we used a 1D Convolutional autoencoder to generate a model with 32 inputs, 16 filtering inputs, and zero padding. In different ideal epochs, the autoencoder model is constructed with a RELU activation function, an Adam optimizer, and a mean-squared-error loss function.

Convolution layer: This is the layer where n number of filters are applied to extract features based on the given size of the kernel.

Batch Normalization: Batch normalization is used to normalize the output of one convolution going as an input to another convolution. This results in efficient training and helps in reducing overfitting.

Dropout layer: This layer is used to reduce overfitting by dropping the specified percentage of features from the model.

Dense layer: This is the final dense layer, also known as the classification layer.

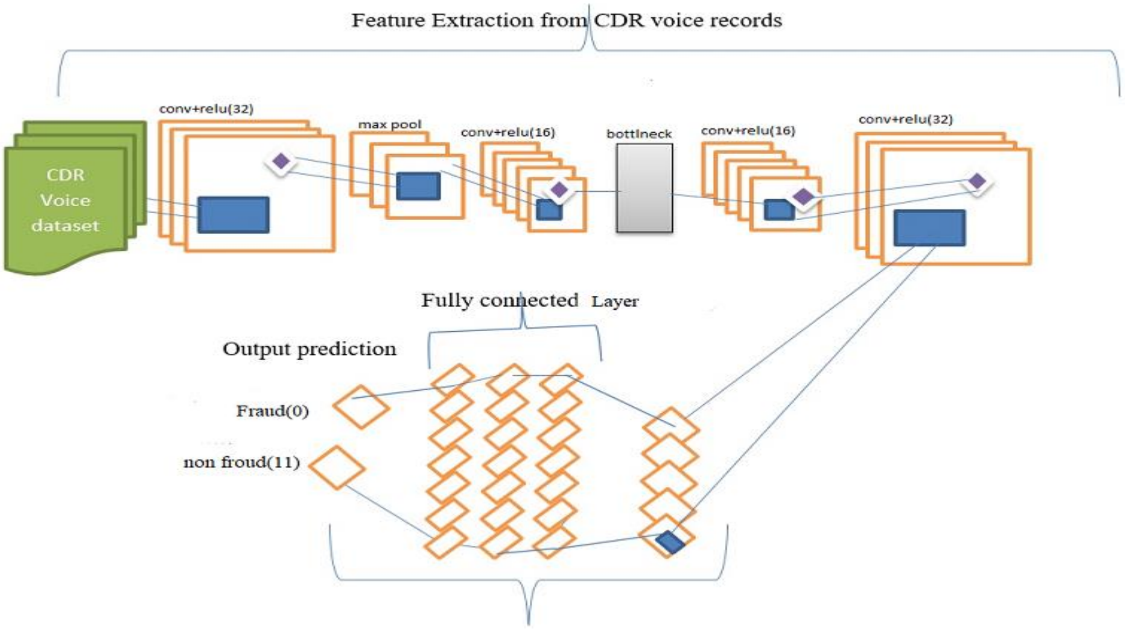


Figure 4: CNN based auto encoder classification

The most significant advantage of deep learning over regular machine learning is that its performance improves as the amount of data increases [31]. A model with seven convolutional layers was created. CDR dataset including call time, call fee, and class characteristics has been provided for the CNN algorithm, as shown in fig 4. The CNN method extracts a representative value from the record and sends it to the autoencoder-based neural network (fully connected layers). The autoencoder receives inputs and compresses the input value before the decoder reconstructs the compressed data as close to the original value as possible. The difference (loss) between the input and output values has been calculated using the mean squared error loss function. Using binary classification algorithms, the neural network determines if the voice record is normal (1) or fraudulent (0).

4.8. Performance Evaluation

Following the training and testing of the algorithms, the next duty is to evaluate the results of the algorithms based on their performance measurement parameters. We assess the performance of the developed fraud detection prototype model in a fair and logical manner, comparing the output to the input. Deep learning-based models may be evaluated using a variety of Performance measure functions, including accuracy, recall, precision, entropy, and F-measure. A confusion matrix, which is a method for examining how effectively an algorithm classifies cases in its classification, may be used to analyze a fraud-based model. The researcher has taken accuracy for measuring the fraud detection model performance. "The simplest intuitive performance metric is accuracy, which is just the ratio of properly predicted observations to total observations (rate of total correct classification). True positives and true negatives are both true "are the observations that are correctly predicted. "A competent classifier reduces the number of false positives and negatives."

True Positives - TP - These are the accurately predicted positive values, indicating that the value of the real class is Fraudulent, as is the value of the anticipated class.

True Negatives - TN - These are the accurately predicted positive values, indicating that the value of the real class is Fraudulent, as is the value of the anticipated class. These are the accurately predicted negative values, which indicate that the value of the actual class is genuine, as is the value of the anticipated class.

False positives and false negatives occur when the actual class contradicts the projected class.

False Positive - FP – When the observed class is negative and the projected class is positive.

False Negative - FN – When the observed class is negative and the projected class is positive. When the observed class is positive while the expected class is negative. We can compute Accuracy after we understand these four factors.,

$$\text{Accuracy} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (1)$$

Eq 4.1 Mathematical Equation [22]

4.9. Work Flow of Proposed Prototype Model

Figure 5 depicts the flow of the proposed prototype system, which accepts CDR as an input. We choose and extract a feature that is relevant to our research and improves the model's performance. The dataset is provided in the same way that it was in the data preparation section. The next job is to build a model using a deep learning methodology, specifically utilizing the auto encoder (AE) method, and then to evaluate the model's performance.

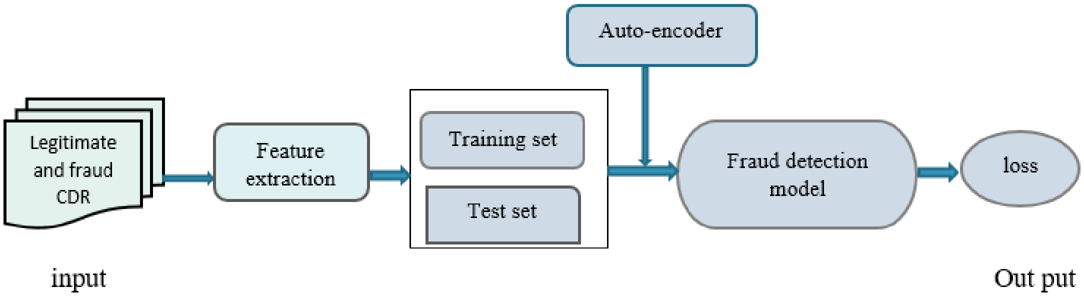


Figure 5: work flow of Prototype model

The detail work flow is discussed as follows:

Step 1: Feature Selection: The efficiency of deep learning is dependent on the usage of an effective collection of characteristics associated with the categories of fraud and regular subscription data. The features chosen are based on a European method to detecting fraud.

Step 2: Feature Extraction: -feature vector which represents words for each instance of a target word, i.e., files of comma-separated values with a file extension of csv.

Step 3: Building Fraud Detection Model and evaluating the Model: - Convolutional auto encoder and deep neural network algorithms has used to build the model. Then the model was evaluated using the test dataset described in the previous sections in order to know how possibly classifying the instances in to their corresponding class.

CHAPTER FIVE

EXPERIMENTATION AND EVALUTION

5.1. Overview

The goal of this research is to develop a model that is the most effective at detecting subscription fraud subscribers. The technique of detecting anomalies in data is known as anomaly detection. Data that deviates considerably from the usual behavior of the data is characterized as abnormal. Anomaly detection may be used for a variety of purposes, including fraud detection, failure detection, and intrusion detection. To meet this goal, features from CDR data were collected and preprocessed. Deep learning is selected for Training and testing. This chapter discusses the experiments, model building, experimental results, and the discussions made during the research. Following the methods, we conduct experimentation on the research questions stated in chapter 1 To address the problem, we collected voice datasets from ethiotelecom, and then we conduct experiments using convolutional autoencoder and DNN fraud detection mechanisms.

5.2. Experimental Setting

Experiments have four categories, experiment I, experiment II and experiment III and experiment IV. Experiment1 were based on voice data with parameters mainly duration and fee. The last column of the dataset was class column. If the value of fee attribute is zero while the calling duration was with value, the record is considered as fraud. The 1048576 voice records were labeled as fraud and non-fraud. At experiment II, after the dataset is prepared in such away, the autoencoder experimentation was undertaken. Then, DNN is also used to detect frauds at this experimentation. Experiment III tried to identify the optimal epoch identification. Finally experiment IV compares methods (algorithms) used in terms of loss, accuracy and running time.

5.3. Data Collection and Data Preparation

The lack of publicly available database has been a limiting factor for the publications on ethiotelecom fraud detection i.e., creating a proper data set for this purpose is very difficult and there are no standard techniques to do this. We collected and organized the voice dataset from ethiotelecom.

Table 4 Dataset for Service type

No	Service type	No of records
1	VOICE	1048576

The total dataset the researcher used for the experimentation of this study was 1048576 data for both training and testing purpose.

5.3.1. System Evaluation

Classification accuracy is defined as the proportion of properly identified items (both normal class '1' and fraudulent class '0') about the dataset. The proportion of successfully identified occurrences is shown in Equation 1 below. The suggested model was assessed using 20% of the test set. In the experiment, we set a zero class for fraudulent subscribers and one class for non-fraudulent subscribers. A classification result has four cases: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) (FN). The measure used to measure performance was Accuracy, which is defined as follows.

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

Equation 1 [22]

5.4. Discussion of Experimental Results

Experiment I: What kind of data features can be used to identify Subscription fraud using convolutional auto-encoder and DNN mechanism?

Because it reduces the number of input variables, feature selection is crucial for every deep learning system. Reducing the input variable helps to reduce the computational cost while also improving the performance of the suggested predictive model. As a result, the duration and call fee was chosen as voice call feature attributes for this study. The voice element's duration is employed as a time stamp, and the fee attribute is used as the payment amount made by subscribers. Call Fee payment values were used to identify fraud utilizing time-series convolutional autoencoder and DNN techniques. Because of the dataset's simplicity, we were able to demonstrate effective fraud detection. As a result, we may conclude that time and call fee are key and most relevant parameters to consider while developing a fraud detection model.

Experiment II: By how much convolutional autoencoder and deep neural network (DNN) methods detect subscription fraud in the case of ethiotelecom?

A dataset is labeled if an annotation exists for each element in the dataset, which determines if it is a normal or anomalous data point. A labeled dataset with normal and anomalous points is the object of supervised fraud detection methods.

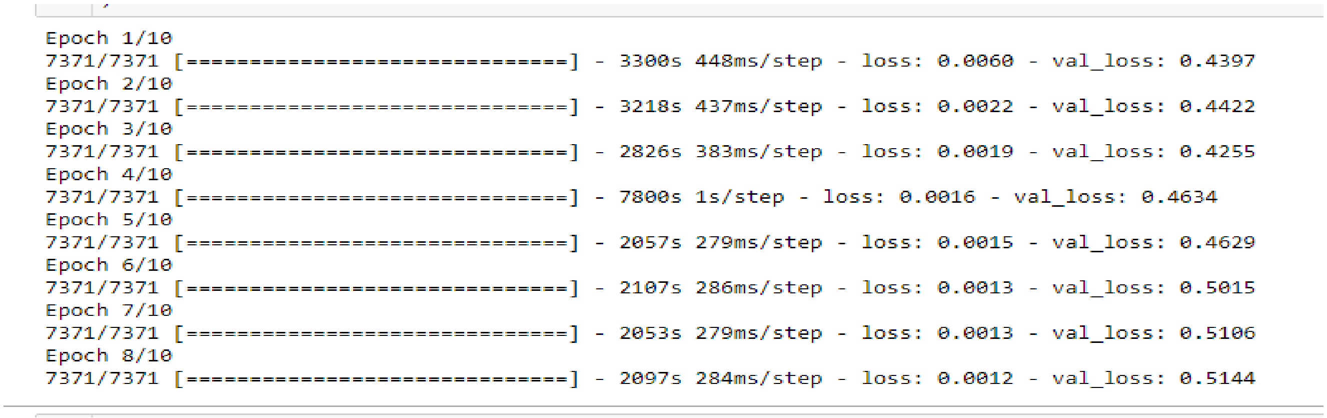
A. Subscription Fraud Using CNN based Auto-encoder

Convolutional autoencoder with 32, 16 filter convolutional layer encoder and 32, 16 convolutional transposed decoder network is employed in our experiment. The reconstruction error is calculated using the Mean squared error between the input and its reconstruction. Then fully connected autoencoder has used for the classification that has an adaptive learning rate starting with 0.0001 and uses Adam Optimization Algorithm for reducing the classification error. We were detecting fraud by determining how well our model can reconstruct the input data as follow.

- Determine the Mean of Auto Encoder loss on training samples.
- Determine the Maximum Mean of Auto Encoder loss value. This is our model's worst sample reconstruction performance ever. This will be the threshold for detecting anomalies.
- If the reconstruction loss for a sample is more than this number, we may conclude that the model is seeing a pattern that it is unfamiliar with. This sample will be labeled as a forgery.

Accuracy and Loss

The loss function indicates how well or badly a predictor classifies input data points in a dataset. When the loss is low, the classifier does a better job of reflecting the relationship between the input data and the output class labels, i.e., classifiers can identify fraud more effectively if the loss becomes too small. Loss is cumulative per epoch. At the beginning of each epoch, loss was 0.0060. For each calculation of the loss, the loss is added to the loss metric on the graph. over time the total loss of the training has decreased to 0.0012. Which indicates the weights of the network are getting more accurate. The more accurate weights provide closer to zero loss value.

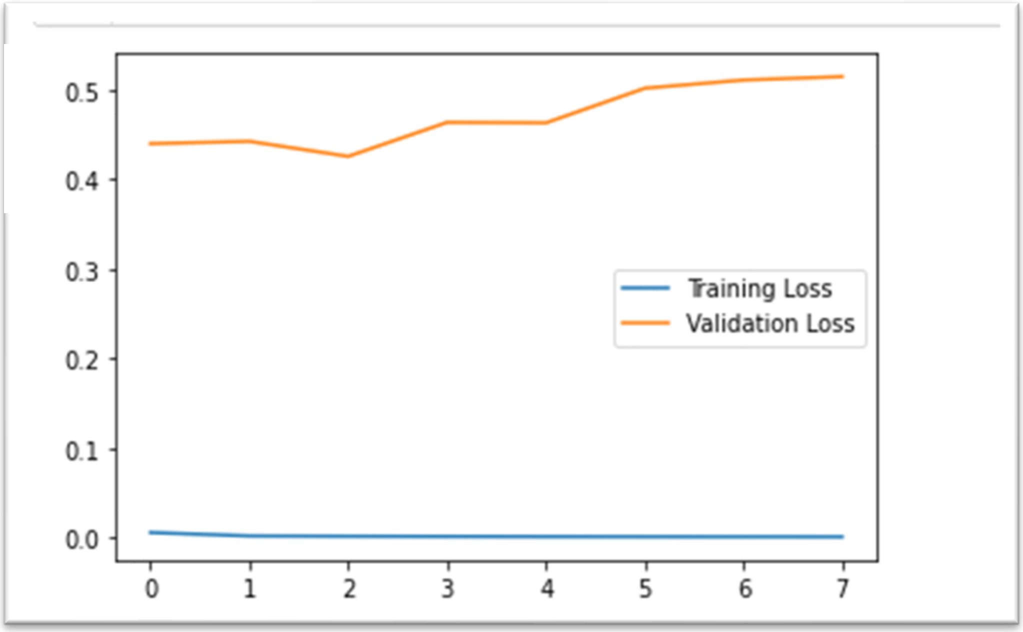


```
Epoch 1/10
7371/7371 [==============================] - 3300s 448ms/step - loss: 0.0060 - val_loss: 0.4397
Epoch 2/10
7371/7371 [==============================] - 3218s 437ms/step - loss: 0.0022 - val_loss: 0.4422
Epoch 3/10
7371/7371 [==============================] - 2826s 383ms/step - loss: 0.0019 - val_loss: 0.4255
Epoch 4/10
7371/7371 [==============================] - 7800s 1s/step - loss: 0.0016 - val_loss: 0.4634
Epoch 5/10
7371/7371 [==============================] - 2057s 279ms/step - loss: 0.0015 - val_loss: 0.4629
Epoch 6/10
7371/7371 [==============================] - 2107s 286ms/step - loss: 0.0013 - val_loss: 0.5015
Epoch 7/10
7371/7371 [==============================] - 2053s 279ms/step - loss: 0.0013 - val_loss: 0.5106
Epoch 8/10
7371/7371 [==============================] - 2097s 284ms/step - loss: 0.0012 - val_loss: 0.5144
```

Figure 6: Loss result using auto encoder

The identification of fraud in non-temporal data, such as spatial data, varies from the detection of anomalies in time-series data. Calculating the divergence of the anomalous points from the rest of the data, for example, is one of the most used ways for finding anomalies in spatial data[39]. Another possibility is to cluster the entire dataset and identify the points in less dense areas as anomalies. This is not true for time-series data. Although the data points in this scenario are not completely independent, it is assumed that the most recent data points in the sequence impact the timestamps of those that follow. Following that, the values in the sequence change in a smooth or predictable manner.

As a result, abrupt changes in the sequence will be seen as fraudulent. To exemplify this behavior, we use the duration points, which exhibit a time-series listing calling duration. When these points are evaluated as separate points, most algorithms will not identify any abnormal behavior, but will find two evenly distributed clusters. To see how the training went, plot the training and validation



loss.

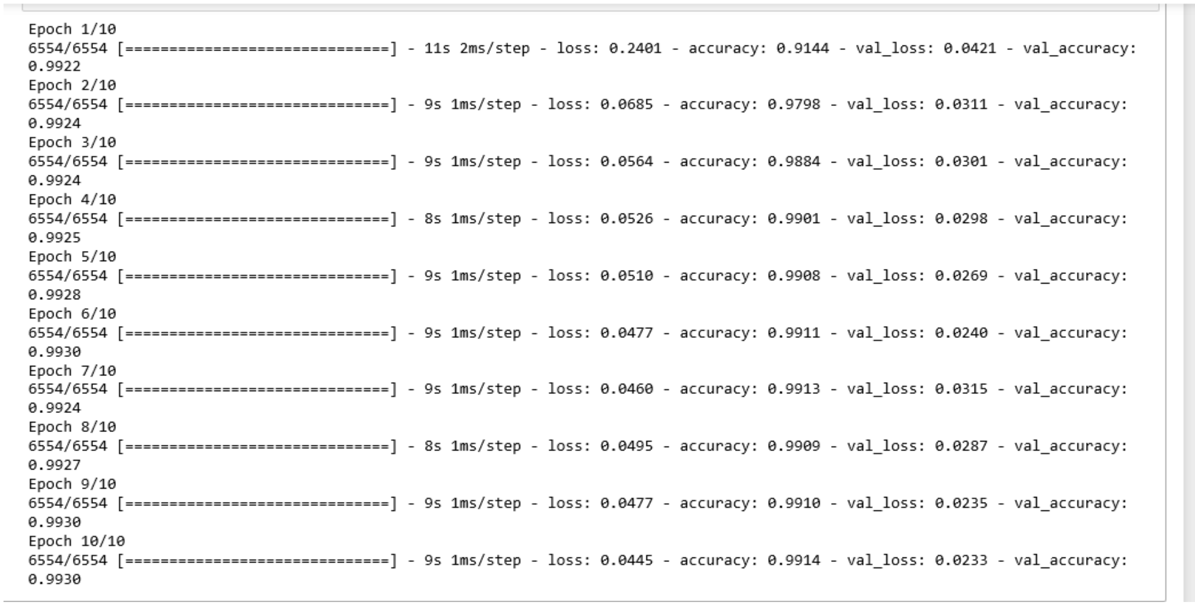
Epoch

Figure 7: training and validation loss

As we can infer from the graph the training loss are decreasing further as the number of epochs reach to 7. The above figure depicts the built model is fitting very nicely the training data but not at all for the validation data, in other words it has not generalizing correctly to the unseen data. Which show some variational on the test data. Perfect training and building 100% performing prototypes are usually not possible.

B. Subscription Fraud Using DNN on Subscriber’s Voice

The best model is obtained from voice dataset as it is seen in Fig 8 below.



```
Epoch 1/10
6554/6554 [==============================] - 11s 2ms/step - loss: 0.2401 - accuracy: 0.9144 - val_loss: 0.0421 - val_accuracy:
0.9922
Epoch 2/10
6554/6554 [==============================] - 9s 1ms/step - loss: 0.0685 - accuracy: 0.9798 - val_loss: 0.0311 - val_accuracy:
0.9924
Epoch 3/10
6554/6554 [==============================] - 9s 1ms/step - loss: 0.0564 - accuracy: 0.9884 - val_loss: 0.0301 - val_accuracy:
0.9924
Epoch 4/10
6554/6554 [==============================] - 8s 1ms/step - loss: 0.0526 - accuracy: 0.9901 - val_loss: 0.0298 - val_accuracy:
0.9925
Epoch 5/10
6554/6554 [==============================] - 9s 1ms/step - loss: 0.0510 - accuracy: 0.9908 - val_loss: 0.0269 - val_accuracy:
0.9928
Epoch 6/10
6554/6554 [==============================] - 9s 1ms/step - loss: 0.0477 - accuracy: 0.9911 - val_loss: 0.0240 - val_accuracy:
0.9930
Epoch 7/10
6554/6554 [==============================] - 9s 1ms/step - loss: 0.0460 - accuracy: 0.9913 - val_loss: 0.0315 - val_accuracy:
0.9924
Epoch 8/10
6554/6554 [==============================] - 8s 1ms/step - loss: 0.0495 - accuracy: 0.9909 - val_loss: 0.0287 - val_accuracy:
0.9927
Epoch 9/10
6554/6554 [==============================] - 9s 1ms/step - loss: 0.0477 - accuracy: 0.9910 - val_loss: 0.0235 - val_accuracy:
0.9930
Epoch 10/10
6554/6554 [==============================] - 9s 1ms/step - loss: 0.0445 - accuracy: 0.9914 - val_loss: 0.0233 - val_accuracy:
0.9930
```

Figure 8: Accuracy result using DNN

In order to make our result clear, we demonstrated in Figure 9: the best accuracy is 99.14%. as it is seen the train accuracy and validation accuracy showed incrementing smoothly. This indicates that the training is not over fitted or under fitted.so that we can conclude that detection of fraud is satisfactory using DNN model.

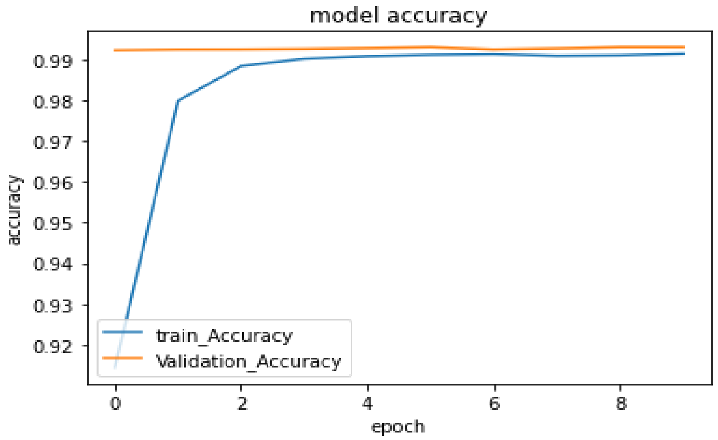


Figure 5.4 Training accuracy and validation accuracy

In Figure 9 below, both the loss and validation loss showed decreasing smoothly. This indicates t hat the training is not over fitted or under fitted.

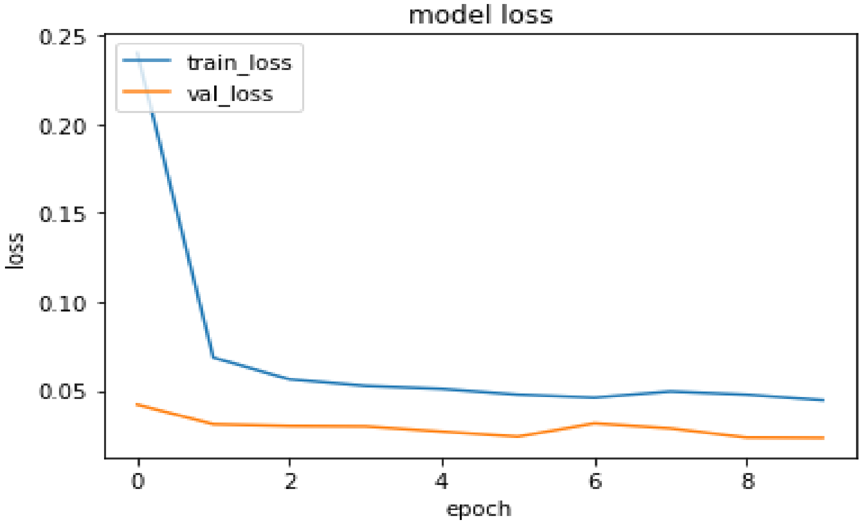


Figure 10: Val loss and Loss

The above figure depicts the built model is fitting very nicely the training data but not at all for the validation data, in other words, it is not generalizing correctly to the unseen data. The model detects fraud 99.14% using epoch 10.

Experiment III. Which epoch becomes most optimal for subscription fraud detection purpose?

Neural networks can learn by modifying the weight distribution in order to approximate a function that is representational of the patterns in the input. The main notion is to re-stimulate the black-box with additional excitation (data) until it achieves a suitably well-structured representation. When the dataset is passed backwards and forward through the whole neural network, it is called one epoch, as after every epoch value of weights assigned is analyzed to make model. The weights are changed, checked, and tested in every cycle for the same dataset simulation. The main memory is keeping the record of all the training data, sometimes it is not possible to keep all the record in main memory, like for larger datasets, so the epoch is brought to memory in divided or batches form, and finally the result is represented as an epoch output. Dealing with epoch is also a challenging task in deep learning[34]. As a result, the dataset is evaluated on several epochal iterations. As a result, we examine several epoch periods to determine the best epoch for voice data. We provided and demonstrated the accuracy mentioned above in table 5 below. The table displays the accuracy and classification results obtained after training with 5, 10, 20, and 60 epochs, respectively.

Table 5 accuracy results using back propagation algorithm

Epoch	Accuracy
5	98.96%
10	99.14%
20	98.88%
60	99.02

As the results shown in Table 5. above when the epoch value increases, the accuracy of system is improved up to epoch 5 to epoch 10. We can see accuracy of the model for the four different epoch values under the same batch size and number of class size fraud and non-fraud.

Therefore, we can conclude that epochs 10 showed the optimal performance for fraud detection model formulation on 1048576 datasets. The reason behind the diminishment of epochs after 10 epochs is that most data set are classified to their category, so classification saturates at pick point until around epoch10 then the classification accuracy decreases since voices are identified and already classified to their class.

Experiment IV. Method (Algorithm) comparison

We tried to include variant algorithms based on their representativeness of algorithmic approaches to our fraud detection task. as we discussed on experiment II, two fraud detection deep learning methods namely CNN based autoencoder and Deep neural network were employed for our detection task. The comparison of these algorithms was undertaken based on the performance and running time taken on the same CDR Dataset. The result in Table 6, shows the performance of algorithms to build a fraud detection model.

Table 6 loss, accuracy and running of methods

Method/algorithm	Convolutional Autoencoders in terms of loss	DNN in terms of accuracy
Performance	0.0012	99.14%
Running time taken in sec.	2097	9

Both Convolutional Auto-Encoders and DNN perform in a promising way because convolution autoencoders performs 0.0012 which is almost zero loss. We also accomplish the fraud detection experiment using deep neural network which scores 99.14%.

DNN is the fastest algorithm of both tested algorithms in the running time of dataset took 9 seconds. Although Convolutional auto Encoders performed well, it took relatively very low speed than DNN. The running time taken was 2097 seconds on same CDR Dataset. Therefore, we can conclude that deep neural network is a promising method to implement for Ethiotelecom subscription fraud detection task considering its fast-running time benefit.

CHAPTER SIX

CONCLUSION AND FUTURE WORK

6.1. Conclusion

In the case of ethiotelecom, we employed the convolutional auto-encoder and DNN fraud detection techniques. The voice dataset allows for the evaluation of supervised fraud detection methods to identify users who use the call service unlawfully and do not pay the expected cost. On the acquired voice dataset, we carefully tested two the state-of-the-art approaches for fraud detection. The assessments are meant to serve as a starting point for the development of future techniques. Our results show that discriminative approaches that leverage descriptors of retrained networks outperform methods that on the fraud-free training data, learn feature representations from scratch. We have provided information based on call duration and call fee values. In this study, hence we identified very valuable features for the fraud detection's sake. we have used common evaluation metrics which is accuracy to evacuate the formulated fraud detection model. Based on our evaluation, the Convolutional auto encoder mechanism scores the minimum loss of 0.0012 and the DNN performance was 99.14% which is promising performance to detect frauds for ethiotelecom. When we compare the state of the arts performance result on fraud detection, the DNN performance result was better and much promising for fraud detection. We tested the performance of the model with varying the number of epochs over the same dataset. Initially during the training process when both the training set and validation set are being learnt by the model. After epoch 10, the performance of the model showed diminishments. So, we concluded that epoch 10 is optimal over 1048576 datasets.

6.2. Contribution

We made the following contributions to this study.Provided insight on subscription fraud detection theoretical perspectives and practical methodologies that allow telecom carriers to identify fake subscriptions. Raising awareness about subscription fraud and demonstrating the limits of Customer Relationship Management (CRM) system approaches for fraud detection. Subscription fraud subscriber behavior was investigated, and relevant components of CDR data, notably call time and service charge, were offered.

To the best of our knowledge, many studies has been conducted with data mining and machine learning algorithms on subscription fraud detection in the case of Ethiotelecom.. In this study We employed a convolutional autoencoder and a deep neural network to detect subscription fraud in the case Ethiotelecom voice call data. We developed a promising model for detecting fraud that help operators to detecting subscription fraud early as possible.

6.3. Future Work

The fraud detection task shall better be tested using generative adversarial network (GAN) because GAN shows good performance for anomaly detection today. Telecom needs not only detection mechanisms but also studies are recommended on mitigating those frauds. This paper was conducted on voice dataset, other researches are needed to consider Internet data and SMS datasets to develop fraud detection model.

.

REFERENCES

[1] S. Qayyum, S. Mansoor, A. Khalid, Khushbakht, Z. Halim, and A. R. Baig, "Fraudulent call detection for mobile networks," *2010 Int. Conf. Inf. Emerg. Technol. ICIET 2010*, 2010, doi: 10.1109/ICIET.2010.5625718.

[2] M. Arafat, A. Qusef, and G. Sammour, "Detection of Wangiri Telecommunication Fraud Using Ensemble Learning," *2019 IEEE Jordan Int. Jt. Conf. Electr. Eng. Inf. Technol. JEEIT 2019 - Proc.*, pp. 330–335, 2019, doi: 10.1109/JEEIT.2019.8717528.

[3] S. Rosset, U. Murad, E. Neumann, Y. Idan, and G. Pinkas, "Discovery of fraud rules for telecommunications---challenges and solutions," pp. 409–413, 1999, doi: 10.1145/312129.312303.

[4] J. Howell, "Fraud Loss Survey CFCA 2019 Fraud Loss Survey A message from the Survey Chairman," no. November, 2019.

[5] S. Wu, N. Kang, and L. Yang, "Fraudulent Behavior Forecast in Telecom Industry Based on Data Mining Technology," *Commun. IIMA*, vol. 7, no. 4, pp. 5–10, 2007, [Online]. Available: http://iima.org/CIIMA/4 CIIMA 7-4-07 Sen 1-6.pdf.

[6] 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, doi: 10.1198/TECH.2009.08136.

[7] H. Farvaresh and M. M. Sepehri, "A data mining framework for detecting subscription fraud in telecommunication," *Eng. Appl. Artif. Intell.*, vol. 24, no. 1, pp. 182–194, 2011, doi: 10.1016/j.engappai.2010.05.009.

[8] Y. A. Ayiza, "Identifying the Reason for Mobile Call," no. November, pp. 1–102, 2018.

[9] M. I. Akhter and M. G. Ahamad, "Detecting Telecommunication Fraud using Neural Networks through Data Mining," *Int. J. Sci. Eng. Res.*, vol. 3, no. 3, pp. 1–5, 2012, [Online]. Available: http://www.ijser.org.

[10] L.Deyu, "data mining approach to analyze mobile telecommunications network quality of service: the case of ethio-telecom" May, 2014 [11] A.A. Ababa. (Mar. 6, 2017). Ethiopia-telecom fraud. A.-A. Ababa, Ed., [Online]. Available: http://apanews.net/en/news/ethiopia-loses-over-52mto-telecom-fraud-official.

[12] fanabc. (Oct. 1, 2019). Telecom fraud. fanabc, Ed., [Online]. Available: https: / / www . fanabc . com / english / 2018 / 10 / ethio - telecom - mulls - over - preventing-telecom-fraud.

[13] F. Tesfaye and A. Ababa, “Near-Real Time SIM-box Fraud Detec- tion Using Machine Learning in the case of ethio telecom,” 2020.

[14] N. Ketkar, “Introduction to Keras,” *Deep Learn. with Python*, pp. 97–111, 2017, doi: 10.1007/978-1-4842-2766-4_7.

[15] P. Hoath, “Telecoms fraud, the gory details,” *Comput. Fraud Secur.*, vol. 1998, no. 1, pp. 10–14, 1998, doi: 10.1016/s1361-3723(97)82712-1.

[16] W. Bogale, “Ethiopian Telecommunications Corporation A Background Paper on Telecom & Telecom Statistics in Ethiopia,” *Regulation*, 2005.

[17] C. Telecommunication, D. Officer, and E. T. Agency, “THE REGULATORY AND INSTITUTIONAL DIMENSION By Ethiopian Telecommunication Agency,” no. March, pp. 17–19, 2010.

[18] Q. Zhao, K. Chen, T. Li, Y. Yang, and X. F. Wang, “Detecting telecommunication fraud by understanding the contents of a call,” *Cybersecurity*, vol. 1, no. 1, pp. 1–12, 2018, doi: 10.1186/s42400-018-0008-5.

[19] T. Fawcett and F. Provost, “Adaptive fraud detection,” *Data Min. Knowl. Discov.*, vol. 1, no. 3, pp. 291–316, 1997, doi: 10.1023/A:1009700419189.

[20] C. M. Held, C. A. Perez, and P. A. Este, “Subscription fraud prevention in telecommunications using fuzzy rules and neural networks - Held, Perez, Este - 2001.pdf,” 2001.

[21] C. S. Hilas and P. A. Mastorocostas, “An application of supervised and unsupervised learning approaches to telecommunications fraud detection,” *Knowledge-Based Syst.*, vol. 21, no. 7, pp. 721–726, 2008, doi: 10.1016/j.knosys.2008.03.026.

[22] T. Fulfillment and T. Engineering, “A Comparative Analysis of Machine Learning Algorithms for Subscription fraud Detection : The case of ethio telecom,” 2020.

[23] K. Hagos, “SIM-Box Fraud Detection Using Data Mining Techniques : The Case of ethio telecom,”p.84,2018,[Online].Available: http://etd.aau.edu.et/handle/123456789/15238?show=full.

[24] G. Maciá-fernández, “Roaming fraud : assault and defense strategies,” vol. 241000, pp. 1–8, 2008.

[25] M. Sahin, A. Francillon, P. Gupta, and M. Ahamad, “SoK: Fraud in Telephony Networks,” *Proc. - 2nd IEEE Eur. Symp. Secur. Privacy, EuroS P 2017*, pp. 235–240, 2017, doi: 10.1109/EuroSP.2017.40.

[26] Shankar, “Subscription Fraud,” [Online]. Available: http://frslabs.com/frsblog/2014/12/19/subscription-fraud-control-can-control-fraud-losses/.

[27] A. L. Caterini, “A Novel Mathematical Framework for the Analysis of Neural Networks,” 2017, [Online]. Available: https://uwspace.uwaterloo.ca/handle/10012/12173.

[28] J. Hollmén, *User profiling and classification for fraud detection in mobile communications networks*, no. 109. 2000.

[29] P. Sanò *et al.*, “The passive microwave Neural network Precipitation Retrieval (PNPR) algorithm for AMSU/MHS observations: Description and application to European case studies,” *Atmos. Meas. Tech.*, vol. 8, no. 2, pp. 837–857, 2015, doi: 10.5194/amt-8-837-2015.

[30] A. Adler, D. Boublil, M. Elad, and M. Zibulevsky, “A DEEP LEARNING APPROACH TO BLOCK-BASED COMPRESSED SENSING OF IMAGES Computer Science Department , Technion , Haifa 32000 , Israel Electrical Engineering Department , Technion , Haifa 32000 , Israel,” *Icassp*, no. 320649, pp. 2–6, 2017.

[31] J. R. Dorronsoro, F. Ginel, C. Sánchez, and C. Santa Cruz, “Neural fraud detection in credit card operations,” *IEEE Trans. Neural Networks*, vol. 8, no. 4, pp. 827–834, 1997, doi: 10.1109/72.595879.

[32] M. E. Aminanto, R. Choi, H. C. Tanuwidjaja, P. D. Yoo, and K. Kim, “Deep abstraction and weighted feature selection for Wi-Fi impersonation detection,” *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 3, pp. 621–636, 2017, doi: 10.1109/TIFS.2017.2762828.

[33] P. Sharma and S. Pote, “Credit Card Fraud Detection using Deep Learning based on Neural Network and Auto encoder,” *Int. J. Eng. Adv. Technol.*, vol. 9, no. 5, pp. 1140–1143, 2020, doi: 10.35940/ijeat.e9934.069520.

[34] A. Mathew, P. Amudha, and S. Sivakumari, “Deep learning techniques: an overview,” *Adv. Intell. Syst. Comput.*, vol. 1141, no. August 2020, pp. 599–608, 2021, doi: 10.1007/978-981-15-3383-9_54.

[35] M. Sanver and A. Karahoca, “Fraud detection using an adaptive neuro-fuzzy inference system in mobile telecommunication networks,” *J. Mult. Log. Soft Comput.*, vol. 15, no. 2–3, pp. 155–179, 2009, doi: 10.1142/9789812709677_0205.

[36] A. Pumsirirat and L. Yan, “Credit card fraud detection using deep learning based on auto-encoder and restricted Boltzmann machine,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 1, pp. 18–25, 2018, doi: 10.14569/IJACSA.2018.090103.

[37] D. Meyer, “Introduction to Autoencoders,” pp. 1–8, 2014, [Online]. Available: http://www.1-4-5.net/~dmm/ml/ae.pdf.

[38] A. M. Ozbayoglu, M. U. Gudelek, and O. B. Sezer, “Deep learning for financial applications: A survey,” *Appl. Soft Comput. J.*, vol. 93, no. May, 2020, doi: 10.1016/j.asoc.2020.106384.

[39] M. Braei and S. Wagner, “Anomaly Detection in Univariate Time-series: A Survey on the State-of-the-Art,” 2020, [Online]. Available: http://arxiv.org/abs/2004.00433.

[40] A. M. Ceylan and V. Aytaç, “Convolutional auto encoders for sentence representation generation,” *Turkish J. Electr. Eng. Comput. Sci.*, vol. 28, no. 2, pp. 1135–1148, 2020, doi: 10.3906/elk-1907-13.

APPENDIX

Auto-encoder code written by python programming

```
importnumpy as np
import pandas as pd
frommatplotlib import pyplot as plt
fromtensorflow import keras
fromtensorflow.keras import layers


df_small_noise_url_suffix = "C:/Users/ET/Voice.csv"
df_small_noise_url = df_small_noise_url_suffix
df_small_noise = pd.read_csv(
df_small_noise_url, parse_dates=True, index_col="Duration"
)

df_daily_jumpsup_url_suffix = "C:/Users/ET/Voicetest.csv"
df_daily_jumpsup_url=  df_daily_jumpsup_url_suffix
df_daily_jumpsup = pd.read_csv(
df_daily_jumpsup_url, parse_dates=True, index_col="Duration"
)

print(df_small_noise.head())
print(df_daily_jumpsup.head())

fig, ax = plt.subplots()
df_small_noise.plot(legend=True, ax=ax)
plt.show()

fig, ax = plt.subplots()
df_daily_jumpsup.plot(legend=True, ax=ax)
plt.show()

# Normalize and save the mean and std we get,
# for normalizing test data.
training_mean = df_small_noise.mean()
training_std = df_small_noise.std()
df_training_value = (df_small_noise - training_mean) / training_std
print("Number of training samples:", len(df_training_value))
```

```
TIME_STEPS = 288

# Generated training sequences for use in the model.
defcreate_sequences(values, time_steps=TIME_STEPS):
output = []
for i in range(len(values) - time_steps):
output.append(values[i : (i + time_steps)])
returnnp.stack(output)
x_train = create_sequences(df_training_value.values)
print("Training input shape: ", x_train.shape)

model = keras.Sequential(
    [
layers.Input(shape=(x_train.shape[1], x_train.shape[2])),
layers.Conv1D(
filters=32, kernel_size=7, padding="same", strides=2, activation="relu"
        ),
layers.Dropout(rate=0.2),
layers.Conv1D(
filters=16, kernel_size=7, padding="same", strides=2, activation="relu"
        ),
layers.Conv1DTranspose(
filters=16, kernel_size=7, padding="same", strides=2, activation="relu"
        ),
layers.Dropout(rate=0.2),
layers.Conv1DTranspose(
filters=32, kernel_size=7, padding="same", strides=2, activation="relu"
        ),
layers.Conv1DTranspose(filters=1, kernel_size=7, padding="same"),
    ]
)
model.compile(optimizer=keras.optimizers.Adam(learning_rate=0.001), loss="mse")
model.summary()

history = model.fit(
x_train,
x_train,
epochs=10,
batch_size=128,
validation_split=0.2,
callbacks=[
keras.callbacks.EarlyStopping(monitor="val_loss", patience=5, mode="min")
    ],
)



plt.plot(history.history["loss"], label="Training Loss")
plt.plot(history.history["val_loss"], label="Validation Loss")
plt.legend()
plt.show()
```