IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# The Role of Data Science in Detection of Fraud and Intrusion over a Network

Umar Mohammed Abatcha[1] , Umar Sule[2], Muhammad Auwal Ahmad[3] , Ouedraogo Pengdwende Leonel Camille[4], Alemnge Nadine Lekeate[5]

[1]Department of Computer Science and Engineering, Sharda University, India

[2]Department of Computer Science, Umar Suleiman College of Education Gashua, Yobe State

[3]Department of Computer Science, Federal University Gashua, Nigeria.

[4]Department of Computer Science and Engineering, Sharda University, India

[5]Department of Computer Science and Engineering, Sharda University, India

## Abstract

There have been several cases of terrorism, fraud and intrusion which are always happening on the internet. These cases are threatening the security and privacy of millions of individuals, companies, government, military and other authorities when the virtual world is of course in thirsty of strong security as well as methods of tackling intrusion over a network. This academic paper comes with some highlights to emphasize the role of data science in detecting intrusion, fraud and other hypocritical activities over a network.

***Keywords: Fraud, Intrusion, Hypocritical Communication, Digital Forensic, Artificial Intelligence, Biometric System.***

## Introduction

Data science has been a big field where data are analyzed, categorized, managed and manipulated in such a way that traditional operations cannot handle. It simply deals with the modern contemporaries in analytics of large amount of data as well as predicting the future based on the previously studied data. Data science is undoubtedly a modern field of research which comprises the use of mathematical modelling such as statistics, calculus, probability, graph & linear algebra, machine learning, data analytics, programming languages and database manipulation.

Data Science has numerous applications in computer sciences, statistics, accounting, businesses, Government and many more, where it constantly expands and penetrates through the private and governmental sectors. One of the great applications of data science is the use of its techniques to detect and thwart hypocritical communications and activities such as fraud, terrorism and intrusion over a network.

The world has been recording the cases of these hypocritical communications and cybercrime over a long period of time. In the current scenario, cybercrime is increasing very fast as the technology is growing very fast (Navneet (2018) where cybercriminals have threatened victims with all sorts of fraud techniques.

According to description from the Federal Trade Commission (FTC), "millennial are uniquely more sensitive to networked deception than seniors, as appalling as it may seem.” The investigation boasts that 40 percent of adults aged 20-29 who have reported fraud ended up losing money in a fraud case. Hence, from the same research, it was reported that 1 in 5 people lost their money where 328 million US dollars has reportedly been lost. Furthermore, same research has identified 23% credit card fraud; adding that, of the 3 million identity theft and fraud reports received in 2018, 1.4 million US dollars were fraud-related, and 25 percent of those cases noted as loss of money. In 2018, consumers reported losing about 1.48 billion US dollars linked to fraud complaints, an increase of 406 million US dollars from 2017. According to Statista Research Department, companies in the United States experience an annual loss of more than 525 million US dollars due to cybercrime.

Internet fraud leads to economic decline in a country. Because this act not only defrauds individuals, but also state-owned companies. For example, in 2006 the United States recorded a loss of US$198.4 million due to internet fraud (Rosenberg, 2007), while in the same year the United Kingdom was reported to be losing up to £150 million annually (BBC News, 2006 ). . Another report from the National Fraud Authority states that fraud costs the UK £73 billion a year. As a result, US-based security firm McAfee estimates that cybercrime currently costs the international economy approximately $600 billion, or 0.8% of global gross domestic product. The Federal Bureau Investigation reported that the total cost of insurance fraud in the United States is estimated at more than $40 billion a year. Despite hundreds of detection systems the world of technology has, internet fraud and intrusion are still increasing. These challenges threaten people from purchasing online, participating in online trade and sometimes even prevent them from almost all the internet activities. Experts suggest that the world of technology needs to develop perfect human-like detection systems with the use of modern technologies to make internet best and safer place.

In a nutshell, the aims of this research aim to;

I. Emphasize the role of data science and related fields in detection of fraud, terrorism and intrusion over a network;
II. Pinpoint some techniques to be used to detect and thwart all hypocritical communications over a network, and at last;
III. Hearten the use of modern scientific breakthroughs to build human-like detection systems.

**Understanding Data Science in Detection of Fraud and Intrusions**

Data science as a scientific discipline is influenced by informatics, computer science, mathematics, operations research, and statistics as well as the applied sciences (Weihs & Ickstadt, 2018). Data science in the world of technology and computer science is considered a science of data which deals with the use scientific methods, strategies and techniques to extract and highlight useful insights from a data set. These useful insights could be an image (or information in it), text, video and other data elements. The information of these data could be extracted using different data science techniques and suggest useful insights in them. The data could be sourced from different data fields like unstructured data such as satellite images, scientific data, photographs or videos; and structured data such as social media data, mobile network, website content and text or documents.

The main focus on data science is structured data which happens to be human daily life activity more especially on the internet. According to Forbes, over 90% of all the data in the world was created in the past few years which resulted the beginning of Big Data era. Domo reported, there are 2.5 quintillion bytes of data created each day at the current pace. According to Visual Capitalist 2019 Research, every minute 188 million emails are sent, 4.5 million videos are watched on YouTube, Facebook holds 1 million logins, total of 41.6 million messages are sent on WhatsApp and Facebook Messenger, 3.8 million searches are processed on Google Search Engine, 347,222 scrolls are done on Instagram, 2.1 million snaps are created, 18.1 million text messages are sent, and almost million dollars are spent online.

Hence, according to Domo’s Data Never Sleeps 5.0 report, every minute 527,760 photos are shared on Snapchat, more than 120 professionals join LinkedIn, 465,000 tweets are sent on Twitter, and 46,740 photos are shared on Instagram. According to Facebook Investor, 1.73 billion people are active on Facebook daily. Zephoria also reported, on Facebook more than 300 million photos are uploaded per day, 510,000 comments are sent and 293,000 statuses are updated every minute. According to InternetLiveStats, Google alone processes an average of 40,000 searches per second, making it over 3.5 billion searches per day and 1.2 trillion searches per year. WhatsApp showed a huge adding in dealing with messages, from handling two billion messages in April 2012 to deal with ten billion information every day, in August 2012 (Iyobor et al., 2020).

The analysis of these huge data flowing on the internet every second, then Big Data technology pop into existence to handle huge amount of data which local data storage facilities could not. Big Data focuses on high volume, high velocity, high veracity and high variety. So, this huge amount of data that are generated by billions of internet users across the globe could not be sighted without the use of modern technologies such as data science and Machine Learning to detect the kind of communication people make online. Furthermore, the excessive rise of intruders, attackers and fraudsters would be tackled to prevent their bad intentions towards harming innocent people. In this technological era, the use of advanced technologies like machine learning becomes common in almost all developed fields. Machine Learning is a subfield of Artificial Intelligence technology that deals with the use of algorithms, processes and other mathematical and scientific techniques to make machines capable of learning, understanding, and detecting objects and information.

Today, many tech giants around the world are investing heavily in the development of technologies such as artificial intelligence, machine learning, and the Internet of Things. This is related to the fact that the amount of data being generated is extraordinarily high, about 20 billion connected devices by 2020, and we will see a trillion connected devices and things by 2050. Machine learning is the intention that there are generic processes that can predict something interesting about a set of information without the desire to write custom code specific to the problem. Machine learning is an application of artificial intelligence (AI) that gives structure the ability to automatically study and improve from experience without being distinctly programmed. Machine learning emphasizes the growth of computer programs that can access data and use it for themselves (Chauhan et al. 2020).

Data Science Techniques Used in Detection of Fraud and Intrusions

As banking is one of the biggest applications of Data Science and with the advancements in Machine Learning, It has become easier for businesses to detect fraud and irregularities in transaction patterns. Fraud detection involves monitoring and analyzing user activity to find common or malicious patterns. With the increasing reliance on the internet and e-commerce for transactions, the number of fraud cases has increased significantly. So, Data Science techniques help in extracting useful insights in data in such a way that understanding those useful insights becomes easier, as it has been playing a key role in automating various financial tasks.

As explained Machine Learning earlier, its algorithms would help Data Science in numerous fields to detect fraud and intrusion. Machine learning helps data scientists efficiently determine which the process are most likely to be counterfeit while originally reducing false positives. The methods are exceptionally effective in fraud prevention and detection, as they allow for the automated detection of patterns in large volumes of streaming transactions. Machine learning models for fraud detection can also be applied to develop predictive and descriptive analytics. Predictive analytics offers a distinct method of fraud detection by analyzing data with a pre-trained algorithm to score a transaction on its fraud riskiness; descriptive analytics uses an algorithm to analyze historical data to answer what has happened till now. Both predictive and descriptive analytics require the same data and training to implement.

Decision Tree: According to Wikipedia, decision tree learning is one of the predictive modeling techniques used in statistics, data mining, and machine learning. It uses a decision tree (as a predictive design) to go from measurement about an idea (represented in the division) to the conclusion about the item's target cost. The decision tree algorithm belongs to the class of supervised studying algorithms. The goal of using a decision tree is to formulate a training model that can be used to predict the class or significance of the target variable by learning simple decision rules derived from previous (training) data. The decision tree algorithm with its partners like regression and classification helps to predict the existence of an event like the next action of a scammer or an intruder by evaluating its previous datasets.

Logistic Regression: logistic regression is a statistical design that has the basic form uses a logistic function to model a binary dependent variable. For example, to call whether the email is spam: 1 or 0. Logistic regression is a predictive analysis, it is used to describe data and explain the relationship between a dependent binary variable and one or more independent variables at a nominal, ordinal, interval, or ratio level. In this case, the data science algorithm helps detect the intrusion where applicable. For example, detecting a fake scam alert and fraudulent online activity might be easier than expected.

Artificial Neural Networks: Artificial Neural Networks (ANNs) are biologically inspired Computational networks (Park & Lek, 2016). Artificial Neural Networks re-produce the complex mathematical equations with summations, exponentials, and parameters to copy neurons (Berry et al. 2000). Artificial Neural Networks are a technology based on studies of the brain and nervous system, as they emulate a biological neural network but they use a reduced set of concepts from biological neural systems (Walczak & Cerpa, 2003). ANNs have been applied in ecology to describe, for instance, the probability of occurrence, species distribution, and abundance (Echelpoel & Goethals, 2015).

The Artificial Neural Networks is strength to learn so rapidly that makes them so impressive and useful for a collection of tasks. Neural Networks study things in exactly in same way as the brain. An artificial neuron is a mathematical function conceived as a design of biological neurons, a neural network. They have been applied to classify crime instances such as burglary, sexual offences, and known criminals' facial characteristics (Mena et al. 2003). Artificial Neural Networks algorithm will be more helpful to identify and recognize recent activities and to even the predict future. It has been in use in many tech giants such as banking and finance to detect the authenticity of a user. For example, in the bank, it helps to identify the authenticity of a customer by analyzing their activities.

**Gradient Boosting Classifier:** Gradient enhancement classifiers are a family of machine learning algorithms that combine many weak learning models to formulate a powerful predictive model. Gradient Boosting is one of the most powerful techniques for building predictive models. It is an iterative functional gradient algorithm, i. H. an algorithm that minimizes a loss function by iteratively choosing a function that points to the negative gradient; a weak hypothesis. Gradient Boosting, the most important component among the three, involves the loss function. Your role is to evaluate how good the model is at making predictions given the listed data.

### How Data Science Plays a Role in Forensic Science

According to The United State Department of Justice, forensics is described as a critical component of the criminal justice system; In addition, forensic scientists examine and analyze facts from crime scenes and elsewhere to develop objective findings that can aid in the investigation and prosecution of criminal offenders or clear an innocent person of suspicion. Henceforth, forensic science means applying scientific methods, processes and techniques to solve crimes. It has become an essential part of the judicial system with its focus on recognition, identification, and evaluation of physical evidences. It utilizes a broad spectrum of science to analyze relevant information to crime and legal evidence. Forensic scientists collect, preserve and analyze scientific evidence and information during the course of investigation to ensure justice extract useful insights. Digital Forensic investigation is done by examining documents, digital media, fingerprinting and autopsy techniques.

These two disciplines having almost same concept and role in extracting useful insights play an important role in detection of fraud, exploitation of intrusions and illegal activity as well as unveiling hypocritical communications and unauthorized accesses over a network. Forensic science has helped in unveiling many culprits and exploiting their heartless deeds. It is undoubtedly playing a significant role in the law system.

According to Statista Research Department, 652,676 women were raped or sexually assaulted in the United States, while the corresponding number of males was 81,956. Digital forensic investigation will become a helpful method of tackling such crimes in the world. In order to investigate rape and murder cases, the investigative and forensic team need to analyze the undisturbed crime scene soon after the incident and collect evidence by getting the cell phone of victim/suspect to analyze call records; to link between victim, crime scene and potential suspect for successful investigation (Pratihari et al., 2019).

Hence, in order to make forensic science's performance better, the advanced knowledge and techniques of data science should be declared upon it. For example, (Pratihari et al., 2019) suggested collection of evidence during forensic investigation by getting the cellphone of victim or suspect to analyze call records. To advance the use of this method, telecommunication and social media companies should implement the use of data science techniques which will predict suspect's next action by analyzing their previously sent data. This method will give an opportunity to detect such criminal activities over the internet before their occurrence.

### Other Methods of Tackling Hypocritical Actions over a Network

### Advanced Intrusion Detection Systems

The Intrusion Detection System monitors network and system traffic for suspicious activity. Once potential threats are identified, the system sends notifications to the company. Among the best intrusion detection systems is the network-based intrusion detection system, which examines and analyzes network traffic. A network-based intrusion detection system must have a packet sniffer, which by default captures network traffic. These advanced systems typically help detect not only intrusions, but also fraud, unauthorized access, and/or illegitimate communications or activities on a network. Today, all developed countries have approached this unavoidable development. In those counties, a fraudster or an intruder would be detected in a short period of time using their modern detection machines with artificial intelligence capability (Ahmad, 2020).

Furthermore, (Rao & K.C. Roy, 2019) have emphasized the technique of encrypting data for a better and secure access over a network, and to help intrusion detection systems to detect unauthorized access. There has been a need of secure transmission and storage of data to protect it from unauthorized access. Encryption is one of the common techniques to validate image security. Encryption of images and videos has a very wide application in various field which includes internet connection, multimedia systems, and the industrial process, for transmitting medical images, telecommunication and military communication, legal images that could contain a lot of confidential information. In the previous times vector quantization was used for the protection of images as an image encryption technique (Rao & K.C. Roy, 2019).

### Modern Biometric Security System

The use of modern biometric security systems will also reduce security challenges and cyber-attacks. (Choudhary, 2012) explains biometrics as the science and technology of measuring and analyzing biological data. A biometric system is basically a pattern recognition system that recognizes an individual based on a set of characteristics derived from specific physiological or behavioral traits that the individual possesses (Prabhakar et al., 2003). These traditional methods are unreliable as keys and cards can be lost or stolen, and passwords can be compromised, forged or hacked

(Omidiora, 2006; Falohun, 2012). Therefore, biometric systems have been used in various applications (Kim et al., 2012). Biometric authentication has the advantage over traditional techniques that it cannot be stolen, forged or forgotten (Adedeji et al., 2018).

In Nigeria now, no one is allowed to have more one bank verification numbers. Individuals' bank verification numbers remain single and constant even when having multiple bank accounts. This happen that whenever someone needs to open a new bank account, their bank verification numbers will be shown by verifying using their fingerprint. In order to implement the same system in the Nigerian security system, citizens must register a new modern national identity card, in which a modern biometric security system will be implemented, in which the fingerprint and facial recognition system of people can be used (Ahmad, 2020).Hence, (Adedeji et al., 2018; Omidiora, 2006; and Falohun, 2012) have emphasized the use of biometric system to recognize, authenticate and identify people not only in security but for preventing loss or stolen of data. Whereas (Ahmad, 2020) highlights the use of modern security system to merge national identity card with bank verification numbers in which exploiting the information can be done by authority only. Furthermore, with the excessive use of this method, fraud and intrusion could easily be detected on a network.

Artificial Intelligence

Technologically advanced countries like the United States, China and the United Arab Emirates have been using modern detection systems to thwart illegitimate activities over a network. Those advanced detection systems include Artificial Intelligence CCTV cameras, for example Facebook's Jarvis and other intelligent systems. Artificial intelligence is redefining the way companies do fraud detection today. The SPD group posits machine learning fraud detection due to the machine learning method's ability to learn from historical fraud patterns and recognize them in future transactions. Machine learning algorithms seem to be more efficient than humans when it comes to data processing speed. So, artificial intelligence fraud detection helps to complete the data analysis within milliseconds and detect complex patterns that are difficult for the fraud analyst to spot in a really effective way.

Furthermore, (Bhowmik, 2008) highlights these artificial intelligence techniques to be used to detect fraud: **Artificial neural networks**: to generate classification, clustering, generalization, and forecasting that can then be compared against conclusions raised in internal audits or formal financial documents,

**Pattern recognition:** to detect approximate classes, clusters, or patterns of suspicious behavior either automatically or to match given inputs,

**Data mining:** to classify, cluster, and segment the data and automatically find associations and rules in the data that may signify interesting patterns, including those related to fraud.

## Conclusion

The conclusion of this research has outcome some methods to be used to thwart all hypocritical communications over a network; in such a way that intruders, terrorists, and attackers could have no possible ways to achieve their goals in performing those illegal activities. Furthermore, the research has suggested the use of modern detection systems in achieving those methods to tackle illegal intrusions and unauthorized access and manipulation of data. With the use of these techniques, frauds and intrusion on a network will be diminished. Security challenges will also be eradicated.

## References

[1] Prabhakar S., Pankanti S. and Jain A. K. (2003) *Biometric Recognition Security and Privacy concerns*, Institute of Electrical and Electronics Engineers, Volume 1 Issue 2

[2] Falohun, A. (2012) *Development of a Feature Extraction Method for Iris Recognition using Enhanced Inverse Analytical Fourier-Mellin Transform*, Unpublished Ph. D. Thesis. Ladoke Akintola University of Technology, Ogbomoso, Nigeria.

[3] Dr. Iyobor, Bamidele, Aaron, Richard (2020) *A Forensic Analysis of WhatsApp on Android smartphone*,. International Research Journal of Computer Science

[4] Kim Y. G., Shin K. Y., Lee E. C. and Park K. R., (2012) *Multimodal Biometric Sytem based on the Recognition of faces and Both Irises*, International Journal of Advanced Robotic Systems

[5] Adedeji, Falohun, Alade & Amusan (2018) *Overview of Multibiometric Systems*, International Research Journal of Computer Science, Volume V, 459-466

[6] Baesens, B.; Vlasselaer, V.V.; and Verbeke W. 2015 *Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science*, North Carolina: SAS Institute Inc. p. 9

[7] Omidiora E. O. (2006) *A Prototype of Knowledge-based System for Black Face Recognition System using Principal Component Analysis and Fisher Discriminant Algorithms*, Unpublished Ph.D. Thesis, Ladoke Akintola University of Technology, Ogbomoso, Nigeria.

[8] Amita, Meenakshi, Simran, Varsha & Ajay (2020) *Object Detection Using Machine Learning*, International Research Journal of Computer Science

[9] Rao & K.C.Roy (2019) *Data Mining Using ID System: Advanced Intrusion Detection System*, International Research Journal of Computer Science

[10] HK Pratihari, SK Chakraborty and Sabyasachi Nath (2019) *Forensic Investigation of a Rape, Sodomy and Murder Case*, Journal of Forensic Science & Criminal Investigation

[11] Claus Weihs & Katja Ickstadt (2018) *Data Science: the impact of statistics*, International Journal of Data Science and Analytics

[12] Y.S. Park & S. Lek (2016) *Ecological Model Types*, Developments in Environmental Modelling

[13] Steven Walczak & Narciso Cerpa (2003) *Artificial Neural Networks*, Encyclopedia of Physical Science and Technology (Third Edition). Cambridge: Academic Press

[14] Navneet (2018) *Introduction of Cybercrime and its Type*, International Research Journal of Computer Science, Volume V, 435-439

[15] Mohiddeen Ahmad (19/7/2020) *Digitalize Nigeria's Biometric Security System*. Accessed 02/08/2020 https://flowdiary.com.ng/2020/07/19/opinion-digitalize-nigerias-biometric-security-system/

[16] The United States Department of Justice (last updated 25/8/2020) *Forensic Science*. Accessed 4/6/2020 https://www.justice.gov/olp/forensic-science

[17] Federal Trade Commission (2017) *Consumer Sentinel Network Data Book 2017: Fraud Reports by Amount Lost, Reported Fraud Losses.* Accessed 8/12/2018 https://www.ftc.gov/system/files/documents/reports/consumer-sentinel-network-data-book-2017/consumer_sentinel_data_book_2017.pdf

[18] BBC News (2006) *Nigeria scams 'cost UK billions.'* Accessed 3/6/2019 https://news.bbc.co.uk/2/hi/business/6163700.stm

[19] Ioana Rijnetu (last updated in May 2020), Andra Zaharia (originally published in January 2016) *Here are the Top Online Scams You Need to Avoid Today*. Accessed 17/9/2020 https://heimdalsecurity.com/blog/top-online-scams/

[20] SPD Group (23/3/2020) *How to Use AI ad Machine Learning in Fraud Detection*. Accessed 7/3/2020 https://spd.group/machine-learning/fraud-detection-with-machine-learning/

[21] Emerj, *AI-Based Fraud Detection in Banking – Current Applications and Trends*. Accessed 2/7/2020 https://emerj.com/ai-sector-overviews/artificial-intelligence-fraud-banking/

[22] Dan Nelson, *Gradient Boosting Classifier in Python with Scikit-Learn*. Acceesed 4/8/2019 https://stackabuse.com/gradient-boosting-classifiers-in-python-with-scikit-learn/

[23] Vihar Kurama (29/3/2020) *Gradient Boosting In Classification: Not a Black Box Anymore!* Accessed 3/9/2020 https://blog.paperspace.com/gradient-boosting-for-classification/

[24] Wikipedia (last edited 31/8/2020) *Artificial Neural Networks*. Accessed 10/9/2020 https://en.wikipedia.org/wiki/Artificial_neural_network

[25] Statista Research Department (2019) *A Minute on the Internet in 2019*. Accessed 7/4/2020 https://statista.com/chart/amp/17518/internet-use-one-minute

[26] InternetLivesStats (2020), *Google Search Statistics*. Accessed 9/6/2020 https://internetlivesats.com/google-search-statistics/

[25] Domo Research (17/7/2017) *Data Never Sleeps 5.0*. Accessed 7/4/2020 https://web-assets.domo.com/blog/wp-content/uploads/2017/07/17_domo_data-never-sleeps-5-01.png

[26] Forbes (21/5/2018) *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read.* Accessed 3/4/2019 https://forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-day-the-mind-blowing-stats-everyone-should-read

[27] Visual Capitalist (2019) *What Happens in an Internet Minute in 2019?* Accessed 8/9/2020 https://visualcapitalist.com/what-happens-in-an-internet-minute-in-2019

[28] Facebook Investor (2020) *First Quarter 2020 Operational and Other Financial Highlights*. Accessed 3/9/2020 https://investor.fb.com/investor-news/press-release-details/2020/Facebook-Reports-First-Quater-2020-Results/default.aspx