Fraud Detection and Prevention in Financial Services Using Big Data Analytics

Anurag Mashruwala
anuragm@bu.edu

Abstract. This paper provides an in-depth analysis of the role of Big Data analytics in fraud detection and prevention within the financial services sector. It explores various methodologies, techniques, and challenges associated with leveraging Big Data to combat financial fraud.

1. Introduction

Financial fraud has become a widespread issue in today's global commerce, evolving in complexity and adaptability over time. It encompasses various activities like identity theft, payment card fraud, and cybercrimes, posing significant challenges that require innovative responses [1]. With technology deeply integrated into financial transactions, both opportunities and challenges arise, as criminals exploit weaknesses for illicit gains [1].

Understanding the current scope of financial fraud is crucial for developing effective prevention strategies [1]. These schemes continually evolve, from traditional scams to sophisticated cyber-attacks, demanding proactive and flexible approaches to detection and prevention. The consequences extend beyond monetary losses, affecting the reputations and trust of financial institutions and businesses alike. The interconnected nature of the global financial system amplifies the impacts of fraud, emphasizing the need for robust detection mechanisms.

Early intervention is vital in mitigating the ripple effects of financial fraud, safeguarding the integrity of financial systems and protecting stakeholders' interests [1]. Leveraging advanced technologies, such as Big Data Analytics, plays a pivotal role in detecting fraudulent activities in real-time by analyzing vast datasets for patterns and anomalies.

This review focuses on the essential role of Big Data Analytics in detecting financial fraud, highlighting its potential to enhance detection capabilities. It also addresses the challenges associated with implementing such technologies while exploring opportunities for leveraging technological advancements in combating fraud in the digital age.

2. Big Data Technologies for Fraud Detection

Big Data is characterized by 5V's: Volume, Velocity, Variety, Veracity and Value. Volume signifies the magnitude of data, often substantial in size. Velocity denotes the rate at which diverse data streams in [2]. Variety pertains to the intricacy of both structured and unstructured data originating from diverse sources. Veracity pertains to the accuracy of obtained data and can tackle quality concerns such as noise or absent values. Value is the subsequent quality aspect crucial for locating pertinent data for analysis. Techniques like Classification, Prediction, Clustering, Dimensionality Reduction, Regression, Artificial Neural Network, and Outlier Analysis from data mining and Machine Learning are employed to discern patterns in fraudulent transactions using gleaned data.

Big Data analysis tools offer an efficient method for identifying unusual patterns in detecting retail fraud [2]. The fraud detection process using Big Data architecture has three main components: data gathering, fraud detection, and communication to users via APIs. The Hadoop framework includes the Hadoop Distributed File System (HDFS) for reliable and cost-effective data storage. It employs MapReduce functions for sorting, filtering, and reducing data, which aids in fraud detection, particularly in the vast and varied retail sector. MapReduce, a programming model consisting of map and reduce components, is utilized for sorting, filtering, and summarizing data. Transactional data, including customer details and retail procedures, are aggregated and processed within the Big Data framework. The processed data are organized into fields and tables, then fed into a search engine where predictive analysis and machine learning algorithms are applied for fraud detection. Recent advancements in Machine Learning (ML) have introduced new models and algorithms, such as NLP and vision-based techniques. Spark, a platform known for its high performance, has gained popularity over the Hadoop framework in recent years. Unlike Hadoop's MapReduce, which frequently writes program states to disk, ML applications often require extensive memory usage [2].

Big data analytics is often employed to analyze user transaction patterns for fraud detection and prevention. Machine learning and big data methodologies have been recommended for detecting and preventing fraud in online transactions. The model proposed by Balasupramanian. N et al. [3] utilizes extensive transactional data, which undergoes cleaning, attribute extraction, and dimensionality reduction using Principal Component Analysis (PCA). The reduced features obtained from PCA serve as inputs to train the model using a machine learning approach. PCA provides a mathematical tool for reducing the dimensions of high-dimensional multivariate datasets. Self-organizing Feature Map (SOFM) or Self-organizing Maps (SOM) are employed to identify hidden patterns and aid in clustering. SOM, a type of artificial neural network, generates a low-dimensional map of input space from training examples, functioning through two phases: training and mapping. In the training stage, input examples are used to construct the map, employing feature extraction techniques in pattern recognition and classification. The primary benefit of employing big data analytics for identifying fraudulent transaction patterns lies in its capacity to swiftly and effectively detect fraud. The Hadoop framework includes a map-reduce component capable of storing vast datasets. Data mining methodologies, including clustering, decision trees, and Naive Bayes, are applied within big data platforms. The majority of techniques available for fraud detection rely on machine learning or data mining algorithms. Some of the machine learning techniques used to detect fraud are [4]:

Support Vector Machine (SVM):

SVM, a supervised machine learning technique, aims to find a maximum margin hyperplane to categorize input training data into two distinct classes [4]. It has the ability to classify new data points by utilizing a labeled training set for each class. Researchers have extensively explored Support Vector Machine (SVM) techniques for fraud detection. They have developed various hybrid approaches combining SVM with other methodologies, such as the fusion Danger theory, spike detections, logic, linear regression, and deep learning. These approaches have been applied to different domains, including credit card transactions, medical billing, and insurance industries, aiming to enhance fraud detection accuracy and efficiency. Experimental results across these studies consistently show promising improvements over existing methods.

A characteristic fundamental to SVM is the drawing of the hyperplane and categorizing the normal from the abnormal hence SVM's are best suited for identifying abnormal bank transactions and alerting the bank and the credit card holder. Indices or features such as unpaid balance, daily transaction, frequency, late payments can be used to create a model. The SVM algorithm usually performs far better than Naïve Bayes, Decision trees, logistic regression and ANN algorithms when dealing with well-separated classes and smaller datasets. Gedela et. al showed that it achieved an f-score of 99.33% and was superior to all the above algorithms [5].

Fuzzy-Logic-Based Method:

Fuzzy logic (FL) serves as a valuable framework for handling data in uncertain and ambiguous contexts [6]. It acknowledges that thinking methods are not precise but estimative [7]. Fuzzy combinations provide effective tools for complex modeling, offering new and improved approaches. Several fraud detection methods utilize FL principles. For instance, the FUZ-ZGY hybrid model integrates fuzzy and Fogg behavioral models to detect anomalous behaviors in credit card transactions. Another method employs fuzzy logic to categorize fraud and non-fraud transactions with reduced false positives, using fuzzy c-means clustering and an ANN model. Additionally, a study [8] introduces a fuzzy logic-based fraud detection method in the banking system, improving accuracy in classifying fraudulent activities by defining rules based on expert experience. This approach is further enhanced by constructing fuzzy rules for improved fraud detection. Other techniques [6] include rule-based methods utilizing algorithms like the firefly algorithm and threshold-accepting to differentiate fraudulent from non-fraudulent transactions, as well as fuzzy-rule-based approaches integrated with genetic feature selections, achieving good performance by removing irrelevant attributes and inducing fuzzy rules [6].

An application of fuzzy logic is in insurance claims. Fuzzy logic enhances insurance claims analysis by handling the inherent uncertainty and complexity in evaluating claims. Unlike rigid criteria, fuzzy logic assigns degrees of membership to various features, such as damage severity and claim history, allowing for a nuanced assessment of claims. It uses fuzzy rules to evaluate claims on a spectrum, flagging those that exhibit patterns of potential fraud without relying on binary classifications. This approach improves the ability to detect anomalies, adapt to new fraud patterns, and integrate expert knowledge, leading to more flexible and accurate fraud detection in insurance claims.

Hidden Markov Model (HMM):

The Hidden Markov Model (HMM) is a sophisticated probabilistic method commonly utilized to handle complex random processes, outperforming traditional Markov models [6]. Numerous studies have applied HMM techniques to financial fraud detection. For instance, Khan et al. [9] employed an HMM-based approach to detect card owners' behaviors by observing incoming transactions, utilizing clustering to differentiate between fraudulent and non-fraudulent patterns. Agrawal et al. [10] introduced a hybrid method combining HMM with Genetic algorithms (GA) to identify credit card fraudulent transactions, effectively preserving transaction logs and determining clustering thresholds. Similar approaches were adopted for internet banking fraud detection, emphasizing user identification and monitoring illicit behaviors. Additionally, HMM was employed to improve fraud detection in credit card operations, demonstrating enhanced accuracy and reduced false-positive rates. Simulation experiments utilizing HMM and K-means methods further validated the effectiveness of the model in bank fraud detection. Overall, HMM-based techniques prove beneficial in enhancing the efficiency and accuracy of credit card fraud detection, often in conjunction with clustering methods like K-means [6].

A Hidden Markov Model (HMM) consists of a finite set of states, each associated with a probability distribution. Transitions between these states are governed by transition probabilities. While each state can generate observations according to its associated probability distribution, only these observations are visible to an external observer; the actual states themselves remain "hidden," giving the model its name. This structure allows HMMs to be particularly effective for detecting fraudulent credit card transactions. An important advantage of using HMMs in fraud detection is their ability to significantly reduce the number of false positives—transactions that are incorrectly flagged as fraudulent by the system, even though they are legitimate. Bhusari et.al have described a HMM that can be employed in credit card fraud detection. Their model looks back at the last ten transactions after having quantized the transaction amount into 3 price ranges (high, medium, low). This model's false positive rate is 7% and effectiveness is 80% over a broad range in input data [7].

Artificial Neural Network (ANN):

Artificial Neural Networks (ANNs) are computational models that mimic the structure and function of the human brain [8]. Neural networks are composed of interconnected nodes, known as neurons, that are arranged in layers. The network facilitates the transmission of information, as each individual neuron receives input, processes it, and transmits output to the subsequent layer. Artificial neural networks (ANNs) acquire knowledge by modifying the synaptic connections between neurons, a procedure referred to as training. This training process entails iteratively modifying the network's parameters in order to decrease the disparity between expected and actual outputs. It commonly employs techniques such as gradient descent and backpropagation. Artificial neural networks (ANNs) are versatile tools that can be applied in various fields such as pattern recognition, classification, regression, forecasting, and decision-making. They possess exceptional proficiency in managing intricate, non-linear connections within data [8].

Artificial Neural Networks (ANNs) are highly effective in fraud detection within financial services due to their ability to recognize complex patterns, handle high-dimensional data, and adapt to evolving fraud tactics. They are used across various applications, including credit card fraud detection, loan and credit scoring, and account monitoring. ANNs offer significant advantages in accuracy, adaptability, and handling non-linear relationships, making them a valuable tool in combating financial fraud. Asha et.al. have shown that they are superior to SVM's and k-NN algorithms to detect fraudulent credit card transactions and tend to achieve an accuracy of 99.92% [9]. The conflict between Asha's et.al. results and Gedela et. al results suggest that the datasets are the culprit. As a rule of thumb ANN will likely outperform SVM in scenarios involving large, complex datasets with intricate patterns, such as in deep fraud detection models where hierarchical feature extraction is beneficial whereas SVM's may be preferable for smaller to medium-sized datasets or scenarios where interpretability and computational efficiency are more critical.

KNN Algorithm:

The K-Nearest Neighbors (KNN) algorithm is an uncomplicated yet potent supervised machine learning technique employed for applications including classification and regression [10]. The operation of this system is based on the idea of similarity, where data points that are similar are expected to belong to the same class or have comparable output values. KNN, a classification algorithm, determines the K closest neighbors to a new data point by utilizing a selected distance measure, such as Euclidean distance, from the training dataset. The new data point is assigned to the class that the majority of these neighbors belong to. When using regression, the K-nearest neighbors (KNN) algorithm calculates the mean (or weighted mean) of the output values of the K closest neighbors in order to make a prediction for the new data point. The selection of K, which represents the number of neighbors, is of utmost importance and has the potential to greatly impact the success of the algorithm. A low value of K can result in overfitting, whereas a high value of K can lead to underfitting. KNN is a non-parametric algorithm, which implies that it does not rely on any assumptions regarding the underlying distribution of the data. It is referred to as a lazy learning algorithm since it does not require explicit training. Instead, it memorizes the full training dataset, which might be computationally expensive for large datasets. Although KNN is a simple algorithm, it can be highly effective, particularly for datasets that have clearly defined clusters or where the decision boundary is non-linear. Nevertheless, its efficacy may diminish in environments with a high number of dimensions or when confronted with datasets that are uneven [10].

KNN works well when the fraud patterns are static and datasets are less complex and fairly small [11]. When the datasets are less complex, KNN usually performs better than SVM [9]. Hence KNN, is usually employed to profile and evaluate the risk of loan seekers. By analyzing historical data on loans and defaults, KNN can help identify applications that resemble those from high-risk or fraudulent cases. This enables financial institutions to prioritize further investigation of suspicious applications.

Bayesian Method:

The Bayesian method is a statistical technique that uses previous information and observed evidence to model uncertainty and make predictions [12]. The central focus is on Bayes' theorem, which offers a structure for revising our convictions on the probability of certain events as fresh data emerges. Bayesian inference commences by establishing initial beliefs or assumptions on the probability distribution of parameters or hypotheses. By analyzing fresh data, we revise our beliefs using Bayes' theorem to derive a posterior distribution. This distribution represents our revised comprehension of the parameters or hypotheses based on the observed data. The Bayesian approach is especially valuable in scenarios where we possess a scarcity of data or previous information that can contribute to our forecasts. It enables the integration of uncertainty and variability into our models, offering a more sophisticated and adaptable approach in contrast to conventional frequentist methods. Bayesian methods are extensively utilized in several domains, such as machine learning, statistics, and decision-making. They are applied in tasks such as estimating parameters, testing hypotheses, selecting models, and creating predictive models. Although Bayesian inference has benefits, it can require significant processing resources, particularly when dealing with intricate models or datasets with a large number of dimensions. Nevertheless, the development of computer tools like Markov chain Monte Carlo (MCMC) and variational inference has significantly enhanced the accessibility and widespread application of Bayesian methods [12].

Bayesian methods are especially useful in fraud detection when prior knowledge is available, when handling uncertainty and variability, and when working with small or imbalanced datasets. They allow for the incorporation of expert insights and prior information, adapt dynamically as new data emerges, and capture complex relationships between features. Additionally, Bayesian techniques offer better interpretability and probabilistic reasoning, making them valuable for understanding the likelihood of fraud in diverse and evolving scenarios. Amit et.al found that Naïve Bayes is superior to Random Forest, Logistic Regression and SVM when the dataset was imbalanced. The accuracy was 80.4% for the imbalanced dataset [13]. In finance, imbalanced datasets can occur in various contexts, particularly where the events of interest are rare compared to the majority of observations. An example could be high-frequency trading event detection. The majority class would be normal trading activity and the minority would be significant price manipulations or rare events. This imbalance between regular trading and rare significant events can cause problems for modelling and prediction and this is where Bayesian methods can help.

Decision Tree:

A Decision Tree is a flexible and easy-to-understand machine learning technique utilized for tasks including classification and regression [14]. The structure is akin to a flowchart, with core nodes representing features or attributes, branches representing decisions based on those features, and leaf nodes representing the outcome or class label. The process of constructing a decision tree entails recursively dividing the data into partitions based on factors that optimize the uniformity of the target variable within each partition. The procedure is commonly directed by metrics such as Gini impurity or information gain for classification jobs, and mean squared error or variance reduction for regression activities. Decision Trees offer several benefits, including their straightforwardness, comprehensibility, and capability to handle both numerical and categorical data without requiring considerable data preprocessing. Neural networks have the ability to capture complex correlations and interactions between different features, which makes them highly reliable and efficient in various applications. Nevertheless, decision trees are susceptible to overfitting, particularly when they are deep or intricate. In order to address this problem, strategies such as pruning, imposing a limit depth, and utilizing ensemble methods such as Random Forests or Gradient Boosting can be implemented. Decision Trees are a potent tool in machine learning that can be used by both novices and professionals. They provide a clear and effective method for addressing classification and regression problems [14].

Decision trees have certain advantages for fraud detection compared to other machine learning algorithms. Decision trees offer high interpretability. Each decision path can be easily traced, and the reasoning behind predictions is transparent, making it easier to understand why a particular transaction or application was flagged as fraudulent. Techniques like neural networks or ensemble methods (e.g., Random Forests) are often seen as "black boxes," providing less intuitive explanations for their decisions. They also handle both categorical and numerical data without the need for extensive preprocessing and they split nodes based on feature values directly while other algorithms such as SVM's and neural networks may require extensive feature scaling or transformation to handle categorical features effectively. This makes them faster to train especially on smaller datasets. Another advantage decision trees have over traditional fraud detection methods is that they naturally handle non-linear relationships between features, making them well-suited for capturing complex patterns in fraud detection data. Linear models (e.g., Logistic Regression) assume linear relationships and may not perform well with complex, non-linear fraud patterns unless combined with polynomial features or kernel methods. Thus, decision trees are ideal when stakeholders need clear explanations for why certain transactions or applications are flagged. They are excellent for detecting suspicious activity transactions and activity. AML and terrorist financing regulations require depository institutions to file suspicious activity records (SAR) if any transaction or customer activity is suspected to have violated the law. Decision trees help here to flag certain activities or transactions and provide an explanation why that activity or transaction was flagged. This in turn can be helpful to file the reports with the Federal agency which in this case would be FinCEN.

Genetic Algorithms (GAs):

Genetic Algorithms (GAs) are optimization methodologies that draw inspiration from the mechanisms of natural selection and genetics [15]. They are a form of evolutionary algorithm that emulates the mechanism of natural selection in order to seek out the optimal solution to a given problem. A genetic algorithm involves a population of potential solutions, which are represented as chromosomes or individuals, undergoing iterative improvement across subsequent generations. Every member of the population represents a potential solution to the problem. These solutions are usually recorded as a sequence of binary digits, real numbers, or other data structures. The algorithm undergoes a repetitive process of selection, crossover, mutation, and evaluation to gradually improve the population towards more optimal solutions. During the process of selection, individuals are picked based on their level of fitness, which is a measure of their performance on the given challenge. Crossover is the process of merging genetic material from specific individuals to produce new children. Mutation is a process that adds random alterations to the genetic material of offspring in order to preserve genetic diversity. Following the creation of a new population via crossover and mutation, each individual is assessed to ascertain its level of fitness. The most physically capable individuals are subsequently chosen to comprise the succeeding generation, and this cycle persists until a termination criterion is fulfilled, such as attaining a maximum number of generations or accomplishing an acceptable solution. Genetic algorithms are highly effective in addressing optimization problems that involve extensive search spaces, intricate objective functions, and numerous local optima. They have effectively been utilized in various domains, including as optimization, machine learning, scheduling, and design. Although genetic algorithms are successful, they can be computationally demanding, particularly for issues that involve large populations and high-dimensional search spaces. Furthermore, the efficacy of genetic algorithms is greatly influenced by the selection of parameters, including population size, mutation rate, and selection technique, which may need to be adjusted for best outcomes [15].

Genetic algorithms are basically used to optimize and tune feature subsets. In credit card fraud detection, GA's might be used to select the most relevant features from a large dataset to improve the efficiency and accuracy of fraud detection models. They could also be used to optimize the hyperparameters of a machine learning model used to detect fraudulent transactions, enhancing model performance. Lastly, in systems that rely on rule-based approaches, GA's can help in refining rules to adapt to new fraud patterns. Thus, genetic algorithms complement the existing models and improve the accuracy of fraud detection [16].

Ensemble Methods:

Ensemble techniques are a technique that combines different models in order to enhance predictive performance. Ensemble approaches can enhance the robustness of predictions by combining the forecasts of multiple independent models, such as decision trees or neural networks, thereby mitigating overfitting. Two well-known ensemble methods are Random Forests, which aggregate many decision trees, and Gradient Boosting, which gradually constructs models to rectify errors created by preceding models [17]. Ensemble methods, such as Random Forests, Gradient Boosting Machines, AdaBoost (adaptive boosting), stacking, and voting classifiers, are highly effective in fraud detection due to their ability to improve accuracy, manage imbalanced data, and enhance robustness. By combining multiple models, these methods reduce overfitting, capture complex fraud patterns, and adapt to new fraud strategies. However, they can be computationally intensive and may complicate model interpretability. Despite these challenges, ensemble methods are widely used in financial services for tasks like credit card fraud detection, loan default prediction, and market manipulation detection, leveraging their diverse strengths to achieve superior performance and reliability.

Clustering-Based Methods:

Clustering-based techniques strive to categorize comparable data points into clusters, with the objective of revealing hidden patterns or structures in the data. K-means clustering and hierarchical clustering are methods that divide data into clusters based on similarity measures, such as distance metrics. Clustering-based techniques are extensively employed in the fields of data exploration, pattern identification, and segmentation tasks [18]. Clustering-based methods are widely used in financial fraud detection to identify unusual patterns and anomalies that may indicate fraudulent activities. Techniques such as K-Means, DBSCAN, Hierarchical Clustering, and Isolation Forest help in grouping similar data points and detecting outliers that deviate from normal behavior. For instance, K-Means can highlight transactions that fall outside typical clusters, while density-based spatial clustering of applications with noise (DBSCAN) identifies transactions in low-density areas as potential fraud. Hierarchical Clustering builds a structure of clusters to spot anomalies, and Isolation Forest isolates transactions quickly to flag anomalies. These methods are effective in detecting novel or unknown fraud patterns, especially when labeled data is scarce, though they may require careful parameter tuning and computational resources [19].

Logistic Regression:

Logistic regression is a statistical technique employed for binary classification tasks, wherein the objective is to forecast the likelihood that an observation belongs to a specific class. Logistic regression differs from linear regression in that it predicts the link between independent factors and the log-odds of the dependent variable belonging to a specific class, rather than predicting continuous outcomes. Due to its simplicity, interpretability, and efficacy in predicting binary outcomes, it is extensively utilized in many domains such as medical, finance, and marketing [20]. Logistic regression has been studied alongside ANN, Naïve Bayes, Decision trees, SVM's etc. to detect credit card fraud. Sahin et.al. reported that ANN models were superior than logistic regression models but the performance was similar [21]. Atchaya et.al. used a novel logistic regression model and compared it with Naïve Bayes method. They found that their novel logistic regression algorithm was 93.59% more accurate than then Naïve Bayes algorithm [22]. As logistic regression is used primarily for binary classification problems, it can be used for risk assessment in financial compliance assessments, credit risk assessments etc. as it can accurately determine the probability of a borrower defaulting on a loan based on credit history, income, employment status, and other financial indicators.

3. Applications of Big Data in Financial Fraud Detection

Transaction monitoring: Real-time detection of fraudulent activities in financial transactions:

Transaction data is pivotal for detecting financial fraud, given its vast volume and rapid generation. To discern fraudulent behavior patterns, advanced analytics are imperative. Big Data technologies facilitate the processing of extensive transaction datasets, enabling the identification of anomalies indicative of potential fraud. Understanding user behavior is critical in identifying deviations from normal patterns, with Big Data Analytics enabling the analysis of user interactions to detect potential fraud. This involves scrutinizing login locations, transaction frequencies, and deviations from typical behavior, contributing to a nuanced understanding of user activities [23]. Integration of external data sources enhances financial institutions' analytical capabilities. Social media activity, public records, and other external sources offer valuable context and insights into financial behavior. Big Data technologies facilitate the integration and analysis of these diverse datasets, enabling a comprehensive fraud detection approach. Machine learning algorithms are pivotal in financial fraud detection, autonomously learning patterns and anomalies from historical data. Supervised learning models, such as decision trees and support vector machines, distinguish between legitimate and fraudulent transactions. Unsupervised learning models, including clustering algorithms, identify anomalies without prior training, adapting to evolving fraud schemes. Predictive modeling focuses on forecasting and identifying anomalies in real-time. By creating models based on historical data, financial institutions predict potential fraudulent behavior by recognizing deviations from established patterns. Continuous refinement of these models enhances predictive accuracy, bolstering defenses against emerging fraud trends [24]. Real-time processing capabilities are imperative due to the urgency of fraud detection. Traditional batch processing may not suffice in identifying and preventing fast-paced, sophisticated fraud schemes. Real-time analytics, powered by Big Data technologies, enable immediate identification of suspicious activities, reducing the time window for potential losses. Addressing challenges such as latency and scalability, Big Data technologies leverage distributed computing frameworks like Apache Flink and Apache Kafka to ensure timely analysis of streaming data without compromising accuracy [25]. In essence, the integration of Big Data Analytics in financial fraud detection represents a revolutionary approach. By harnessing diverse data sources, employing advanced machine learning algorithms, and embracing real-time processing, financial institutions fortify their defenses against an ever-evolving landscape of financial fraud.

Identity verification: Biometric authentication, behavior analysis, and device fingerprinting

Biometric authentication, behavior analysis, and device fingerprinting are essential components in the field of financial fraud detection, as they help protect against fraudulent activity. Biometric data, such as fingerprints, facial traits, provides a very safe method for confirming a user's identification. Biometric authentication in financial transactions ensures that only authorized individuals are able to access accounts or carry out transactions, hence minimizing the danger of identity theft or impersonation [26]. Surveillance of user conduct during financial transactions aids in detecting irregularities that could suggest fraudulent behavior. Behavior analysis algorithms have the ability to identify and signal any deviations, which then prompts alarms for additional study [27]. Each device possesses distinct attributes, including its IP address, operating system, browser version, and hardware configuration. Device fingerprinting identifies and logs these characteristics in order to generate a distinct profile for every device that accesses financial services. This allows for the identification of potentially suspicious actions, such as logging into an account from devices that are not recognized or have been previously marked as suspicious. These actions may suggest unauthorized access or an attempt to take control of the account [28]. Financial institutions can greatly improve their ability to identify and prevent fraudulent activities by incorporating biometric authentication, behavior analysis, and device fingerprinting into their fraud detection systems. This will help protect the interests of their customers and reduce financial losses.

Social network analysis: Detection of fraudulent networks and collusion schemes

Social network analysis (SNA) has become instrumental in understanding the intricate relationships and structures within various social systems. Its application extends to the identification and mitigation of fraudulent activities and collusion schemes, which pose significant challenges across multiple domains. By examining network metrics, community detection algorithms, and anomaly detection techniques, we can uncover suspicious patterns and behaviors within complex networks [29].

Network analysis plays a crucial role in preventing fraud by identifying and analyzing fraudulent activity patterns within networks. It involves examining relationships and interactions among entities like individuals, transactions, or accounts to detect anomalous behavior or fraud patterns. Network analysis provides several key capabilities for fraud prevention:

- Data Collection: Collecting data on transactions and user interactions allows network analysis tools to represent relationships between entities accurately.

- Graph Analysis Techniques: Centrality measures and anomaly detection algorithms identify nodes with high centrality or deviations from the norm, signaling potential fraudulent activities.

- Behavioral Analysis: Analyzing past interactions establishes a baseline of normal behavior, helping detect deviations indicative of fraud.

- Link Analysis: Examining relationships between entities uncovers hidden or indirect links between potentially fraudulent actors, aiding in the detection of fraud rings.

- Community Detection: Algorithms identify fraud rings through the detection of clusters representing organized fraudulent activities.

- Real-time Monitoring: Analyzing network data in real-time enables proactive detection of suspicious activities and improves predictive analysis capabilities.

- Integration with Other Security Measures: Network analysis is often integrated with rule-based systems, anomaly detection, and identity verification to create comprehensive fraud prevention strategies.

Fraud networks can take various forms, such as credit card fraud rings, identity theft networks, phishing networks, online scam networks, and cybercrime networks. Network analysis can uncover different types of fraud, including transaction and payment fraud, identity theft, credit card fraud, e-commerce fraud, phishing, account takeover, money laundering, and online scams. Financial institutions (FIs) utilize network analysis for fraud detection and prevention, anti-money laundering (AML) compliance, customer relationship analysis, credit risk assessment, operational risk management, cross-border transactions monitoring, customer due diligence (CDD), network-based credit scoring, and cybersecurity. Overall, network analysis provides valuable insights into relationships and behaviors within systems, enabling organizations to effectively detect and prevent fraudulent activities [29].

4. Challenges and Limitations

Data quality and imbalanced datasets

Decision trees, Bayesian networks, and support vector machines (SVM) are commonly used in fraud detection. However, they overlook a crucial aspect of fraud data: the imbalance between the number of valid records and fraudulent records. This imbalance poses a challenge as the dataset is skewed [30]. While fraud detection is typically seen as a binary classification issue, it actually presents itself as an n-class problem due to the unique nature of each fraudulent activity. The majority of transactions in the dataset are considered normal, with fraudulent ones potentially accounting for less than 0.1% of the total. Designing a model for this task poses challenges. If we solely evaluate performance based on accuracy, a model that always predicts "regular" transactions would achieve a notably high accuracy score. Challenges stemming from datasets with extremely small class ratios manifest in three distinct areas: Understanding and capturing the correlation properties of features for underrepresented classes during modeling and learning, identifying relevant feature class distinctions, such as recognizing unique features associated with each class, introducing significant bias to "standard" evaluation metrics, which are typically formulated for datasets with comparable class sizes [30].

Scalability and real-time processing requirements

The challenges associated with scalability and real-time processing in detecting fraudulent financial transactions involve the ability of the system to handle large volumes of data and analyze them rapidly to identify fraudulent activities as they occur [31]. Modern-day cloud systems and big data analytics frameworks are extensively utilized in detecting fraudulent transactions. Nevertheless, as data volume grows and workloads and resources become more diverse, coupled with the dynamic nature of user requests, the uncertainties and intricacies of resource scheduling and service provisioning escalate significantly. This often leads to suboptimal resource utilization, compromised system reliability, and adverse effects on user-perceived performance. Resource scheduling involves matching demand, which consists of requests to allocate resources for running processes of specific tasks or applications, with supply, which consists of available resources on cluster nodes. As a result, the complexity of resource management is directly influenced by the number of concurrent tasks and the number of server nodes within a cluster. Furthermore, other factors, such as supporting resource allocation across multiple dimensions (such as CPU, memory, and local storage), enforcing fairness and quota constraints among competing applications, and scheduling tasks in proximity to data, also contribute to the complexity. To address the growing volume of running tasks and the scale of clusters, computing systems operating at massive scales must prioritize scalability concerns [31].

Privacy and regulatory compliance concerns

In recent years, both academic literature and market solutions have primarily focused on accumulating and consolidating large volumes of transaction data, including user behavior data, and enhancing algorithms aimed at detecting fraudulent activities [32]. Concurrently, legislation within the European Union, such as PSD2, has been enacted to mandate stakeholders to detect fraud. However, while the legislation outlines this legal requirement broadly, market solutions are diversifying in terms of the data they collect and their efforts to consolidate data to achieve more precise outcomes. This presents an unexplored issue in both academic literature and legislative discussions: the privacy implications associated with profiling and aggregating data for fraud detection purposes, and the accountability of stakeholders in detecting fraud within the framework of their obligations under data protection laws. Determining the level of intrusiveness of data processing is crucial in the context of fraud detection, the objective is to safeguard an individual's financial account. At first glance, it may seem justifiable to process various types of data for this security purpose. However, it's imperative to weigh the benefits of fraud detection against the potential impact on individuals. This includes the collection of extensive data on their behavior, including financial activities, sharing such data with other entities like credit/payment institutions or merchants, and the direct or indirect consequences on individuals, such as discrimination and bias, the denial of services, and the creation of inaccurate profiles [32].

Adversarial attacks and evasion techniques used by fraudsters

Ensuring the security of transactional systems is a top priority for all institutions handling transactions. This is essential for safeguarding their businesses from cyberattacks and fraudulent activities. Adversarial attacks, which have been demonstrated as effective in deceiving image classification models, can also be utilized with other forms of data. Adversarial attacks involve creating modified inputs, known as adversarial examples, that manipulate the Artificial Intelligence (AI) system to produce incorrect outputs beneficial to the attacker. Fraudsters continuously devise novel methods to deceive these systems, a phenomenon referred to as concept drift. Consequently, a fraud detection system often demands substantial maintenance to keep up with these evolving tactics. Fraudsters employ various methods to circumvent fraud detection systems. Among these tactics, adversarial attacks represent a cutting-edge approach that could elevate financial fraud to a more sophisticated level. The objective of adversarial attacks is to create inputs, known as adversarial examples, that closely resemble legitimate data but are misclassified by the machine learning model. Recent studies have demonstrated the significant effectiveness of algorithms in generating adversarial examples to deceive Machine Learning models, particularly Deep Neural Networks (DNNs) in Image Recognition. This poses a concern for numerous applications reliant on these technologies, such as self-driving vehicles or facial recognition systems. Adversarial examples arise due to the disparity between how humans and machines perceive knowledge and relationships among visual elements in object recognition tasks. This disparity allows attackers to manipulate the pixels of an image in a manner imperceptible to humans, yet capable of misleading an image classifier into making an incorrect interpretation. For example, by subtly altering the color of a few pixels, an attacker can prompt an image classifier to confidently identify a gibbon in an image depicting a panda. By knowing how a bank or a financial institution employs different machine learning algorithms, it is very easy to come up with a set of inputs that "trains" the machine to incorrectly classify a fraudulent transaction as a regular transaction [33].

5. Evaluating the Effectiveness and Limitations of Big Data Analytics in Fraud Detection: Real-World Examples

This discussion will explore the effectiveness of big data analytics in detecting fraud, examining case studies to showcase successful applications and the impact of analytical tools. At the same time, it will address the limitations by focusing on areas which requires the current technology to catch-up to. By evaluating these aspects, we can gain a comprehensive understanding of how big data analytics contributes to fraud detection in the real-world and where improvements are needed to enhance its effectiveness.

Big data based fraud risk management at Alibaba:

Alibaba has advanced fraud risk management by developing a real-time payment fraud prevention system called CTU (Counter Terrorist Unit), one of the most sophisticated in China. CTU tracks and analyzes user behavior to identify suspicious activities and apply appropriate measures through intelligent arbitration. Supporting this system, fraud risk models use big data and statistical techniques to assess risk associated with accounts, users, and devices by analyzing numerous attributes and their correlations with fraud. These models are integral to various Alibaba procedures, including account opening, identity verification, order placement, and transaction monitoring, ensuring comprehensive fraud detection and management [34].

Over the past decade, Alibaba has experienced rapid growth, with daily transaction volumes soaring from under 10,000 in 2005 to 188 million by November 2013. This expansion has led to the development of an advanced data platform that supports various business functions, including targeted marketing, customer service, and fraud prevention. Alibaba's fraud risk management system, a key part of its online payment services, employs a multi-layered risk prevention framework to secure transactions. This framework includes five layers: Account Check, Device Check, Activity Check, Risk Strategy, and Manual Review. Each layer performs specific checks to detect and prevent fraudulent activities. For instance, Account Check evaluates past suspicious activities, Device Check assesses transaction patterns from the same device, and Activity Check analyzes historical behavior and links among accounts. Risk Strategy aggregates results to make final decisions, and Manual Review scrutinizes suspicious cases that cannot be automatically resolved [34].

Alibaba utilizes the CTU (Counter Terrorist Unit), a real-time payment fraud monitoring system launched in 2005, which has evolved to handle various fraud types such as money laundering and marketing fraud. The CTU processes hundreds of event types, making real-time risk decisions within milliseconds. A key component is the RAIN risk model (Risk of Activity, Identity, and Network), which evaluates risk across three dimensions: -

- Activity Analysis: Algorithms examine historical transaction data, user behavior patterns, and device usage. This helps identify deviations from normal behavior that may indicate fraudulent activity.
- Identity Verification: Models assess the legitimacy of user identities by analyzing patterns associated with known fraud cases, cross-referencing multiple data points to confirm the authenticity of identities.
- Network Analysis: Alibaba employs graph theory and network-based analysis to detect fraudulent networks. By mapping connections among accounts, devices, and transactions, these algorithms uncover hidden relationships that may indicate organized fraud schemes [34].

Alibaba leverages a variety of ML algorithms such as decision trees, random forest, logistic regression, graph theory and network-based analysis to detect fraudulent activity. The model employed depends entirely on the dimension described above. For instance, to detect fraudulent networks, network-based analysis helps reveal hidden connections among accounts, which is crucial as fraudsters increasingly attempt to obscure their activities by avoiding detectable patterns like shared names or addresses. Network analysis is essential for fraud prevention as it helps uncover and assess patterns of fraudulent activity within networks. This technique involves scrutinizing the relationships and interactions among entities such as individuals, transactions, or accounts to identify unusual behaviors or signs of fraud [34].

In practical applications, accounts are represented as nodes in a graph, while connections between these nodes (such as shared IP addresses or phone numbers) are shown as edges. For instance, if multiple accounts share common details, network analysis can expose these relationships, revealing connections between different groups of accounts. Graphs illustrate how shared attributes among accounts can highlight fraudulent networks. To manage the complexity of real-world data, advanced graph algorithms and specialized storage solutions are employed. Betweenness centrality, a concept from network analysis, identifies key nodes that bridge different parts of the network, providing crucial insights into account connections and helping prevent fraudsters from establishing undetected networks [34].

At Alibaba, the model-building process has become highly refined through extensive development and iterative improvement. The approach begins with selecting two types of samples: "white" samples, representing low-risk or good entities, and "black" samples, which are deemed high-risk or problematic. The goal is to develop a model that can effectively distinguish between these two categories. To create the model, Alibaba collects behavior and activity data from both white and black samples to generate initial variables. These variables are derived from aggregating and abstracting data, with only the most validated variables being used in model construction. Based on Alibaba's success with big data to detect fraud, decision tree algorithms like C5.0 and Random Forest are used. These algorithms are preferred due to their ability to balance bias and variance without assuming a specific data distribution, unlike traditional data models. Once a model shows strong performance in differentiating between good and bad samples, it must be validated to ensure its applicability across various scenarios. Successful models, which demonstrate effectiveness and efficiency through rigorous testing and validation, are then deployed into the production environment. In this setting, they integrate with Alibaba’s

Counter Terrorist Unit (CTU) and other fraud prevention strategies and rules to enhance overall fraud risk management [34].

Thus, different machine learning algorithms are employed to detect fraudulent activity based on the activity and the specific type of fraud. In the case of Alibaba, C5.0 and Random Forest are used because they handle data variability without making assumptions about the data distribution, unlike some other statistical models. Validation of the model is crucial such that it works well in different scenarios and is not overfitted to the specific data it was trained on. Effective models are those that perform well across a range of test cases and scenarios. Finally, we also learn from Alibaba's fraud detection tool that network-based analysis is used for risk control and to uncover fraudulent networks and keep a check on the level of fraud happening in the system.

Synthetic Identity Fraud:

Synthetic identity theft involves using a stolen Social Security number (SSN) in conjunction with fabricated personal details—such as a name, date of birth, and contact information—to create a new, false identity. This type of fraud is particularly challenging for traditional monitoring systems to detect because it often targets vulnerable populations like children, the elderly, and homeless individuals, who are less likely to regularly check their credit history. Criminals might either steal an SSN or buy one from the dark web, and then combine it with invented information through identity compilation, manipulate existing personal details, or completely fabricate new ones [35].

The fraudulent identities generated through these methods are frequently used to commit various forms of financial fraud, including applying for loans, opening bank accounts, and accessing credit cards. Synthetic identities can also be employed to file false tax returns, obtain medical services, or claim unemployment benefits. Detecting synthetic identity theft is challenging because these fabricated identities can appear legitimate, making it difficult for traditional big data tools to flag them as suspicious. Moreover, victims of synthetic identity theft may experience fragmented credit files, where negative credit actions associated with the synthetic identity are mistakenly linked to their real credit history, causing long-term damage to their credit scores [35]. The widespread availability of online generative AI tools is enabling scammers to transform the fraud landscape. These AI technologies are facilitating the rapid and widespread execution of deepfake and spoofing attacks. When such attacks are successful, they can escalate quickly and spread through criminal networks or the dark web.

Synthetic identity fraud presents a significant challenge for traditional machine learning algorithms for several reasons. These fabricated identities often appear genuine, incorporating believable personal details like names, addresses, and Social Security numbers, which can blend seamlessly with legitimate accounts. Traditional models, which rely on identifying anomalies or deviations from standard patterns, may struggle to distinguish these false identities from real ones. Additionally, synthetic identities often lack substantial historical data or past behaviors that could trigger suspicion, complicating the detection process. The gradual, subtle nature of synthetic identity construction further limits the effectiveness of models that look for abrupt behavioral changes.

Moreover, synthetic identities may use convincing but unremarkable features that traditional models may not flag as suspicious, and the constantly evolving tactics of fraudsters can outpace static models trained on historical data. The fragmentation of data sources and the vast volume of data associated with synthetic identities also pose difficulties for traditional algorithms, which may struggle to process and analyze such extensive information effectively. Furthermore, traditional models can experience high rates of false positives, misidentifying legitimate accounts as fraudulent due to the subtlety of synthetic identities.

6. Conclusion

The integration of Big Data analytics into financial fraud detection marks a significant advancement in ensuring the integrity of the financial industry. To enhance big data analytics for fraud detection, several strategies can be employed to address current limitations and leverage emerging technologies. Advanced machine learning techniques, such as deep learning and sophisticated anomaly detection, can reveal complex patterns in large datasets that traditional algorithms might miss. Integrating data from multiple sources, including transaction logs, social media, and public records should provide a more comprehensive view of activities and relationships, while improving data quality through cleansing and enrichment enhances fraud detection. The integration of real-time analytics and response systems, facilitated by stream processing frameworks, enables quicker detection and reaction to fraudulent activities. Automated response systems can expedite the process by flagging anomalies or alerting users. Blockchain technology also offers benefits like immutable transaction records and decentralized identity verification, which improve transparency and security. Additionally, refining behavioral analytics with dynamic user profiles and behavioral biometrics provides deeper insights into normal and abnormal behaviors. This paper has examined various facets of big data analytics in financial fraud detection, discussing its applications, benefits, challenges, and future trends, and providing valuable insights for practitioners and researchers. Challenges in big data analytics include data privacy, quality issues, limited data sharing support, load balancing, and data aggregation. Future research should tackle these challenges and develop a framework for aggregating data from diverse sources, applying analytical tools and business intelligence to detect transaction anomalies.

References

[1] N. P. O. Shoetan, N. A. T. Oyewole, N. C. C. Okoye, and N. O. C. Ofodile, “REVIEWING THE ROLE OF BIG DATA ANALYTICS IN FINANCIAL FRAUD DETECTION,” *Finance & Accounting Research Journal*, vol. 6, no. 3, pp. 384–394, Mar. 2024, doi: 10.51594/farj.v6i3.899.

[2] B. K. Jha, G. G. Sivasankari, and K. R. Venugopal, "Fraud Detection and Prevention by using Big Data Analytics," IEEE Xplore, Mar. 01, 2020. https://ieeexplore.ieee.org/abstract/document/9076536

[3] Balasupramanian. N, Ben George Ephrem, and Imad Salim Al-Barwani, "User Pattern Based Online Fraud Detection and Prevention using Big Data Analytics and Self Organizing Maps", ICICICT, pp. 691-694, 2017.

[4] A. Ali et al., "Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review," Applied Sciences, vol. 12, no. 19, p. 9637, 2022, doi: https://doi.org/10.3390/app12199637.

[5] Bhargavi Gedela and P. R. Karthikeyan, "Credit card fraud detection using support vector machine algorithm in comparison with various machine learning algorithms to measure accuracy, sensitivity, specificity, precision and f-score," AIP Conference Proceedings, Jan. 2023, doi: https://doi.org/10.1063/5.0150792.

[6] Pradeep, G.; Ravi, V.; Nandan, K.; Deekshatulu, B.L.; Bose, I.; Aditya, A. Fraud Detection in Financial Statements Using Evolutionary Computation Based Rule Miners. In Proceedings of the International Conference on Swarm, Evolutionary, and Memetic Computing, Hyderabad, India, 18–19 December 2015; pp. 239–250.

[7] Bhusari, V.; Patil, S. Study of Hidden Markov Model in credit card fraudulent detection. In Proceedings of the 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), Coimbatore, India, 29 February–1 March 2016; pp. 1–4.

[8] D. Anderson and G. McNeill, ‘‘Artificial neural networks technology,’’ Kaman Sci. Corp., vol. 258, no. 6, pp. 1–83, 1992.

[9] A. RB and S. K. KR, "Credit Card Fraud Detection Using Artificial Neural Network," *Global Transitions Proceedings*, vol. 2, no. 1, Jan. 2021, doi: https://doi.org/10.1016/j.gltp.2021.01.006.

[10] A. Moldagulova and R. Bte. Sulaiman, "Using KNN algorithm for classification of textual documents," May 2017, doi: 10.1109/icitech.2017.8079924.

[11] P. Raghavan and N. E. Gayar, "Fraud Detection using Machine Learning and Deep Learning," *IEEE Xplore*, Dec. 01, 2019. https://ieeexplore.ieee.org/document/9004231

[12] M. E. Glickman and D. A. Dyk, "Basic Bayesian Methods," in *Methods in molecular biology*, 2007, pp. 319–338. doi: 10.1007/978-1-59745-530-5_16.

[13] A. Gupta, M. C. Lohani, and M. Manchanda, "Financial fraud detection using naive bayes algorithm in highly imbalance data set," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 24, no. 5, pp. 1559–1572, Jul. 2021, doi: https://doi.org/10.1080/09720529.2021.1969733.

[14] S. Suthaharan, “Decision Tree Learning,” in *Integrated series on information systems/Integrated series in information systems*, 2016, pp. 237–269. doi: 10.1007/978-1-4899-7641-3_10.

[15] M. Srinivas and L. M. Patnaik, “Genetic algorithms: a survey,” *Computer*, vol. 27, no. 6, pp. 17–26, Jun. 1994, doi: 10.1109/2.294849.

[16] M. H. Özçelik, E. Duman, M. Işik, and T. Çevik, “Improving a credit card fraud detection system using genetic algorithm,” IEEE Xplore, Jun. 01, 2010.

[17] T. G. Dietterich, “Ensemble Methods in Machine Learning,” in *Lecture notes in computer science*, 2000, pp. 1–15. doi: 10.1007/3-540-45014-9_1.

[18] S. Kaushik, “Clustering | Different Methods, and Applications (Updated 2024),” *Analytics Vidhya*, Jun. 13, 2024. https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/

[19] A.S. Sabau, Survey of clustering based financial fraud detection research, Inform. Econ. 16 (1) (2012) 110–122.

[20] M. P. LaValley, ‘‘Logistic regression,’’ Circulation, vol. 117, no. 18, pp. 2395–2399, May 2008, doi: 10.1161/CIRCULATIONAHA.106.682658

[21] Sahin, Y.; Bulkan, S.; Duman, E. A cost-sensitive decision tree approach for fraud detection. Expert Syst. Appl. 2013, 40, 5916–5923.

[22] P. Atchaya and K. Somasundaram, ‘‘Novel logistic regression over Naive Bayes improves accuracy in credit card fraud detection,’’ J. Surv. Fisheries Sci., vol. 10, no. 1S, pp. 2172–2181, 2023.

[23] Malini, N.; Pushpa, M. Analysis on credit card fraud identification techniques based on KNN and outlier detection. In Proceedings of the 2017 third international conference on advances in electrical, electronics, information, communication and bio-informatics (AEEICB), Chennai, India, 27–28 February 2017; pp. 255–258.

[24] “The Role of Big Data in Fraud Detection and Prevention for Payment Providers,” paylinedata.com. https://paylinedata.com/blog/big-data

[25] F. Carcillo, A. Dal Pozzolo, Y.-A. Le Borgne, O. Caelen, Y. Mazzer, and G. Bontempi, “SCARFF : A scalable framework for streaming credit card fraud detection with spark,” Information Fusion, vol. 41, pp. 182–194, May 2018, doi: https://doi.org/10.1016/j.inffus.2017.09.005.

[26] Nurul Afnan Mahadi, Mohamad Afendee Mohamed, Amirul Ihsan Mohamad, Mokhairi Makhtar, Mohd Fadzil Abdul Kadir, and Mustafa Mamat. 2018. A survey of machine learning techniques for behavioral-based biometric user authentication. Recent Advances in Cryptography and Network Security (2018), 43–54.

[27] R. Rieke, M. Zhdanova, J. Repp, R. Giot, and C. Gaber, “Fraud Detection in Mobile Payments Utilizing Process Behavior Analysis,” *2013 International Conference on Availability, Reliability and Security*, Sep. 2013, doi: https://doi.org/10.1109/ares.2013.87.

[28] N. Nikiforakis, A. Kapravelos, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna, “Cookieless Monster: Exploring the Ecosystem of Web-Based Device Fingerprinting,” *2013 IEEE Symposium on Security and Privacy*, May 2013, doi: https://doi.org/10.1109/sp.2013.43.

[29] “Network Analysis: Unveiling Fraud Patterns with AI,” *DataVisor*. https://www.datavisor.com/wiki/network-analysis/

[30] Bian, Y.; Cheng, M.; Yang, C.; Yuan, Y.; Li, Q.; Zhao, J.L.; Liang, L. Financial fraud detection: A new ensemble learning approach for imbalanced data. In Proceedings of the 20th Pacific Asia Conference on Information Systems (PACIS 2016), Chiayi, Taiwan, 27 June–1 July 2016; p. 315.

[31] R. Yang et al., “Reliable Computing Service in Massive-Scale Systems through Rapid Low-Cost Failover,” IEEE Transactions on Services Computing, vol. 10, no. 6, pp. 969–983, Nov. 2017, doi: https://doi.org/10.1109/tsc.2016.2544313.

[32] L. G˘abudeanu, I. Brici, C. Mare, I.C. Mihai, M.C. Şcheau, Privacy intrusiveness in financial-banking fraud detection, Risks 9 (2021) 104, https://doi.org/10.3390/risks9060104.

[33] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, “The Limitations of Deep Learning in Adversarial Settings,” *IEEE Xplore*, Mar. 01, 2016. https://ieeexplore.ieee.org/abstract/document/7467366

[34] J. Chen, Y. Tao, H. Wang, and T. Chen, “Big data based fraud risk management at Alibaba,” *The Journal of Finance and Data Science*, vol. 1, no. 1, pp. 1–10, Dec. 2015, doi: https://doi.org/10.1016/j.jfds.2015.03.001.

[35] “Articles,” *www.equifax.com*. https://www.equifax.com/personal/education/identity-theft/articles/-/learn/synthetic-identity-theft/