Journal for Re Attach Therapy and Developmental Diversities
eISSN: 2589-7799
2023 December; 6 (10s) (2): 1954-1963

A Data-Driven Framework For Real-Time Fraud Detection In Financial Transactions Using Machine Learning And Big Data Analytics

Murali Malempati[1*]

[1*]Senior Software Engineer, Mastercard International INC, mmuralimalempati@gmail.com, ORCID: 0009-0001-0451-9323

Abstract

The rapid growth of electronic commerce and the gradual increase in customer confidence in the security of electronic payments have led to a persistent increase in the number of on-line transactions in the past years. Credit and debit cards account for the majority of the on-line payments. Consequently, the credit card financial ecosystem growth has been accompanied by a similar growth of illicit actions, by which con-men try to benefit from this huge financial exchange. Fraud detection is critical for credit institutions, merchants, and national services to minimize money losses. In recent years, several initiatives to enhance systems aimed at detecting fraudulent credit card transactions have been taken. Detecting fraud is very challenging since machine learning approaches rely on training sets limited to the observations that were available at the moment of the training. The newly triggered events, which were never observed in the training phase, can lead to severe issues such as alarm fatigue, in which frauds are detected only after a large amount of incurred losses.

Detecting fraud requires applying scalable learning techniques able to analyze the huge amount of streaming data generated by the transactions and able to mitigate the two main situations complicating the problem: the class imbalance and the concept drift since the world evolves and consequently the fraud patterns change. The need to detect frauds in real-time gives rise to several challenges. Recent advances in analytics, and the availability of open source solutions for storage, processing, and analytics of Big Data, have opened new perspectives for the real-time detection of frauds in massive amounts of transactions. In this paper, the SCAlable Real-time Fraud Finder (SCARFF) framework is presented. SCARFF is a distributed (anomaly) detection machine learning approach for the fraud detection that integrates Big Data tools both for the massive storage and processing of the transactions and for the predictive analysis.

SCARFF makes a contribution to the literature in four directions. First, the integration of the Hadoop and Spark ecosystems and of the sophisticated learning approach, addressing the inherent problems of imbalance, nonstationarity, and feedback latency, is a unique contribution. Second, the capability of handling a never-seen-before massive dataset of real credit card transactions is a unique achievement. Third, the formal description of the methods implemented to tackle data imbalance in real-time is presented. Fourth, the implementation in real-time of an ensemble learning engine capable of detecting credit card frauds at the rate of records−1, with large computational savings compared to batch implementations, is a further unique achievement.

Keywords: Big Data, Deep Learning, Fraud Detection, Financial Transactions, Linear Regression of Scoring, Real-Time.

1. Introduction

Fraud is described as abusing a profit organization's system. The prosperity of financial institutions and customer relationship management are linked to financial instruments. Individuals are deceived by promises of good fortune, investments, tax reductions, and so on. As a result, fraud detection is a fascinating research area that spans technology, business, and finance. Money laundering is an illegal activity aimed at concealing the ownership of illegally acquired property. Transactions can occur in various forms, and fraud can take several forms. Organizations base their decisions on analyses of vast volumes of data, or big data. Financial fraud datasets fall within this paradigm. Data is bigger than ever. Integration with emerging technologies enables cybersecurity protection strategies. Companies implementing financial systems are a key target for fraudsters. To make informed investment decisions, individuals and organizations rely on financial statements manipulated by fraudsters. Deterioration of security in mobile banking and other platforms leads to an increase in fraud. All of these problems require sustainable and scalable detection solutions.

Fraud is any act that deceives a victim to gain the victim’s resources—usually wealth or abundant access to the victim’s assets—and with the intent of keeping it. In other words, fraud is abuse of one's role in a profit organization’s system. There are many types and forms of fraud. To be a good fraudster, one must be an experienced expert. The difficulty and consequences of such acts are a huge barrier to individuals committing fraud. Using knowledge of others’ weaknesses is the foundation of fraud. Accordingly, fraud is a dynamic discipline. Since the time covers a span of decadence hundreds of thousands years, the techniques and knowledge of fraud change, evolve, and improve. A victim is without a chance to evade complicated acts of fraud. Machine learning, a great method for data analytics, can be used to detect fraud. There

are several types of machine learning: supervised, unsupervised, semi-supervised, and reinforcement learning. Any type of data can be input into machine learning, provided it is in a suitable form. Unlabeled data can be sequenced to approach the output labels step by step.

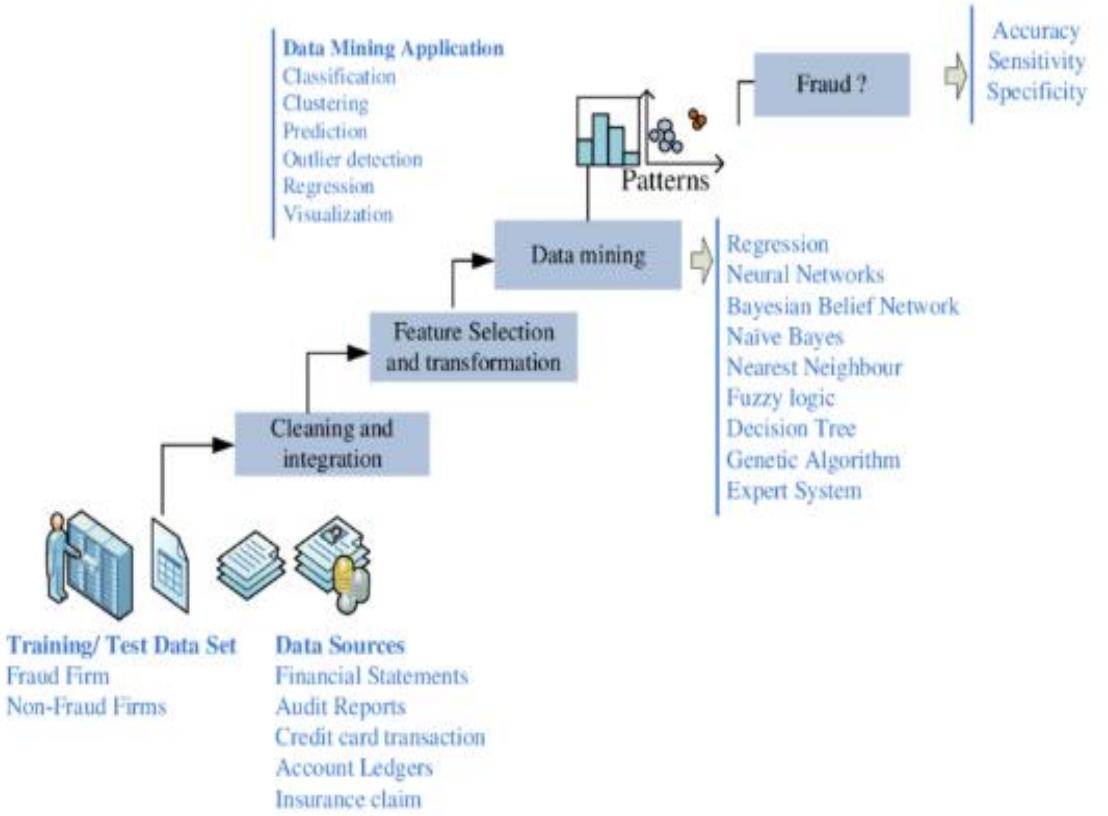

Fig 1: Framework for financial fraud detection

1.1. Background and Significance

Fraud has been described as the systematic abuse of the system of any profit-oriented organization, with the intention of taking undue advantage from the system. The prosperity of a financial institution or an organization mainly relies on financial instruments like deposits, loans, insurances, and live accounts. In a short span of time and with increasing business size, several organizations' customer base can swell to millions, and the large volume of data that is built up due to this customer base needs to be analyzed for proper customer relationship management. Organizations develop policies, decisions, and strategies based on the analyses of vast volumes of data or big data. In fact, financial fraud datasets also fall within this with so many organizations being involved in it. The primary characteristics of big data are popularly known as 5V's: Volume, Velocity, Variety, Veracity, and Value.

With the application of the internet, world computing capabilities have increased and transactions are being made using the internet-based system, which is very fast as well. Here, the transactions change from time-consumed offline based paper transactions to online/faster bit transactions but this change opened the doors to frauds too. Even as criminals are inventing innovative ways to perpetrate fraud, they are happily supported by the use of technology. Fraud detection has become critical in the financial domain. The rapid growth of online transactions has widened the scope of credit card fraud. One of the greatest challenges faced in financial transactions is the identification of fraudulent transactions. A Novel solution methodology is needed for this ATM fraud detection problem, which is done using Machine Learning in Big Data-Based Framework. It has been observed that banks fail to take quick actions on frauds occurring in their ATM networks. Consequently, customers face heavy monetary loss. Fraud detection is the responsibility of Banks in the ATM network. Most of the existing works in ATM fraud detection find fraudulent transactions based on the history of transactions. These works are unable to identify the frauds when channels are different (e.g. the transaction is made from a different ATM). This limitation has led to a research gap in ATM fraud detection problem, which is thus targeted in this study.

In recent years, the rapid growth of the use of electronic payments has opened new perspectives for fraudsters. Criminal activities concerning the fraudulent use of payment cards have grown considerably. Fraud detection methods would require the analysis of very large amounts of data generated by the transactions. Typically, precautions against fraud are applied to the most important entities such as spending transactions for customer credit cards and withdrawal transactions for bank accounts. Traditional fraud detection methods are no longer sufficient. Off-line fraud detection methods estimate the fraudulent score of transactions using only the features embedded in their descriptive variables. However, they take little account of the dynamic fraud environment, which results in the worsening of performance over time as new criminals make efforts to disguise their fraudulent action. In addition, on-line fraud detection methods make decisions through a scoring system provided by an off-line fraud detection method. So, on-line methods may fail to detect novel fraud types as well.

Equ 1: Compliance & Encryption Overhead

$$E_t = \frac{E \cdot D}{B}$$

Let:

- E_t: Time to encrypt + transmit data
- E: Encryption strength factor
- B: Bandwidth
- D: Data volume

2. Literature Review

With the increase in digital transactions, the amount of money lost in fraudulent transactions is also growing. Better machine-learning models are necessary to optimize accuracy given the high number of records to analyze. Credit cards are one of the most common payment methods worldwide. Credit card transactions give customers access to convenient payment all the time, but unfortunately, this convenience also opens doors to fraud. The goal of credit card fraud detection is to determine whether a transaction is fraudulent or genuine based on the information collected related to it. This was originally done by asking the customer about the transaction, a costly process to do with the growing number of transactions each second. Hence, a machine learning model is needed to automatically detect fraudulent transactions at the source.

Different classifiers exhibit different performance levels for credit card fraud detection. The Random Forest Classifier and Extreme Gradient Boost Classifier exhibited the best accuracy, time complexity, precision, recall, f-score, and confusion matrix, confirming looming accuracy in the model. The increasing speed of cashless transactions raises the demand for automatic fraud detection systems. The case study on UK credit cards proves classifiers with high levels of precision, recall, and f-score yields a greater confidence increase when detecting fraud. Even with a low number of transactions and no detailed records of customer transactions, accurate predictions can still be provided. Performance can be improved with more memory.

Automated Fraud Detection aims to find suspicious behavior from historical data and to define those behaviors as rules. Data mining techniques are used with great success in fraud detection, and several classification algorithms for detecting fraudulent transactions have been developed and tried on real datasets. The main challenge of many classification approaches is the dynamic character of frauds and thus, the concept drift that can occur. Hence, model retraining strategies must be considered. In streaming data environments, learning requires new model construction techniques because of the need to process data within a limited time and often under resource constraints. In a description, the challenges posed by debit networks and how the proposed framework SCARFF can aid in tackling these challenges, particularly in streaming environments, are described.

2.1. Research design

Table 1 outlines the research phases and the researchers' intentions at each stage. The phases of sample and dataset gathering begin on 10 Nov 2020 and conclude on 4 Dec 2020. The dataset file is called `creditcard.csv`. The size of the complete dataset is `284807 x 31`, and it is divided into three data sets, of which 5000, 10000, and 50000 datasets have been created for further analysis. All analysis and graphs are drawn from the dataset's CSV file, which has hung in the application for deep learning model training. Raw transactional data is in csv format; exploratory data analysis is performed using Excel to quickly understand the types of data, and any scrapping or missing data analytics is done using Python and proper libraries. Visualizations are created using the seaborn library, pylab library, and matplotlib to analyze

the dataset distribution and prepare for machine learning model selection. The last step of data preparation is taking the complete dataset, splitting it into 70 % for training and the remaining 30 % for validation. This general research design includes a problem definition, a search for relevant literature, the formulation of hypotheses, the choice of a research approach, research strategy, and research methods, followed by the analysis of data and drafting reports. Finally, institutions, individuals, and organisations make their conclusions and recommendations. In these types of writings, a logical and coherent thought flow is the prime focus, so that readers can easily understand the research framework. Research design provides a conceptual framework for data collection, measurement, and analysis. A research design is the arrangement of conditions or directly related to the collection and analysis of data. It shows a general plan of the problem to be studied, detailing how the research will be conducted. These smart facilities enhance the delivery of infrastructure services and improve the quality of everyday life. Big data plays an important part in societies globally.

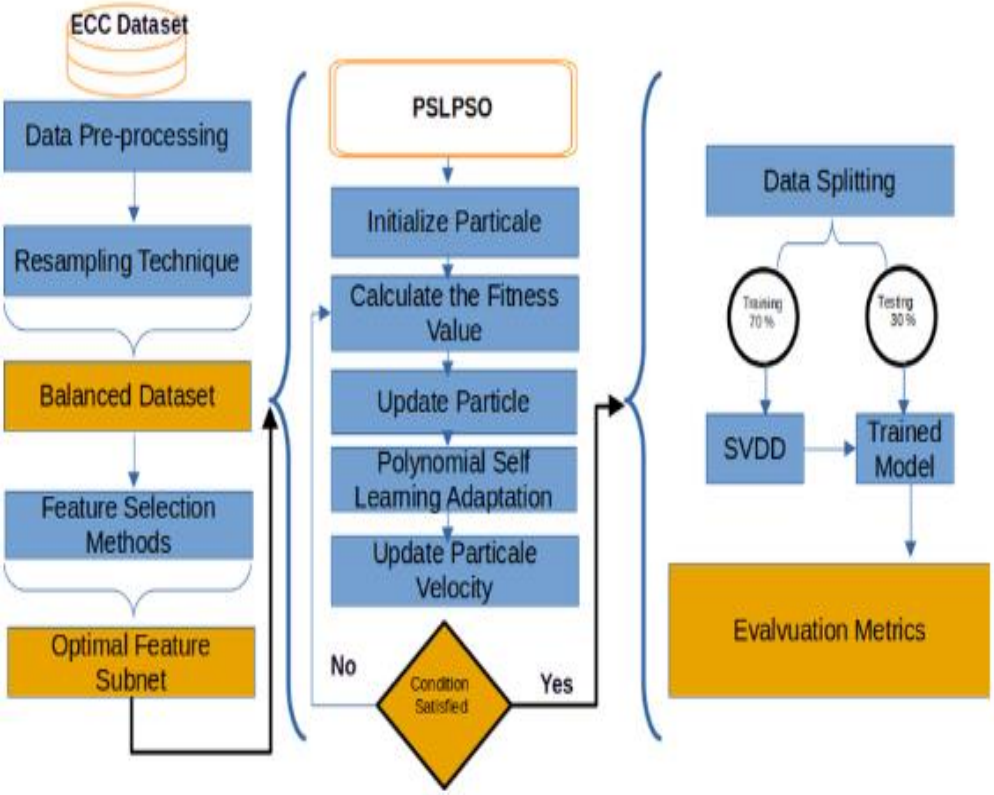

Fig 2: Research design of financial fraud detection

3. Understanding Fraud in Financial Transactions

Fraud in financial transactions can be defined, in a wide sense, as an ill-justified suspension of a transaction made by the customer of a financial institution. The customer can be an individual or a corporation, and the institution can be a bank that manages checking accounts that are charged for any expense, or a payment processor that operates credit cards during electronic payment transactions. Fraud may take multiple degrees of magnitude, from simple, habitual cheating on complex contracts like mortgages or insurances, to more dramatic crimes like bank robberies or butchering of cash carrying trucks. Any financial institution endeavors principally to avoid fraud detection and action policies because so far this enormous cost has not been substantially reduced by technological development. Fraud detection is studied by operations research, statistics, computer science, sociology, and psychology.

In this paper, newly developed technologies like Machine Learning and Big Data Analytics are ameliorated to use large amounts of streaming near real-time data, that does not fit any static model nor can be usefully reduced in detail or number of features. Probabilistic models with maximum likelihood inference upgraded with Gradient Boosted Machines boost decision trees are able to capture this evolving information and build a one-hour market view, updated every minute through robust online learning techniques. A Hidden Markov Model keeps into consideration the agent banks while filtering systems capture the attitude of the fraud/legitimate process.

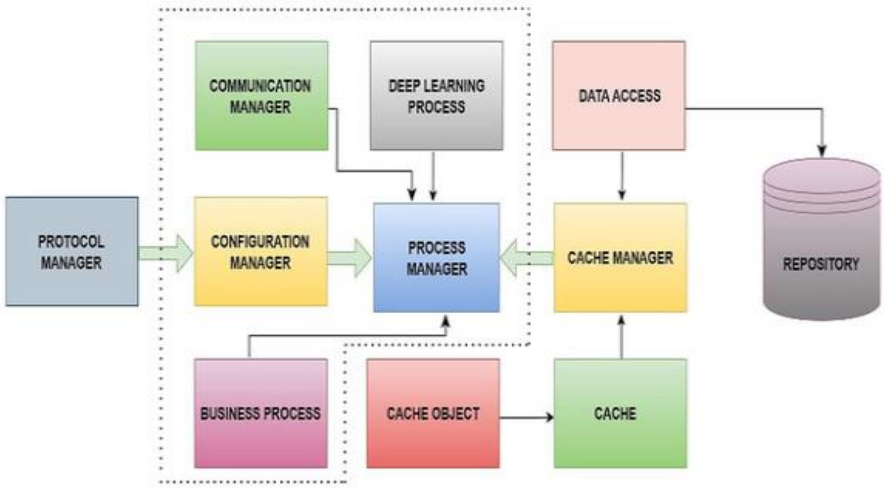

Fig 3: Credit card fraud transaction detection

3.1. Types of Fraud

Although data and knowledge are abundant in the banking sector, the emergence of big data and complex financial transactions also increases the challenge for banks in detecting unplanned behavior in the system. A fraud detection system is essential to protect customers' data and money loss for banks. A financial transaction is flagged as fraud when customers make unexpected transactions, such as by geolocation, amount difference, or time. Financial fraud detection systems present extensive problems.

• Credit card fraud: Credit card fraud is one of the most common fraud types today. In this case, consumers face transaction issues either from online service providers or merchants. For instance, an online service provider may flag a transaction from a foreign country as high risk for fraud and may refuse to deliver a requested product. Merchants face a chargeback request from the payment processor, and the whole amount will return to the cardholder. Cardholders can deny a transaction, which causes a loss on both merchants and sales.

• Insurance fraud: This is another type of fraud scenario in the financial domain. Insurance fraud is insurance claim requests that are noncompliance with rules and regulations or fraud. Some stakeholders may agree to create fictitious incidents to achieve benefits. An insurance fraud detection system aims to improve the level of security and trustworthiness inside the insurance claim process.

• Market manipulation: Merchants send continuous queries for asset price history, while fraudsters take advantage of this processing by creating sudden withdrawal orders to the product. A sophisticated behavior detection is necessary here to protect the merchants when some orders were exactly the same but at different timestamps and returns.

• Money laundering: Money laundering is a process where criminals conceal their illicitly obtained money, also known as dirty money. Criminals use money laundering to remove any trace of activity on illegal revenue and will invest it into legal financial transaction systems, such as purchasing stocks or home mortgages and opening credit cards, while exiting their law-break activities. A money laundering detection system aims to trace back financial transactions to fulfill the requirements of investigation by specialists.

3.2. Impact of Fraud on Financial Institutions

Fraud refers to "the abuse of a profit organization's system without the strict liability consequences of criminal fraud." This term encompasses any fraudulent activity that aims to benefit the perpetrator directly or via an accomplice with the attendant economic loss of the victim. In general, fraud is considered the most common and pernicious computer crime and has significantly impacted financial business valuation. Payment fraud, ATM fraud, credit card fraud, stock exchange fraud, and intrabank fraud constitute a small portion of banking fraud. Finance fraud also includes hacking the informatics infrastructure of institutions conducting buying and selling transactions through the internet. Technological developments, the widespread use of electronic banking channels, the use of automated teller machines (ATMs), and even internet-selling banks, led to an increase in potential fraud acts. Transaction data that is rejected during the control periods is stored as big data and determines potential frauds. Fraud can no longer be sensed without the use of qualified big data analysis. Subsequent fraudulent behaviours follow each fraud act when those fraud acts are not detected. These behaviours can only be noticed with the right analysis of big data containing years' worth of transaction data.

Saved data must be analysed and rules must be determined for that analysis. Fundamental behaviours or acts are grouped as rules. Basic behaviours are achieved using classification using machine learning. However, seeing behaviours that are not covered by rules is difficult or even impossible. Potential fraud candidates react to basic behaviour rules by malicious

behaviours or a fraud maraud in a way to avoid being noticed (sustaining fraud behaviour thanks to learnt rules). This study proposes a two-stage system that consists of the identification of the companies that are subject to potential frauds and the scoring of them based on the probability of being a fraud victim. Companies conduct a larger man on the side, involving a time slot where fraud and legitimate actions occur concurrently. Total money transactions during the pre-fraud time are taken as constant and deviations from it are flagged. Such actuations reveal serious fraud behaviours. This study only evaluates behaviours expressed in money actions, however other actions should not be neglected.

Equ 2: Resilience and Fault Tolerance

$$F = 1 - (\delta)^R \cdot A$$

Let:

- F: Fault tolerance level
- R: Redundancy factor (replication across zones)
- A: Availability zone count
- δ: Failure probability per zone

4. Machine Learning Fundamentals

Machine learning (ML) is a sophisticated technology capable of modeling complex relationships in data, handling non-linear dependencies, and discovering relevant features automatically without explicit human intervention. This area of Artificial Intelligence (AI) achieves good performances with Big Data (BD), which has opened new challenges to tackle using Machine Learning, including data streaming. ML algorithms can be classified into three major paradigms: (1) supervised; (2) unsupervised; and (3) semi-supervised or self-supervised ML algorithms. Supervised ML algorithms learn to predict a response variable Y using a feature space or predictor variables X. In this sense, these algorithms learn a mapping function or score function between the explanatory variables, also called covariates, features, or predictors, and the response variable.

Unsupervised ML algorithms deal only with predictor variables, and there is no response variable to be predicted. These algorithms can cluster the observations sharing similar characteristics and project a high-dimensional feature space into a lower-dimensional representation space that is easier to interpret or visualize. Unsupervised techniques have many applications, including Customer Relationship Management (CRM), which segment customers according to their demographic, purchasing, and behavioral data in order to better target marketing campaigns. Semi-supervised or self-supervised ML learning performs better than supervised ML algorithms based solely on labeled data. Typically, only a small portion of data has been labeled due to both the high cost of data labeling and the exponential growth of data volume. This concerns a variety of application domains, notably in healthcare, web-content classification, and fraud detection, among others.

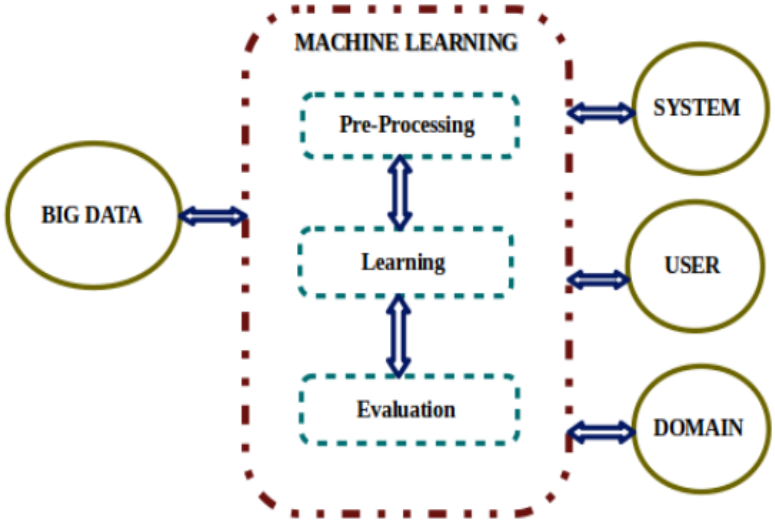

Fig 4: Machine Learning Fundamentals

4.1. Supervised Learning

Fraud is explained as the abuse of a profit organization's system. Fraudsters enter false details or exploit loopholes in an organization's system to obtain unauthorized advantages. They make transactions using the details they generated, often committing fraud worth lakhs and crores of rupees. Countries lose vast amounts worth bigger scales of rupees because of fraudulent activities. Organizations collect huge volumes of data for all the transactions ever made by their clients. Based on the analyses of these vast volumes of data, organizations form policies regarding fraud detection. Digital payment fraud datasets fall within the ownership of the big data paradigm characterized by the Five Vs: Volume, Velocity, Variety, Veracity, and Value. Numerous big data processing tools have been developed to collect, manage, and process huge volumes of data. Out of all these tools, Spark is a unified data analytics engine for big data processing with built-in modules for streaming, machine learning, SQL, and graph processing.

ATM fraud detection has attracted huge interest in research. ATM is the term for a cash point or automated teller machine. The ATM transaction history, which contains information like the ATM ID, the transaction time, the account number, the customer ID, the status (whether the transaction is successful or failed), and the amount of money to deposit or withdraw, can be examined to detect fraudulent transactions. Various possible scenarios exist to identify a fraudulent transaction; a few scenarios include the stolen cards, unusual transaction history of an account, unusual transaction locations, and transaction amounts worth a lakh or more. ATM fraud detection is modeled as the binary classification task wherein the fraudulent transactions are considered as the positive class and the non-fraudulent transactions as the negative class. There is an urgent need to automate the ATM fraud detection process using machine learning techniques. This has not been addressed at length previously except for a few researches. As ATM datasets fall within the big data paradigm, the computer application has been developed to efficiently perform the ATM fraud detection process using scalable machine learning algorithms. The experimental results, based on the real ATM fraud datasets obtained from a bank in India, show that the developed system detects the fraudulent transactions in real-time with the highest classification accuracy and performance in terms of scalability.

4.2. Unsupervised Learning

From the perspective of the function, there are three mainstream ideas of credit card fraud detection: supervised learning, semi-supervised learning, and unsupervised learning. In supervised learning, fraud and legitimate transactions are both needed and the code for legitimate transactions is much larger than the code for fraudulent transactions. This sort of label data imbalance could already cause difficulty in model construction. However, it is duller in real cases as only the legitimate data could be obtained. Lots of fraud cases even cannot be found before. In this regard, there are two sorts of semi-supervised learning methods. On one hand, a small sample of fraudulent transactions could be used to guide the learning of rich legitimate transactions. On the other hand, generative adversarial networks could synthesize artificial fraudulent transactions according to a set of legitimate transactions. Both of the semi-supervised methods are based on the strong prior knowledge which is hard to satisfy in reality. Most notably, fraud patterns change over time while supervised methods rely on past labelled data.

In contrast, unsupervised learning is intended to model the data distribution of one class (i.e., either normal or fraud) and determine whether the test sample belongs to this class or not. Although unsupervised learning is not as attractive as the supervised one, it is suitable for credit card fraud detection as it does not require balanced label data. This sort of unsupervised learning model, even the simplest ones, could be more prominent if the label data is insufficient and the data imbalance is so severe. Another advantage for unsupervised learning is that a fraudulent credit card use could be detected promptly. Once it is put into production, the unsupervised model can be updated in a low latency manner by using online unlabeled data in banks and financial institutes. These transactions could be garnered on a light-green shelf and most of them could be justiciable. They could be sent to the model using a Paillier homomorphic encryption scheme. So the proposed automated system is able to continuously modify the model by using new added transactions in this manner. Such a merit is realized in fact as the self-organizing map model does not require priori information.

4.3. Reinforcement Learning

Prior research has agreeably exhibited the applicability of traditional supervised and unsupervised approaches for detecting online bank fraud transactions using machine and deep learning. Although the catastrophic nature of financial crime has detected fraudulent transactions ahead of time, fraud detection is a continuing battle. The old-fashioned fraud detection algorithms become obsolete with the evolution of large-volume transaction data due to the emergence of evolving fraud patterns and human habits. Therefore, it is necessary to continually update the model with the strategy of automatic, timely, and non-intrusive model upgradation. Reinforcement learning is characterized for adapting the model according to evolving data and for timely learning, implementing, and adjusting the model without rebaking. It deals with the much-needed ingredient for real-world online systems, the learning agent’s ability to self-regulate the learning speed by adapting to changing conditions. The online transaction fraud dataset was used for assessment in conjunction with Commercial Bank of Qatar financial data. The detailed implementation and training of the proposed architecture were

extensively documented to evaluate the performance of the proposed model systematically. A novel approach for credit card fraud transaction detection using a deep reinforcement learning scheme was proposed. With the prevalence of credit card usage, which enables consumers to make purchases and payments in a more timely and simpler manner, a huge number of credit card transactions take place all over the world at different times and places. Fraud transactions tend to be more common due to the large volume of credit card transactions. Meanwhile, the development of advanced algorithms has led to changes in online banking and the increasing emergence of new fraudulent transactions. These fraud patterns need to be detected in real time, which has posed a challenge for detecting credit card fraud transactions.

5. Big Data Analytics in Finance

The banking sector has witnessed several significant changes in its operational infrastructure involving the automated banking solution, where cash and traditional banking halls have been replaced by cash-recycling automated teller machines (ATMs). The advent of advanced ATM technologies has brought in numerous challenges and facilitated attacks on the automated ATM banking sector. Financial institutions and banks are largely dependent on customer satisfaction, which directly depends on customer relationship management. Fraudulent activities within the banking sector result in huge losses for banks and financial institutions, and these losses increase when these fraudulent activities are not identified in time. Automated teller machine (ATM) fraud detection focuses on in-depth analysis to analyze the customer's behavior, transactions, ATM usage history, and transaction pattern to classify ATM usage as legitimate or fraud.
Fraud detection in financial transactions is a well-researched topic among the data mining and machine learning (ML) community. Several credit card companies and banks have invested heavily in fraud detection systems focusing on online automated intelligent solutions to provide real-time alerts for suspicious transactions. However, as financial operations have migrated to online platforms where users perform transactions using their credit card or banking details in an automated electronic fashion, telematic frauds have evolved and put the onus on banks to make proactive investments in fraud detection systems based on streaming processing and machine intelligence computing systems.
Since each transaction may have many features, which is usually above 1000, the curse of dimensionality creates a challenge in model building. Moreover, fraudulent transactions are rare when compared to non fraudulent transactions. This leads to attribute sparsity and class imbalance, which leads to poor performances of the supervised learning model if tried as is. So deep learning-based unsupervised learning is used to automatically extract the better features/attributes which address the dimensional problem better and produce some helpful clusters where closer transactions belong to the same cluster which is helpful in understanding the data distribution and also used for pre-processing before supervised learning is carried out.
These financial fraud datasets fall within the big data paradigm, whose primary characteristics are the five Vs – Volume, Velocity, Variety, Veracity, and Value. Spark has gained popularity among organizations owing to its unified data analytics engine which can handle massive datasets. Spark offers extensions for various programming languages, support for SQL with SparkSQL, developing scalable machine learning algorithms with SparkMLlib, and streaming analytics with Discretized streams (DStreams).

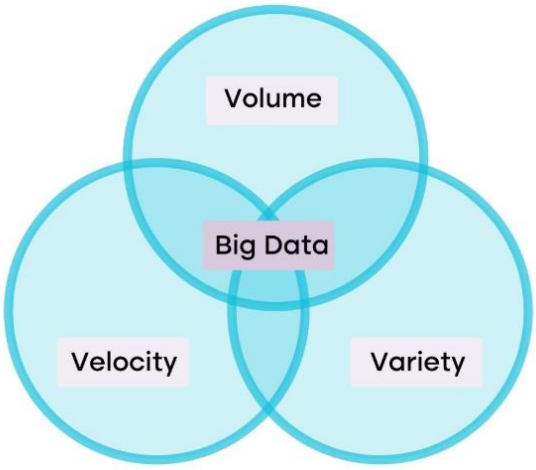

Fig 5: Big Data Analytics for Fraud Detection and Prevention

5.1. Data Sources

Financial institutions increasingly rely on the internet and electronic systems to perform banking and payment services. This rapid transformation offers fraudsters new possibilities for fraudulent actions and has a considerable financial impact

on institutions as well as consumers. Consequently, the need for automatic systems capable of detecting fraud is rapidly increasing. Credit cards are among the most popular means of electronic payments, which poses a real challenge for institutions. In addition to the inherent constraints due to the complexity of the fraud detection task, usual data mining obstacles need to be taken into account as well, notably class imbalance, high dimensionality, concept drift and delayed transaction labeling.

ATM fraud detection is indeed a binary classification task, where the objective is to identify, given some features about a transaction, whether this transaction is fraudulent or not. Among the types of financial frauds, ATM fraud detection has captivated the authors' interests. The ATM transaction history can be examined to detect fraudulent transactions. Scenarios that might assist in identifying fraudulent transactions include loss or theft of the debit/credit card, unusual and unmatched transaction history of the ATM card, transactions initiated from an unusual location, and losses involving bulk amounts of money. Most financial institutions rely on manual intervention to detect ATM fraud scenarios. There is a need to automate this process by employing various machine learning techniques. Organizations base their Fraud policies on the analyses of vast volumes of data, or "Big Data".

5.2. Data Processing Techniques

Data preparation techniques constitute an essential part of developing data-driven applications, because the raw data collected usually require to be transformed before being fed into the machine learning model. This preparation consists of 1) feature extraction from raw data and 2) data normalization or standardization to improve the performance and convergence speed of the model. The feature extraction process is mainly additive as it requires enriching the raw attributes of the transaction data with additional information regarding the user behavior and the transaction context. A pre-processing step is firstly performed on the transactions to collect statically and dynamically aggregated features that reflect how much and how often a card is used during the time. Statistical limits such as mean and standard deviation are further computed on some of these aggregated features to capture the stability of the user behavior. Since the splits are equivalent regarding the input format of the machine learning model, it can be selected on arbitrary experimental splits. General standards were established to tune the critical parameters for different scenarios. For instance, the maximum depth parameter is set to be 13, the number of estimators is determined to be 16, and a learning rate of 0.1 and maximum depth of 5 is determined. Concerning allowed parameters of 8 neighbors for K-Nearest Neighbors, maximum iterations of 360 for Logistic Regression with maximum iteration checks of 20, and maximum depth of 16 with Gini impurity are adopted.

Equ 3: Security Risk Reduction via Cloud Infrastructure

$$R_c = \beta \cdot R_b \cdot g(V)$$

Let:

- R_b: Baseline risk in a traditional on-prem system
- R_c: Residual risk in cloud infrastructure
- β: Cloud security enhancement coefficient, $0 < \beta < 1$

6. Conclusion

Due to the widespread adoption of e-commerce platforms, electronic payments have gained significant traction, leading to the emergence of various types of financial fraud. Automated systems to counter such criminal activities are a necessity for banks, payment service providers, and other institutions in charge of managing electronic payments. Current anti-fraud systems comprise a first layer of supervised machine learning classifiers based on historical information. These classifiers assess the risk of transactions in near real-time, triggering additional security checks on flagged samples, with a delay before the transaction takes place. As a consequence, predicting false positives has high social costs for merchants, since buyers risk dropping their shopping baskets and switching providers.

To address these challenges, this study presented a framework for streaming credit card fraud detection in big data settings. This framework exploits classification forests in a big data context to learn from highly imbalanced data streams and adapt to the arrival of new, previously unseen transaction patterns. Its integration with a popular engine for big data processing allows it to be a complete and open-source implementation of an operational fraud detection system for financial institutions. Research efforts were grouped in three main areas to enhance the state of the art of credit card fraud detection. First, an extensive empirical study of techniques for improving and stabilizing the performance of random forests for imbalanced classification was conducted. Next, a modeling strategy particularly suitable for imbalanced and non-stationary settings was proposed, including a heuristic for dealing with the fact that labels on transactions are delayed. Finally, the framework was built on top of a popular engine for big data processing.

The framework is thoroughly assessed in one of the largest commercially available datasets, including 7 million transactions over a period of 39 days, demonstrating its suitability in detecting credit card fraud in imbalanced and non-stationary environments. In particular, bucket drift detection was shown to detect 50% of the fraud with only 0.0543% false positives, two times better than the standard baseline. It concludes by discussing the limitations of the proposed approach and presents possible directions for future investigations.

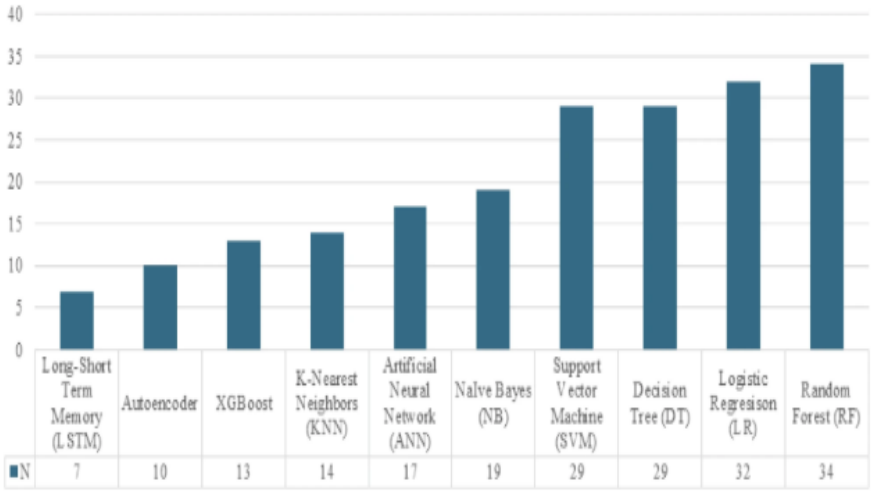

Fig : Financial fraud detection

6.1. Emerging Technologies

The emergence of new technologies such as ERP, data warehouses, big data processing, and analytics are concurring with the need for financial institutions to find new ways to deal with the risk factors threatening their survival and continuity. A review of the fraud detection solutions reported in the literature is done in order to be compared with the solution offered in here, a Data-Driven Framework for REAL-Time Fraud Detection in Financial Transactions. Fraud Detection Solutions Addressing Big Data Scenarios, Machine Learning Methods for Fraud Detection, and Big Data Technologies are part of a quick review of the large solutions that can be applied to big data scenarios. A special interest is devoted to financial frauds, like the most disruptive ones: HFT, insider trading, and credit card frauds. Solutions addressing credit card fraud detection in big data contexts are surveyed, paying a large attention to the features of the proposed methods, namely the adopted strategies to address the problems intrinsic to credit card fraud detection and the performance on real datasets differing on 10 orders of magnitude in size.

The volume of transactions processed by modern financial institutions has been increasing steadily, as there has been a steady adoption of payment cards by the public and the advent of e-commerce. As an economic consequence, financial frauds have been proliferating: it is estimated that in Europe more than €1.5 bn were lost in electronic payment fraud in 2010 which involve payments made with payment cards, such as credit & debit cards or digital wallets, and more than 3000 active clones of phishing websites abusing the brands of the leading online payment service providers. Add in this the European Directive on payment services effective since January 13, 2016 requiring Payment Service Providers to guarantee the technical and organizational security measures to protect infrastructures against frauds, and the fine imposed for non-compliance, and it becomes evident that fraud detection is becoming critical. Detecting frauds in near real-time is a challenge imposed by the large volume of transactions that need to be analyzed.

7. References

[1] JPMorgan Chase. (2024). *AI-Driven Fraud Prevention and Risk Management*. Yallo+1 Leading Tech Trends+1
[2]Mastercard. (2024). *AI-Powered Decision Intelligence Solution.* Yallo
[3] PayPal. (2024). *AI-Driven Fraud Detection Algorithms.* SoftStream.ai
[4] FBD Insurance. (2024). *AI-Based Security Software for Real-Time Threat Analysis*. Financial Times
[5] MPowered Mortgages. (2023). *AI in Lending Processes*.