



An Overview of Credit Card Fraud Detection Learning Techniques

Norah Aljalawi

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

December 14, 2023

**THESIS AN OVERVIEW OF CREDIT CARD FRAUD DETECTION LEARNING
TECHNIQUES**

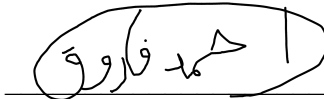
A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science
at Virginia State University, 2021.

by

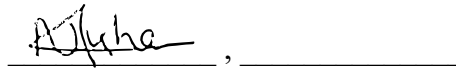
NORAH M. ALJALAWI

Bachelor's of Science, Al Majmaah University, KSA – 2011

Director: Thesis Ahmed F. Mohammed, Ph.D, Assistant Professor, Department of Computer
Science

A handwritten signature in Arabic script, "أحمد فاروق", enclosed in an oval shape, positioned above a horizontal line.

Committee members: Naha Farhat, Ph.D., and Ju Wang, Ph.D.

A handwritten signature in Arabic script, "Naha", positioned above a horizontal line.

©Norah M. Aljalawi, July 2021 All

Rights Reserved.

Abstract

By Norah M. Aljalawi

THESIS AN OVERVIEW OF CREDIT CARD FRAUD DETECTION LEARNING TECHNIQUES

Under the direction of (Ahmed F. Mohammed, Ph.D., Naha Farhat, Ph.D., and Ju Wang,
Ph.D.)

Credit card fraudsters are becoming more creative, altering their behaviors, and finding new ways to trick computer systems. Card fraud has become a major national and global threat to e-commerce causing losses of great amounts of money. Immediate attention needs to be directed towards improving existing techniques, or creating new methods for pinpointing fraudulent transactions. Supervised classification algorithms have proven to be accurate measures for predicting illegal transaction with more than 90% accuracy. This work reviews existing techniques and compares their reliability by examining their accuracy and speed on their application to three deferent data sets.

Acknowledgements

I would like to thank God first and foremost for giving me the energy to finish this work as well as my family for supporting me through this journey. Special thanks to Dr. Ahmed F. Mohammed for guiding me through this process and preparing for my final defense. Many thanks to my thesis committee members, Drs. Naha Farhat and Ju Wang for being supportive advisors. Many thanks to Dr. Joon-Suk Lee for advising me on course selection and motivating me to work hard. All of whom have greatly impacted me in a positive manner and helped me grow to make better decisions in academia. Finally, many thanks to Virginia State University for giving me this opportunity. I hope I have met your expectations.

TABLE OF CONTENTS

Chapter	Page
Abstract.....	i
Acknowledgements.....	ii
TABLE OF CONTENTS.....	iii
LIST OF TABLES.....	v
LIST OF FIGURES.....	vi
INTRODUCTION.....	1
1.1 Background.....	1
1.2 Challenges.....	3
1.3 Proposed Solutions.....	5
1.4 Motivation.....	7
TYPES OF FRAUD.....	9
2.1 Credit card fraud.....	10
2.2 Telecommunication Fraud.....	10
2.3 Computer Intrusion.....	11
2.4 Bankruptcy Fraud.....	11
2.5 Application Fraud.....	12
LITERATURE REVIEW.....	13
METHODOLOGY.....	17
4.1 Assessing accuracy and performance.....	18
4.2 Random Forests.....	19
4.3 Support Vector Machine.....	20
4.4 Logistic Regression.....	21

4.5	Naive Bayes.....	21
4.6	Multi-layer Perceptron.....	22
4.7	AdaBoost.....	23
4.8	Dataset.....	24
	RESULTS	25
5.1	Abstract dataset	25
5.2	European dataset	27
5.3	PaySim dataset	28
5.4	Comparing performances	29
	CONCLUSION.....	32
	FUTURE WORKS.....	35
	Appendix A.....	37
	ABBREVIATIONS	37
	Appendix B.....	38
	DATA DOWNLOAD SOURCES	38
	REFERENCES	39
	VITA.....	44

LIST OF TABLES

Table	Page
1 Dataset demographics	23
2 Learning model accuracies and training speeds for the Abstract dataset.....	25
3 Learning model accuracies and training speeds for the European dataset.....	26
4 Learning model accuracies and training speeds for the PaySim dataset.....	28
5 Mean accuracies for model performances per dataset.....	29
6 Mean training speed for model runs per dataset.....	29
7 Mean of mean accuracies and performances.....	30

LIST OF FIGURES

Figure	Page
1 A depiction of a k-fold cross validation estimation. The first iteration considers a random 10% as test data while the remaining 90% are used to train a model. This process is repeated k times.	17
2 A sample decision tree that splits the data according to it's dimensions: (1) color, and (2) is underlined.	18
3 A depiction of Support Vector Machine where a line is placed at equal distance between the support vectors, separating the points into two groups.	19
4 An example Multi-layer perceptron showing a mesh of connections, or neurons, between the it's layers.....	21
5 A depiction of data boosting using the AdaBoost algorithm. Each model can make weak decision on a data set, but collectively, the different learners 1-N can make a powerful decision.	22

CHAPTER 1

INTRODUCTION

1.1 Background

The term fraud refers to obtaining money through deceptive means or illegal revenue resulting in achieving personal financial growth. Since illegal revenues pose as legitimate transactions, generally, a minimal defect in the transaction will label it as ‘illegitimate’, and although deceptive, revealing those defects will make illegitimate transactions stand out. However, revealing those events is difficult, making the act of conducting fraudulent activities a primary attraction for committing crime. The inflated use of the internet in everyday activities ranging from paying bills to ordering goods and services has widened the door for more credit card frauds. The term “credit card fraud” refers to successful and unsuccessful attempts to gain illegitimate funds during transactions involving the use of credit cards. Most commonly, credit card information is stolen by a criminal and used for personal gains, without the consent of the card’s owner. E-commerce makes it possible for criminals to use stolen information at any place, without any geographical restrictions and at any time, which makes the problem even worse [1]

In 2018, **AP News** reported 24.2 billion US dollars were lost due to credit fraud, while **Quartz** reported that 47% of all credit card fraud events occur in the US. An approximate number of 163,000 have experienced credit card fraud in the year 2018. According to **The Nilson Report**, card fraud is expected to have cost the world 30 billion US dollars in 2019 and 40 billion by 2027 [2]. This evidence is further supported by **The Federal Trade Commission (FTC) Yearbook Data**, where it is reported that credit card fraud ranked the highest forms of frauds with a reported number of 2,184,531 cases, while 1,387,615 cases for identity theft fraud. These numbers are reported to be growing since 2001, where credit card fraud cases scored 137,306, while identity theft totaled 86,250. In other words, the past 20 years credit card fraud in the US has grown by 1591%. A graph presented in the FTC report suggest that this number will continue to grow for many years to come [3]. All these reasons demand immediate attention towards creating and improving existing detection schemes that will improve E-commerce application building and management. The research is necessary as detection of frauds can be cumbersome and time consuming using the naked eye and personal logic. With the plethora of event occurrences, fraudulent activities can be easily hidden, therefore, fraud detection models are of paramount importance.

The three most important types of E-commerce are (1) B2B, or Business to Business, where one business entity or organization conduct a commercial transaction with another. B2B companies embody a noteworthy portion of the US economy, where it is projected that 72% of businesses operate for other businesses [4]. B2B E-commerce usually involves the digital sale or trade of goods and services between businesses via an online portal [5]. A considerable percentage of e-commerce industries leverage credit card purchases to

maintain profitability [6]. (2) B2C, Business to Consumer, or B2M, Business to Many have more sustainable chances of receiving funding as they receive exposure through word-of-mouth and maintain contracts with their customers that are likely to remain loyal. An example of B2C is a book publishing company, where the working relationship between the publisher and author is a key factor to creating a successful product. Such companies institute themselves more quickly on the market [7]. Lastly, C2C, or Consumer to Consumer, where businesses facilitate an online service for consumers to purchase goods and services from each other. Prime examples can be like using Craigslist or auctioning sites like eBay. Marketing is almost nullified because customers can locate their needed items at a click of a button and are allowed to communicate with other costumers to meet their needs, eliminating the middleman. Such model is prone to fraud, for example, retailers can create several accounts on an auction site to deceive interested buyers. One account is used to sell an item while the others bids on the very same item, driving the buyer to bid more money. Other theft scenarios, for example, identity theft scam artists can create websites containing popular domain names such as eBay that attracts loyal eBay costumers. These sites commonly ask costumers for personal information such as credit card numbers. Frequent cases are documented where clients find themselves presented with exaggerated bank statements for unknown purchases [8].

1.2 Challenges

There are several challenges facing credit card fraud detection methods [9]. Some of these challenges are associated with the data sets, or information about previous credit card transactions. Other challenges are due to fraudsters changing strategies and lack of technologies. This list provides the reader with popular examples.

1. **Unavailability of a real data set:** it is one of the most important problems facing credit card fraud detection methods that many researchers have faced. The reason is that financial institutions and banks do not allow disclosure of customer data and transactions, because they consider it a breach of privacy.
2. **Unbalanced data set:** credit card fraud data is considered perverted data because it is legitimate data, and some data is considered deceitful. Thus, legal transactions different from fraudulent for example 98% of transactions are legal while 2% are fraudulent. This also presents reliability issues because available data are highly imbalanced which can result in creating biased solutions.
3. **The size of the data set:** daily efforts incorporate processing multiple credit card transactions. Therefore, this analysis creates certain limitations for researchers as it demands high techniques and computing power that can accommodate large datasets and high intensity computations. It is later shown that large datasets can overthrow the performance of notable algorithms like the Multi-layer Perceptron.
4. **Determining the appropriate evaluation criteria:** there are two popular scales for fraud detection techniques: false positive and false negative rates. These measures have an adverse connection, decreasing one and increasing the other. Therefore, precision is not considered an appropriate scale to detect credit card fraud because the data set is not balanced. Hence, with high precision, all deceitful transactions can be incorrectly classified. The fault cost of incorrectly classifying fraudulent cases is higher than the fault cost of wrongly classifying legitimate cases. Therefore, it is important to consider the sensitivity and accuracy of the correct classified fraud cases for each case.

5. **The dynamic attitude of the fraudster:** fraudsters mean dynamic attitudes, that is, those who alter their behavior over time to bypass any new detection system and adjust fraud patterns. This is the reason why fraud has become progressively more sophisticated, eventually experts become almost unpredictable.
6. **Lack of standard metrics:** there is no standard to measure and compare the results of fraud detection systems [10].

1.3 Proposed Solutions

Several models have been proposed to aid in defensive actions which avoids credit card fraud and reduces financial threats such as theft of money, or intangible properties such as bonds and stocks. Various methodologies have been applied such as Data Mining, Artificial Intelligence, Genetic Algorithms, Hidden Markov Models, Cryptography, and Sequence Alignment [11, 12].

Data Mining is the process of deriving valuable information from large data sets available on databases and repositories. The application of the subject requires knowledge of interdisciplinary concepts, which can be mathematical or computational. Data numbers can be extensive, for example, as large as millions and billions of credit transactions a year, phone calls, etc., can be in the form of Gigabytes and Terabytes. Since size is a concern, data is generally downloaded and refined using database technologies and general-purpose programming. Information can be interpreted using Statistical Analysis, Machine Learning, Neural Networks, and pattern recognition approaches [13]. Moreover, Data mining refers to Knowledge Discovery in Databases (or KDD), an area that detects useful and new information from a large set of data. There are many areas in which data mining has been applied, including retail sales, bioinformatics, and counter-terrorism [14].

The proposed methods in this paper include a variety of Data Mining classification techniques. Composite data analysis tools are employed to discover unknown insight, associations, and recurring patterns among sizable datasets. Learning tools include Machine learning methods such Neural Networks, Support Vector machines and Random Forests. The process involves the collection of data represented in the form of text and numbers, where each feature can be recorded in the form of a String or a Number. Target values are represented as 1's and 0's, which is a hit or miss fashion for “fraud” and “not fraud”. Data are examined for estimation and predication using cross-validation errors for different classifiers [15].

The classification categories of credit card fraud detection techniques are fraud analysis (abuse detection) and user behavior analysis (deviation detection). The supervised classification deals with the techniques of the first group. Based on historical data, transactions are classified as fraudulent or normal. Then, classification models are created that can predict the status of new records whether they are probable or normal from that data. Examples of known techniques or methodologies are Extrapolation, Decision Trees, and Neural Networks. There is an approach known as abuse detection that reliably detects most fraud scams [15].

While the second group deals with the unsupervised methodologies that are based on account behavior. Fraud is detected when the fraudulent transaction conflicts with the user's normal behavior. Because it is natural for the fraudster to not behave in the same way as the account owner. The legitimate model of user behavior is extracted as a user profile, and then fraudulent activities are detected to achieve the goal. Through this method, new behaviors are compared to the model, and then it is easy to identify and distinguish the various activities as fraud. This is known as anomaly detection because it contains personal files such as merchant types, amount, location, and time of transactions.

There are two groups that solve technical data problems through: supervised learning and unsupervised learning. As mentioned above, in supervised learning, data points have target values. The assessment of the model accuracies are determined by cross-validation errors- the lower the error, the more accurate the model. The process is phased, first, by using some Data Mining algorithm, a dataset is trained, and a classifier is created. The quality of the classifier will be determined using the errors, and if the errors are too high, some preprocessing can be applied to the data and then the cycle is repeated. This cycle continues until a good error is reached or the algorithm is determined to fail the desired expectation. Preprocessing involves data cleaning such as removal of incomplete data points or having empty feature values, removing outliers and data balancing- ensuring that the number of data points corresponding to distinct target values are comparable in number.

Unsupervised learning combines tasks to make use of descriptive models. It is characterized by the ability to solve problems without prior knowledge of the data analyzed. Therefore, unlike supervised, the data points are not tagged to any target values. The training phase of the prediction models focuses on finding correlations or patterns among the features. These correlations are later converted to rules. For example, a geographical location analysis can sometimes reveal clusters, and hence, each cluster can be assumed to be a tag of its own [16].

1.4 Motivation

Most research focuses on improving the prediction efficiency of credit card fraud on single datasets. Minimal work considers the time needed to build these models, which is a considerable drawback. For this thesis we compare different supervised learning techniques and their efficiencies in terms of building prediction models, and for practicality purposes, the time it takes to train the data. The study of unsupervised techniques will be considered elsewhere. To improve the comprehensibility of our work, we consider efficiency and time

for 3 different datasets. Moreover, this work provides an overview of the current research of the topic, various learning techniques associated with the research, as well as a description of the most popular datasets for credit card frauds.

Our research show that there are a few learning techniques that can be superior to others. However, most algorithms can suffer in terms of reliability when presented with larger datasets.

The rest of this paper is organized as follows. In Chapter 2, we shed light on the different types of frauds and how they can be associated with credit cards or card fraud. The different types of credit card frauds are discussed. In Chapter 3, a brief history is provided in the context of a literature review. A timeline is given outlining the evolution of techniques used to tackle credit card fraud. In Chapter 4, goes over the different methods, tool(s) and data sets considered in the experiments that will determine which of the methods to be more reliable. The results are shown in Chapter 5 with comparisons of model accuracies and speeds across the different datasets. Finally, chapters 7 (CONCLUSION) and 8 (FUTURE WORKS) provide discussions for concluding remarks and possible directions for this line of research.

CHAPTER 2

TYPES OF FRAUD

Fraud can fall into two specific categories: internal fraud or extrinsic fraud. Internal fraud occurs when an employee perpetrates an organization, business entity or community via earning it's or their trust. Extrinsic frauds include a broad domain of planners, including vendors, customers, or theft by third parties. There are three kinds of Extrinsic fraudsters: 1) the average perpetrator known as soft fraud, 2) the criminal offender, and 3) the organized crime perpetrator known as hard fraud [17].

There are several types of frauds, famous among them, are relating to insurance fraud, internal fraud detection, credit card fraud, telecommunications fraud, computer intrusions, bankruptcy fraud, application fraud, and check forgery [18]. Internal fraud works in identifying fraudulent financial reports through management, while retail transactions are conducted by employees. Insurance fraud has several types including health insurance, home insurance, auto insurance, and crop insurance.

In this we focus mainly on credit card fraud. Here is a brief description of credit card fraud and the most likely types of fraud that can eventually lead to it.

2.1 Credit card fraud

Offline frauds are committed by using credit cards that are physically stolen, while online frauds are committed via internet, shopping, phone, and web. The latter does not require the physical card on-site for successful transactions to go through, the card's information will suffice. E-commerce growth has provided suitability for such transactions to be universal, and often undetected. During shopping, fraudsters can use stolen information or harass Merchant and banks by:

1. in case a fraudster does not intend to buy anything from the shop, he/she provides wrong information and payment was done via cash on delivery to harm the merchant.

Or,

2. in case a fraudster acquired card information like credit card number, CVV number, etc., the process of making online payments or purchases will be straight-forward [6].

2.2 Telecommunication Fraud

Gaining trust can be done through various means, such as through telecommunication devices. Perpetrators can pose as salespersons that eventually work their way to stealing personal information such as mailing addresses and social security numbers. Sufferers from this type of fraud are consumers, companies, and telecommunication service suppliers.

There are several forms of fraud against users by phone companies, like cramming, where additional charges are billed to the client's account for services that were never ordered. Slamming is another example, where any unauthorized alteration to the default carrier or Internet service for a subscriber, most often made by deceitful vendors keen to steal business from competing providers. Frauds against customers by third parties are

abundant and not limiting to company representative impersonations, call forwarding scams, telemarketing frauds, fraudulent customer owned coin-operated telephones, caller Id spoofing and many more.

Moreover, phone companies can use phone to fraud against one another like in cases of interconnect fraud, where records are falsified to purposely miscalculate the money owed by one telephone network to the other. Frauds can be conducted by users against phone companies, for example, subscription fraud, or signing up with false information, or intentional no return of equipment. Frauds can be committed against phone companies by third parties like in the case of phreaking, where acting callers or ‘phreaks’ ask systematic questions to company representatives which can reveal company strategies and methods of operations.

2.3 Computer Intrusion

Computer hacking is known as intrusion. Anyone accessing computers without permission leading to endless possibilities of fraud by stealing information or intentionally tampering with it. Intruders can be trusted personnel or knowledgeable hackers who know the system’s design. Stolen information can be used as fake identities to apply for credit cards and loans.

2.4 Bankruptcy Fraud

Filing for bankruptcy can be a major relief for those acquiring tremendous amounts of debts, yet, can be accomplished with fraudulent intentions. The majority who file for bankruptcy are truthful and plainly disclose all assets. However, there are recurring cases where someone yields to temptation and deliberately hide assets from their creditors. Some of these examples are purposely not listing an asset, providing false documents to courts, destroying documents, or paying third parties to hide properties. All are examples of ways

to conceal assets that can be sold for the benefit of the creditors. It is important to note that this type goes hand-in-hand with card fraud. Once one is capable of falsifying information, the very same deceitful strategy can be utilized to falsify financial documents used to support a credit request.

2.5 Application Fraud

Another form of identity fraud where fraudsters apply for new accounts to online services or products using conjured or stolen identities. This presents businesses with many challenges as many of their customers are fakes. Fake clients negatively influences new customer gains because most of the budget runs out on fake users, which distorts marketing campaigns and eventually companies lose large amounts of money. Moreover, the creation of sleeper accounts will initiate reputation issues.

Bad actors use stolen identities to apply for credit lines with no intention of fully settling their debts. On the long run, a fraudster can manage to acquire several lines of credit and build an authentic looking credit activity. Systematic build ups on credit activity associated with the fake name will eventually achieve good credit limits. When the time is right, the fraudster maxes out on all credit lines and is not to be heard of again.

CHAPTER 3

LITERATURE REVIEW

In attempt to decrease fraudulent credit card activity, researchers have developed several fraud detection techniques.

In 1986 **Shen et al** presents the increasing risk of credit card fraud. The efficiency of using classification techniques like decision trees, neural networks and logistic regression are shown to be promising. A framework for selecting the most applicable model for the fraud transaction type is proposed [19].

In 1993 **Quinlan** illustrates a conclusive description of his complete system of ID3 C4.5- a decision trees algorithm. The description includes latest developments, and various issues relating to decision trees such as missing features, tree conversions, tree pruning and producing an initial tree. For each of the above a clear description is given, as well as descriptions for the limitations of the algorithm. The methods presented are highly flexible in terms of working with data distributions and provides robust results [20].

In 1994, **Ghosh and Reilly** trained a neural network using a large data set of a credit card issuer labelled data set. The data consisted of account activity that lasted a period of two months. Examples of frauds due to lost cards, application fraud, stolen cards, counterfeit cards, mail-order fraud and non-received issue (NRI) fraud were considered. The system showed to be superior when rule-based detection procedures were installed on an IBM 3090 at Mellon Bank [21].

In 1999, **Chan et al** surveys and evaluates three main issues with credit card issuer data, namely, scalability due to large sample sizes, skewness of the distributions and nonuniform cost per error. Distributed data mining was proposed to reduce loss due to fraud using data boosting with multiple learners. A framework is set for organizing highly distributed databases [22]. **Lane** examines user profiles using Hidden Markovian Models. A user identity classification system is formulated founded on the anomaly detection parameters likelihood and give an approximation that permits this quantity to be estimated with trusted accuracy. The research concluded that the behavior of a trusted client is more consistent than that of an imposter [23]. **Stolfo et al** devised an AI-based approach that merges inductive learning and meta-learning proposition for improving classification of fraud data [24]. Inductive learning goes over distributed data sets to recognize unique patterns or outlier behaviors, while meta-learning is used to derive knowledge from these distinctive cases. Experiments were done on two different datasets supplied by two different financial institutions. This approach was considered appropriate for fraud detection.

In 2002, **Syeda et al** developed a parallelized neural network with the main intention of speeding the knowledge discovery process. They concluded that their algorithm gives fewer average training errors with larger data, however, the higher the error, the more likely the transaction being fraudulent [25].

In 2003, **Hoang et al** developed a multi-layer anomaly intrusion detection system based on Hidden Markovian Models and Enumerating Methods- an improvement over single layer approaches such as the mentioned above [26].

In 2008, **Srivastava et al** shows that Hidden Markovian Models can be used to represent credit card transactions as stochastic processes. The accuracy of the system presented is close to 80% for large variations of data [27].

In 2013, **Alekhyia et al** stressed the importance of achieving higher accuracies in credit card frauds due to the surging increase in e-commerce. Various techniques were explored such as Machine Learning, Fuzzy logic, Sequence Alignment and Genetic Algorithms. It is concluded that Genetic Algorithms were best suited for protecting e-commerce as it can easily adapt to changing behaviors of fraudsters. A prototype system was implemented [28]. **Ingole et al** uses Hidden Markovian Models to analyze order of operation in credit card transactions while involving the use clustering algorithms. The Hidden Markovian Models are trained using Baum-Welch algorithm that detects if an incoming transaction is fraudulent or not. The accuracy of the system was determined to be 75% [29]. **Meshram et al**, proposed an authentication mechanism that is composed of multiple security layers prior to entering pin numbers [12]. Secret questions are asked to the user for verification purposes forcing the user to pass layers of security before entering

a pin. The multiple layers of security composite is improved and implemented by des 3-des algorithm. **Shabbir et al** use genetic algorithms and scoring mechanisms to deduct fraudulent transaction and minimize the number of false alerts [30]. Essentially, odds of fraud attempts can be predicted before the credit card transactions. A series of anti-fraud approaches can be utilized to prevent banks from heavy losses and reduce risks.

CHAPTER 4

METHODOLOGY

All experiments were run using WEKA- a tool developed by the University of Waikato, New Zealand, that provides options for classification, regression and minimal visualizations such as viewing the number of data points per label [31]. Moreover, WEKA provides summary reports for individual runs which includes the time taken to train a model.

For a small to medium size data set, the product is efficient, However, for larger data sets like 10,000+ points, this product could run into speed issues. For this reasons, a few experiments were discarded.

Minimal preprocessing was done to the data before the runs such as balancing the data. Generally, the data is overflowing with N's, or “not fraud” data points. A JAVA module is written to balance the number of ‘N’ points with the number of ‘Y’ points= a crucial step to avoid overfitting. Other preprocessing were implemented such as removing irrelevant features, like line numbers and those who have same values for all the rows.

All experiments are run on a 64-bit stand-alone machine, processor: Intel core i-7, 3.2 GHz with 6 cores. The rest of this section goes over the different concepts considered for assessing the accuracy of the models, the learning methodologies, and the data considered.

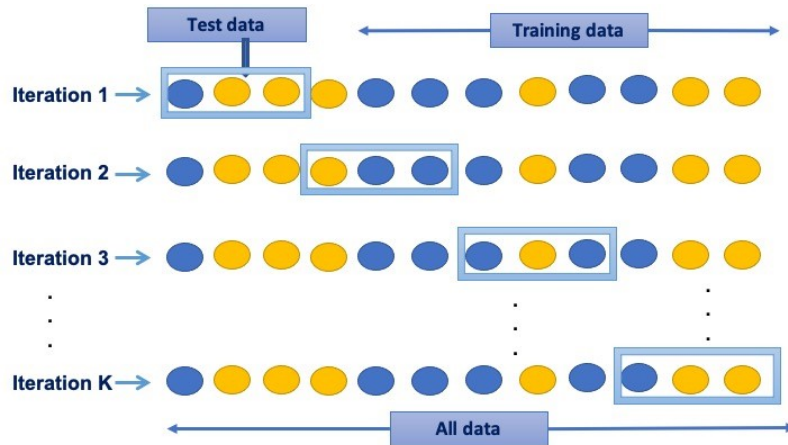


Fig. 1. A depiction of a k-fold cross validation estimation. The first iteration considers a random 10% as test data while the remaining 90% are used to train a model. This process is repeated k times.

4.1 Assessing accuracy and performance

The accuracy of the model is evaluated using a 10 k-fold cross-validation error. This means the process is done in iterations. As shown in Figure 1, the first iteration or 1st fold divides the data into a random 10% and 90%. The 90% are used to train a model, while the 10% are used for prediction. The error is then recorded and then the process is repeated 9 more times for 9 more models. By the end of the process, there will be 10 different errors for the 10-fold models. The errors are averaged to give us the 10 k-fold cross-validation error. This process is known to be completely unbiased due to the randomizations and has been used for several years.

Other metrics to assess accuracy are the number of correctly classified points and the number of incorrectly classified points for both data types (or tags)- precision, and recall.

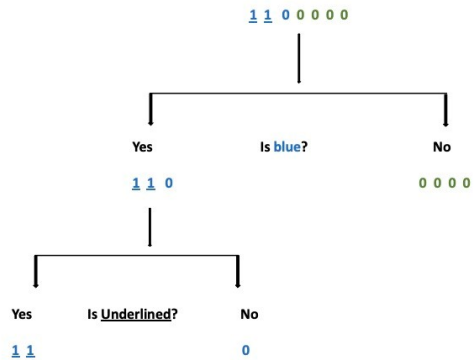


Fig. 2. A sample decision tree that splits the data according to it's dimensions: (1) color, and (2) is underlined.

4.2 Random Forests

Random forests are an ensemble learning method for classification that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees [32].

A decision tree is best explained by this example given in Figure 2. Imagine that we are trying to separate the data using their features. The features are color (blue vs. green) and whether the observation is underlined or not. The first split will decide which string goes to the left and which string goes to the right based on the question: is it blue or is it green? The second split will decide on whether a character is underlined or not. So “Yes” - “Yes” gives us underlined blue characters, “Yes” “No” gives us not underlined green characters, and “No” gives us blue characters.

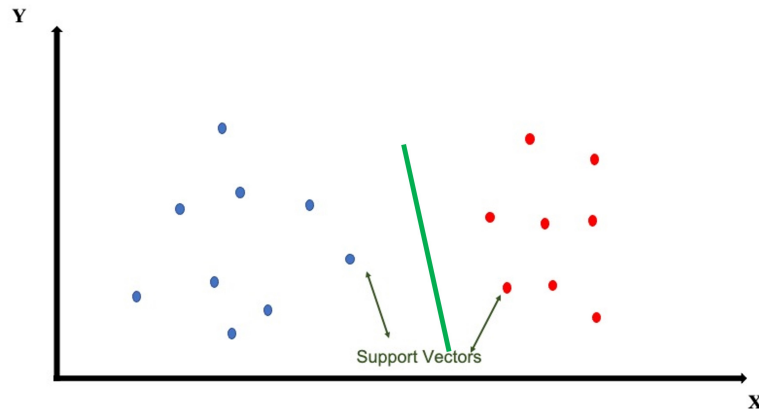


Fig. 3. A depiction of Support Vector Machine where a line is placed at equal distance between the support vectors, separating the points into two groups.

4.3 Support Vector Machine

A Support Vector Machine (or SVM) is a supervised machine learning algorithm that can be used for both classification and regression. SVMs is a quadratic optimization problem. The main idea is finding a hyperplane that best divides a dataset into two classes, as shown in Figure [3]. A line divides the blue set from the red set, the line is positioned at an optimal position, right in the middle, to separate the closest blue and red points by an equal distance. Having the separation distance (or margin) at an optimal distance which separates the two closest points of opposing classes (or support vectors), is the optimization criteria for SVMs.

In this example you can see a line separator because the data points are 1dimensional or have one feature. SVMs can also produce curves and circles. 2dimensional data are separated by planes, spheres, and cubes. 3 or more-dimensional data cannot be visualized by the naked eye, so they are referred to as hyperplanes.

4.4 Logistic Regression

A contradiction appears when we declare a classifier whose name contains the term ‘Regression’ is being used for classification, but therefore Logistic Regression has an advantage: it uses linear regression equation to produce discrete binary outputs. Similar to perceptrons inputs are entered in an activation function of the form shown in Equation 4.1- the Sigmoid Function. The function shrinks the data inputs into binary values of 0’s and 1’s. This makes our values normalized which helps us reach more consistent coefficients for curve or plane separators.

$$P(Y|X) = \frac{1}{1 + e^{-f(x)}} \quad (4.1)$$

4.5 Naive Bayes

Naive Bayes is similar to Logistic Regression in the regards that it has a formula that maps your input to specific values. The formula is given in Equation 4.2. It is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

For example, some fruit may be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all these properties independently contribute to the probability that this fruit is an apple and that is why it is known as ‘Naive’.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. Moreover, they are considered one of the fastest classifiers.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (4.2)$$

4.6 Multi-layer Perceptron

An Multi-layer perceptron (MLP) consists of at least three layers of nodes: an input layer, a hidden layer, and an output layer. The hidden layer can be further consisting of more layers depending on the design. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training [33].

As shown in Figure 4, input nodes (on the far left) accept data point values, which are passed to the hidden layer where they get multiplied by random weights, and these values keep transitioning forward until the output layer. The error is assessed and if it is not sufficient, these output values become new inputs for the next iterations. This is basically the propagation process. The weights will be continuously adjusted until the error stops improving. This concept is derived from the Gradient Descent approach in calculus.

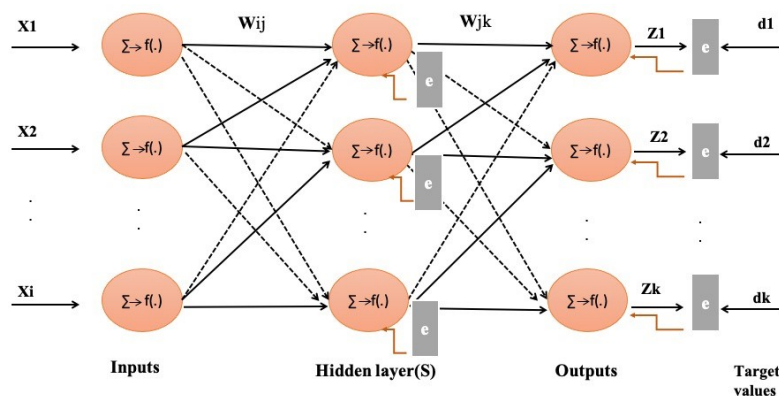


Fig. 4. An example Multi-layer perceptron showing a mesh of connections, or neurons, between the it's layers.

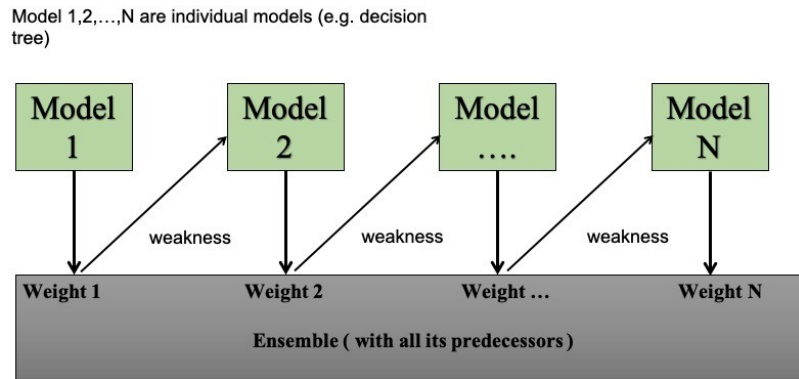


Fig. 5. A depiction of data boosting using the AdaBoost algorithm. Each model can make weak decision on a data set, but collectively, the different learners 1-N can make a powerful decision.

4.7 AdaBoost

AdaBoost algorithm, short for Adaptive Boosting, is a Boosting technique that is used as an Ensemble method in Machine Learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights to incorrectly classified instances. Boosting is used to reduce bias as well as the variance for supervised learning. It works on the principle where learners are grown sequentially.

Figure 5 depicts the concept of data boosting. AdaBoost works by having a base learner that works on a data. The model will have some incorrectly classified points. These points will be priority for a next learner. This learner might be able to correctly classify these points and miss on others. So basically, a model is created by running the data through several learners. The learners might be individually weak but collectively they can all form powerful decisions.

4.8 Dataset

This study considers three data sets for training and prediction purposes.

1. **European data set:** The dataset comprises credit card transactions in September 2013 by European cardholders. The transactions occurred in two days.
2. **PaySim data set:** A synthetic data set generated using a simulator called PaySim. PaySim uses aggregated data from existing data sets, and AI to generate a mock data set which resembles existing data behaviors.
3. **Abstract data set:** A minimized version of the European data set. Preprocessed to be reduced in number of columns and features by removing repeated data points and correlated features.

Table 1. Dataset demographics

C1	C2	C3	C4	C5	C6	C7
Set	Feature count	Point count	Feature count post preprocessing	Point count post preprocessing	Nature	Source
European	31	284808	31	984	Real	European cardholders
PaySim	11	1048576	10	2284	Synthetic	PaySim generator
Abstract	12	3076	12	896	Real	European cardholders

All data sets can be found on Kaggle.com. Their download links are provided in

Appendix B. The different set demographics are given in Table 1.

CHAPTER 5

RESULTS

Several experiments have been conducted to test different supervised learning algorithms on the datasets mentioned above. Some of these algorithms performed poorly, while others performed up to standard. In this section, we report results for those algorithms that performed relatively well in terms of predicting transaction outcomes, and the speed it takes to train a model.

5.1 Abstract dataset

Table 2 illustrates the accuracies and performances for different learning techniques applied to the Abstract dataset. Accuracies are resulting from a 10 k-fold cross validation during training. For relevance, precision and recall are reported. The third column reports the normalized speed of the training phase- speed is divided by the size of the data, or the number of cells. This information will be useful for obtaining the averages, which are shown below.

Individual analysis show that all algorithms perform well in terms of predicting target values. The top predictors of this list are Random Forrest and logistic Regression.

All algorithms are relatively fast, However, this seems to be the case because the dataset is relatively small, only 896 data points. The precision and recall usually correlates with the cross-validation errors.

Table 2. Learning model accuracies and training speeds for the Abstract dataset.

Model type	Accuracy (%)	Build Time (s)	Time/Datasize	Precision	Recall	Class
RF	97.21	0.06	$5.58 \rightarrow 10^6$	0.98	0.964	Y
				0.965	0.98	N
SVM	96.09	0.02	$0.537 \rightarrow 10^6$	0.971	0.961	Y
				0.951	0.961	N
LR	97.88	0.05	$4.65 \rightarrow 10^6$	0.971	0.983	Y
				0.9986	0.971	N
NB	91.74	$\leftarrow 0$	$\leftarrow 0$	0.888	0.955	Y
				0.879	0.914	N
MLP	96.76	0.49	$4.56 \rightarrow 10^5$	0.962	0.972	Y
				0.972	0.962	N
AdaBoost	90.51	0.01	$9.3 \rightarrow 10^7$	0.861	0.967	Y
				0.962	0.844	N

In regards to speed, second to none is that of Naive Bayes. The use of the probability functions makes model generation much faster than regular predictors, almost no time. This usually because the probability function has fewer parameters than other activation functions like for example the Multi-layer perceptron (MLP). All models have relatively

good speeds expect MLP. At it's current state it appears to be okay, however, when the data size increases, this will eventually reveal disadvantages.

5.2 European dataset

Table 3 illustrates the accuracies and performances for different learning techniques applied to the European dataset. Results are consistent for speed and accuracy where Random Forest and Logistic Regression are leading. All models perform relatively well. While Logistic Regression performs better than AdaBoost for the Abstract dataset, it seems that data boosting came to an advantage with larger datasets.

Naive Bayes is highest in speed while MLP is lowest. Training larger data with MLP will take longer as there will be several input triggers to the activation functions. In addition, WEKA tries to find the best mesh of neurons suited for the training which could take a long time.

Table 3. Learning model accuracies and training speeds for the European dataset.

Model type	Accuracy (%)	Build Time (s)	Time/Datasize	Precision	Recall	Class
RF	93.29	0.15	$4.92 \rightarrow 10^6$	0.9	0.974	Y
				0.971	0.892	N
SVM	92.28	0.01	$3.29 \rightarrow 10^7$	0.874	0.988	Y
				0.986	0.858	N
LR	93.7	0.03	$9.83 \rightarrow 10^7$	0.923	0.953	Y
				0.952	0.921	N
NB	89.94	$\leftarrow 0$	$\leftarrow 0$	0.857	0.959	Y
				0.954	0.839	N
MLP	92.58	2.75	$9.01 \rightarrow 10^5$	0.917	0.937	Y
				0.936	0.915	N
AdaBoost	91.87	0.4	$1.3 \rightarrow 10^6$	0.901	0.941	Y
				0.938	0.896	N

5.3 PaySim dataset

Table 4 illustrates the accuracies and performances for different learning techniques applied to the PaySim dataset.

As the size of the data increases, we observe that AdaBoost was the only algorithm that maintained an over 90% accuracy. The use of several learners in one single algorithm makes this possible. All other algorithms performed within the 80% range

Table 4. Learning model accuracies and training speeds for the PaySim dataset.

Model type	Accuracy (%)	Build Time (s)	Time/Datasize	Precision	Recall	Class
RF	86.51	0.22	$7.5 \rightarrow 10^4$	0.872	0.799	Y
				0.814	0.883	N
SVM	84.06	17.14	$3.29 \rightarrow 10^7$	0.874	0.988	Y
				0.986	0.858	N
LR	83.58	29.13	$1.28 \rightarrow 10^3$	0.9	0.756	Y
				0.789	0.916	N
NB	84.63	$\leftarrow 0$	$\leftarrow 0$	0.864	0.822	Y
				0.83	0.87	N
MLP	NA	NA	NA	NA	NA	Y
				NA	NA	N
AdaBoost	92.51	0.02	$8.76 \rightarrow 10^7$	0.935	0.914	Y
				0.916	0.936	N

Table 5. Mean accuracies for model performances per dataset.

Dataset	RF	SVM	LR	NB	MLP	AdaBoost
Abstract	97.2	96.1	97.9	91.7	96.8	90.5
European	93.3	92.3	93.7	89.9	92.6	91.9
PaySim	86.5	84.1	83.6	84.6	N/A	92.5
Average	92.3	90.8	91.7	88.8	94.7	91.6

Table 6. Mean training speed for model runs per dataset.

Dataset	RF	SVM	LR	NB	MLP	AdaBoost
Abstract	$5.6 \rightarrow 10^6$	$0.5 \rightarrow 10^6$	$4.7 \rightarrow 10^6$	$\leftarrow 0$	$4.6 \rightarrow 10^5$	$9.3 \rightarrow 10^7$
European	$4.9 \rightarrow 10^6$	$3.3 \rightarrow 10^7$	$9.8 \rightarrow 10^7$	$\leftarrow 0$	$9.0 \rightarrow 10^5$	$1.3 \rightarrow 10^6$
PaySim	$9.6 \rightarrow 10^6$	0.0008	0.0013	$\leftarrow 0$	NA	$8.8 \rightarrow 10^7$
Average	$6.7 \rightarrow 10^6$	0.0002	0.0004	$\leftarrow 0$	$6.9 \rightarrow 10^5$	$1.0 \rightarrow 10^6$

Table 7. Mean of mean accuracies and performances.

Model	Average Accuracy	Avg Time
RF	92.34	$6.71 \rightarrow 10^6$
SVM	90.81	0.0002
LR	88.77	$\leftarrow 0$
NB	91.72	0.0004
AdaBoost	91.63	$1.04 \rightarrow 10^6$

CHAPTER 6

CONCLUSION

Credit card fraud is an example of employing deceptive means to acquire illegal income. Although such means are hard to detect because they are often overlooked by the naked eye, there are ways to detect these abnormalities electronically. Sadly the growth of e-commerce has made this a difficult task, statistics show that growth of e-commerce has resulted in more opportunities for fraudsters to steal. Abnormal fraudster behavior is still evolving, which begs for immediate attention towards developing more means for stopping illegal fraudulent transactions. There are several challenges associated with detecting fraudulent transactions such as limitation and ambiguity of the available data. Moreover, fraudsters continue to manage to maneuver the system in their favor.

One approach to limit the success of fraudsters is to utilize Data Mining where computer systems can learn from previous transactions how to predict futuristic fraud attempts. Several methods have been proposed in the past like Machine Learning, Neural Networks, and pattern recognition. These algorithms fall under the supervised learning category because they work with labelled data, It is characterized

by the ability to solve problems without earlier understanding of the data analyzed. They have proven to be successful in the past which motivated the research.

Most research focuses on obtaining high accuracy in terms of predicting illegal transactions from a single dataset, while barely reporting the performances in these models. Our work explores several supervised classification techniques used to train prediction models on three different datasets. Moreover, the speeds for model generations are reported.

This work serves as a comprehensive review for fraud detection methodologies, hence, a discussion of the various fraudulent activities relating to credit card fraud are discussed. Credit card fraud, either online or offline, usually is initiated with some form of theft, whether the physical card or its information. Fraudsters can gain such advantage by means of telecommunication, device hacking, bankruptcy scams, and fake application. This paper enforces the understanding of the problem by alerting the reader to these forms of misuse of trust. Moreover, a timeline of the advancements to tackle the problem is given in the context of a literature review. Several methods have been proposed in addition to supervised learning such as Hidden Markovian Chains, security questions and Genetic Algorithms.

Our findings suggest that there are five main supervised learning techniques that can predict fraudulent transactions with +85% accuracy. Namely, Random Forests, Support Vector Machines, Logistic Regression, Multi-layer Perceptrons (MLP), and AdaBoost. All models are relatively reliable, however, scalability can affect the accuracies. This is due to the fact that the data is highly inconsistent with the range of feature values resulting from varying fraudster behaviors. As the size of the data increases, the possibility of including more inconsistent values rises and the training process suffers with the exception of AdaBoost. The algorithm thrives on weak data points, which is the essence of data boosting.

Successive training isolates weak points during the boosting phases and use them to make strong decisions.

MLP is a powerful algorithm with cases of 96% accuracy, however, it has proven to be slow when training models for large datasets. This is due to the fact that finding the best hidden layer structure can take considerable time. Given suitable equipment, the algorithm should produce competitive results for larger datasets, however, this remains to be seen. Naive Bayes (NB) have proven to be reliable for smaller datasets, but most importantly, extremely fast as compared to the others. However, NB is at a drawback because the accuracy suffers when creating models for large datasets. AdaBoost combines competitive accuracy along with speed, but most importantly, it is resilient to data variability which makes it the most reliable.

CHAPTER 7

FUTURE WORKS

The main goal here is to achieve more reliable results. This can be done by including more datasets to the fold, which will add to the comprehensibility of the results for testing model resilience. One data set that can be used is that generated by Sparkov an AI data set generator which creates synthetic data that mimics existing sets, much like PaySim. Moreover, the narrative can be expanded by including unsupervised learning. Comparing both supervised and unsupervised techniques will provide a more comprehensive view for model reliability.

Although this study has explored training across three datasets, the sizes of the data is still relatively small. The preprocessing removes several data points in order to balance the data and avoid bias model creation. A strategy should be devised to merge balanced datasets by feature elimination, data discretization and normalization. It would be interesting to train large datasets using AdaBoost. Not to mention, if more data emerges in the community, they should be considered as well.

The high accuracy resulting from Multi-layer Perceptron (MLP) cannot be disregarded. WEKA takes a long time trying to find the best hidden layer structure that gives the least amount of error. However, a different direction of research can be dedicated for this purpose, and design a MLP using other sources like MATLAB

for example. Although the possibility of missing powerful structures is high, there is a high chance of converging on a relatively good predictor.

Appendix A

ABBREVIATIONS

B2B	Business to Business
B2C	Business to Consumer
B2M	Business to Many
FTC	Federal Trade Commission
LR	Logistic Regression
MLP	Multi-layer Perceptron
NB	Naive Bayes
RF	Random Forest
SVM	Support Vector Machine

Appendix B

DATA DOWNLOAD SOURCES

1. **European data set:** <https://www.kaggle.com/mlg-ulb/creditcardfraud>
2. **PaySim data set:** <https://www.kaggle.com/ealaxi/paysim1>
3. **Abstract data set:** <https://www.kaggle.com/shubhamjoshi2130of/abstract-data-set-for-credit-card-fraud-detection>

References

- Abdallah, Aisha, et al. "Fraud Detection System: A Survey." *Journal of Network and Computer Applications*, vol. 68, 2016, pp. 90–113. *Crossref*, doi:10.1016/j.jnca.2016.04.007.
- Adams, Niall M., et al. "Data Mining for Fun and Profit." *Statistical Science*, vol. 15, no. 2, 2000. *Crossref*, doi:10.1214/ss/1009212753.
- Baker, Ryan S. J. D. "Data Mining for Education." *International Encyclopedia of Education (3rd Edition)*, 2010, www.columbia.edu/~rsb2162/Encyclopedia%20Chapter%20Draft%20v10%20fw.pdf.
- Bousquet, Olivier, et al. *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2–14, 2003, Tübingen, Germany, August 4–16, 2003, Revised ... (Lecture Notes in Computer Science, 3176)*. 2004th ed., Springer, 2004.
- Cegielski, Kelly Rainer; Brad Prince; Casey. *Introduction to Information Systems - Fifth Edition*. 5th ed., Wiley, 2021.
- Chan, P. K., et al. "Distributed Data Mining in Credit Card Fraud Detection." *IEEE Intelligent Systems*, vol. 14, no. 6, 1999, pp. 67–74. *Crossref*, doi:10.1109/5254.809570.

- Chaudhary, Khyati, et al. “A Review of Fraud Detection Techniques: Credit Card.” *International Journal of Computer Applications*, vol. 45, 2012, citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.677.970&rep=rep1&type=pdf.
- “Federal Trade Commission. Consumer Sentinel Network. Feb. 2021.” *Consumer Sentinel Network*, Data book 2020, 2020, www.ftc.gov/system/files/documents/reports/consumer-sentinel-network-data-book-2020/csn_annual_data_book_2020.pdf.
- “Fortune 500.” *Fortune*, 2015, www.googleadservices.com/pagead/aclk?sa=L&ai=DChcSEwiKlb7EkKjyAhUN13cKHffKBFcYABAAGgJIZg&ae=2&ohost=www.google.com&cid=CAESQOD2qcC-mM08XiwXC1GJo3m5h25eArh8qyhsDxeuo_1JV2anBQrfobEyGGQa8f7ZkkEJJAVVN58kwOQGzDrBZSI&sig=AOD64_1L917rP9PWLgjb-z0JpGULCfeg7g&q&adurl&ved=2ahUKEwi-37LEkKjyAhUNCRoKHfbwB2oQ0Qx6BAgDEAE&dct=1.
- Ghosh, and Reilly. “Credit Card Fraud Detection with a Neural-Network.” *Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences HICSS-94*, 1994. *Crossref*, doi:10.1109/hicss.1994.323314.
- Guadamuz González, Andrés. “EBay Law: The Legal Implications of the C2C Electronic Commerce Model.” *Computer Law & Security Review*, vol. 19, no. 6, 2003, pp. 468–73. *Crossref*, doi:10.1016/s0267-3649(03)00604-6.
- Ingole, A. “Credit Card Fraud Detection Using Hidden Markov Model and Its Performance.” *Semantic Scholar*, 2013, www.semanticscholar.org/paper/Credit-

[Card-Fraud-Detection-Using-Hidden-Markov-and-Ingole-Thool/8420e69562c7fbd6eb443f20207355fa556d7137.](https://doi.org/10.21961/8420e69562c7fbd6eb443f20207355fa556d7137)

- Joshi, Kaneeka. “A Review of Credit Card Fraud Detection Techniques in E-Commerce.” *Xournals*, Academic Journal of Forensic Sciences, 2018, www.academia.edu/39529497/A_review_of_Credit_card_Fraud_Detection_techniques_in_e_commerce.
- JRana, Priya, and Jwalant Baria. “A Survey on Fraud Detection Techniques in Ecommerce.” *International Journal of Computer Applications*, vol. 113, no. 14, 2015, pp. 5–7. *Crossref*, doi:10.5120/19892-1898.
- Lane, Terran. “Hidden Markov Models for Human/Computer Interface Modeling.” *School of Electrical and Computer Engineering and CERIAS*, 1999, citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.158.4810&rep=rep1&type=pdf.
- Lorette, Kristie. “How ECommerce Can Reduce Business Transaction Costs.” *Small Business - Chron.Com*, 21 Nov. 2017, smallbusiness.chron.com/ecommerce-can-reduce-business-transaction-costs-3503.html.
- Pedro, Hugo T. C., et al. “Mathematical Methods for Optimized Solar Forecasting.” *Renewable Energy Forecasting*, 2017, pp. 111–52. *Crossref*, doi:10.1016/b978-0-08-100504-0.00004-4.
- Petrov, Christo. “21+ Credit Card Fraud Statistics to Keep You Aware in 2021.” *SpendMeNot*, 19 May 2021, spendmenot.com/blog/credit-card-fraud-statistics.

- Pratiksha, Ms., et al. "Credit and ATM Card Fraud Prevention Using Multiple Cryptographic Algorithm." *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 8, 2013, ijarcsse.com/Before_August_2017/docs/papers/Volume_3/8_August2013/V3I7-0546.pdf.
- Prodromidis, Andreas, and Salvatore Stolfo. "Agent-Based Distributed Learning Applied to Fraud Detection." *Researchgate*, 1999, www.researchgate.net/publication/2800734_Agent-Based_Distributed_Learning_Applied_to_Fraud_Detection.
- Quinlan, Ross. *C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning)*. 1st ed., Morgan Kaufmann, 1992.
- Shabbir, Syed Ahsan, and Kannadasan R. "A Survey of Methodology of Fraud Detection Using Data Mining." *International Journal of Scientific and Research Publications*, vol. 3-1, no. 5, 2013. *Crossref*, www.ijsrp.org/research-paper-0513/ijsrp-p1771.pdf.
- Shen, Aihua, et al. "Application of Classification Models on Credit Card Fraud Detection." *2007 International Conference on Service Systems and Service Management*, 2007. *Crossref*, doi:10.1109/icsssm.2007.4280163.
- Sorournejad, Samaneh, et al. "A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective." *Researchgate*, 2016, arxiv.org/pdf/1611.06439.pdf.

Srivastava, A., et al. “Credit Card Fraud Detection Using Hidden Markov Model.” *IEEE Transactions on Dependable and Secure Computing*, vol. 5, no. 1, 2008, pp. 37–48. *Crossref*, doi:10.1109/tdsc.2007.70228.

Syeda, M., et al. “Parallel Granular Neural Networks for Fast Credit Card Fraud Detection.” *2002 IEEE World Congress on Computational Intelligence. 2002 IEEE International Conference on Fuzzy Systems. FUZZ-IEEE’02. Proceedings (Cat. No.02CH37291)*, 2002. *Crossref*, doi:10.1109/fuzz.2002.1005055.

Tin Kam Ho. “Random Decision Forests.” *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 2002. *Crossref*, doi:10.1109/icdar.1995.598994.

Vadim, Kochetov. “Overview of Different Approaches to Solving Problems of Data Mining.” *Procedia Computer Science*, vol. 123, 2018, pp. 234–39. *Crossref*, doi:10.1016/j.procs.2018.01.036.

Witten, Ian, et al. *Data Mining: Practical Machine Learning Tools and Techniques (Morgan Kaufmann Series in Data Management Systems)*. 4th ed., Morgan Kaufmann, 2016.

Xuan Dau Hoang, et al. “A Multi-Layer Model for Anomaly Intrusion Detection Using Program Sequences of System Calls.” *The 11th IEEE International Conference on Networks, 2003. ICON2003.*, 2003. *Crossref*, doi:10.1109/icon.2003.1266245.

Zareapoor, Masoumeh, and Pourya Shamsolmoali. “Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier.” *Procedia Computer Science*, vol. 48, 2015, pp. 679–85. *Crossref*, doi:10.1016/j.procs.2015.04.201.

VITA

Norah Mohammed Abdulaziz Aljalawi earned her Bachelor's of Science degree in the field of Computer Science in 2011 from Al Majmaah University in Riyadh, Kingdom of Saudi Arabia (KSA). In 2013, Norah was employed in KSA's Almaalee Institute in Al Majmaah to prepare candidates for The Interdisciplinary Council on Development and Learning (ICDL) certification. In 2018, Norah taught middle and high school grades in Manarat Al Majmaah. Her research focuses on applications of using Datamining techniques.