Fraud Detection Techniques in the Big Data Era

Hanae Abbassi[a], Imane El Alaoui[b] and Youssef Gahi[c]
Ecole National des Sciences Appliquées, Ibn Tofail University, Kenitra, Morocco

Keywords: Fraud, Fraud Detection, ML, DL, data-mining, Big-data, big data analytics.

Abstract: In this age of digital technology, the methodology, complexity, and extent of fraud are drastically increasing, comprising various sectors such as credit card transactions, insurance claims, et cetera, resulting in significant financial losses. To detect fraudulent activities, organizations and financial institutions have implemented different models basing on several techniques, including data mining, machine learning (ML), and deep learning (DL). However, with the advent of big data, the traditional approaches have shown many limits, such as real-time detection and false-positive alerts. These limits have been solved mainly by advanced big data solutions. This paper aims to provide a state of the art of fraud detection techniques in a meaningful data context. Then, we review the traditional methods and identifying their limits by taking advantage of Big Data analytics patterns.

1 INTRODUCTION

Although technology has improved consumer comfort in digital transactions as online payment, transferring money and et cetera, it has also opened new forms of fraud such as bank account takeover fraud, internet fraud, and mail fraud. Financial fraud figures indicate that every year millions of dollars (Raghavan & Gayar, 2019) of penalties for the institution and public entities are triggered by credit card fraud, insurance fraud, and other fraudulent acts belonging to various fields. It is therefore essential to detect fraudulent activities as quickly as possible to mitigate risk and losses. Several approaches allow to detect and prevent frauds in various areas. In the past, the most common method of fraud detection was based on the "rules-based" system. (Allan & Zhan, 2010). However, this technique is severely limited. Since it is built on known models, it can only detect known fraudulent patterns. Therefore, rules-based only partly mitigated the issue.

Later, other techniques have been developed to enhance fraud detection, such as Outlier Detection (OD) techniques and Machine Learning. In OD, the used methods have their way of handling the issue by suggesting that most instances in the dataset are regular and checking for outliers. On the other side, in ML techniques, fraud detection is typically treated as a supervised classification problem, where observations are classified as "fraud" or "non-fraud" based on those observations' characteristics. All these techniques have shown good performances in detecting frauds in various domains. However, with the enormous growth of data, these techniques became limited because of their inability to build accurate models. Therefore, there is an excellent chance of triggering false alarms. Handling massive data, known as big data, need special requirements (such as collection, storage, analysis, and exploitation) due to their following characteristics, called big data 7Vs: (Ana-Ramona Bologa et al., 2010) (El Alaoui & Gahi, 2019) (El Alaoui et al., 2019).

- Volume: refers to the vast amount of data that is generated and processed.
- Variety: Refers to the various data formats, which might be structured, semi-structured, or unstructured.
- Velocity: Refers to the processing speed of data.

[a] https://orcid.org/0000-0002-6451-3441
[b] https://orcid.org/0000-0003-4428-0000
[c] https://orcid.org/0000-0001-8010-9206

Abbassi, H., El Alaoui, I. and Gahi, Y.
Fraud Detection Techniques in the Big Data Era.
DOI: 10.5220/0010730300003101
In *Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning (BML 2021)*, pages 161-170
ISBN: 978-989-758-559-3

- Veracity: refers to the data's credibility and the degree to which it can be trusted to make decisions.
- Variability: refers to data that is continually changing.
- Value: refers to the ability to transform raw data into meaningful information.
- Visualization: refers to the ability to visualize data in readable graphical charts or records.

Several solutions, called big data analytics, which occur as an excellent analysis, have been developed. Big data analytics are considered a cutting-edge way for fraud detection in a significant data context. They provide the ability to capture, store, analyze, reliably visualize voluminous and heterogenic data. They help develop a predictive model that can trigger an alarm as soon as an entry point for fraudulent activity is detected.

For this aim, several researchers focus on using big data analytics to detect fraud and develop preventive measures. They have proposed several detection models in various domains such as healthcare, finance, and networks. In this contribution, we discuss these exciting models that help to protect big data environments against fraud.

The remainder of this article is structured as follows: Section 2 gives an overview of some fraud cases. Section 3 reviews studies that have proposed traditional methods for fraud detection and also shows their limitations. Section 4 examines the tasks that have tackled big data analytics within the fraud detection context. Section 5 discusses the benefits and the challenges faced by the big data analytics techniques for fraud detection. Finally, in section 6, we conclude our review.

2 FRAUD CASES

When information becomes ever more openly shared and consumers demand almost immediate access to goods and services, the war against crime is relentless, posing new challenges. In fact, From the past until now, Fraudsters continue to evolve their strategies and techniques. Fraudulent activities are not limited to a specific field; they relate to all domains, including insurance, healthcare, and networks, as shown in Figure 1.

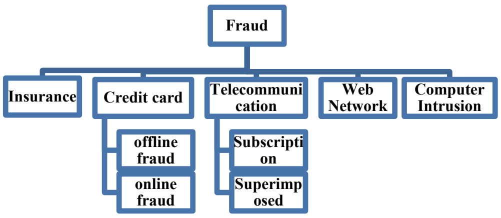

Figure 1: Fraud cases.

Insurance Fraud: Insurance fraud is when someone tries to take advantage of an insurance policy for financial gain. It refers to the improper use of insurance policies or applications to profit or obtain illegal benefits. Insurance fraud includes Healthcare, Automobile, Home, and life insurance.

Credit Card Fraud: Credit card (CC) fraud occurs when someone uses a credit card or credit account to purchase without the person from its holder. This activity can occur in various ways; offline copy, which refers to using a physical card stolen, and online fraud committed through the web or phone.

Telecommunication Fraud: refers to the misuse of telecommunications services to illegally acquire money from a communications operator or its customers. It can be classified into two types: i) Subscription fraud happens when a fraudster uses their own or stolen identity to get services with no intention of paying. In such instances as cellular cloning, the irregular use is superimposed on the legitimate customers' regular usage. ii) In Superimposed fraud, a legal account is taken over by fraudsters. (Yufeng Kou et al., 2004).

Computer Intrusion: An intrusion is any action that threatens to violate the credibility, security, or availability of a resource (file system, user account, et cetera.). Computer intrusions are divided into two types:

i) Misuse intrusions, which are attacks against known weak spots in a system.

Among them, we find denial of service attacks, malicious use, et cetera. ii) Anomaly intrusions are described as anomalies from a regular system that are observed. (Yufeng Kou et al., 2004).

Web Network Fraud: Web fraud is described as using network resources or applications with Internet access to defraud or exploit victims. It includes

i) The web advertising network fraud, a go-between for Internet publishers/advertisers.

ii) Internet auction fraud involving the misrepresentation or non-delivery of a commodity available for sale.

These fraudulent activities have a negative impact not only on financial losses but also on institutions' reputations. For these reasons, researchers have shown a great interest in fraud detection by proposing different models basing on machine learning, deep learning, and data mining. In the next section, we review attractive fraud, detection models.

3 TRADITIONAL FRAUD DETECTION METHODS

A variety of fraud detection methods have been suggested in the literature. We have noticed that ML, data mining, and DL techniques are the leading used models by reviewing traditional fraud methods. We present these models in the following.

Machine Learning: ML is a field of artificial intelligence (AI) that allows machines to learn and evolve independently without being programmed explicitly. We present here two forms of ML algorithms which are supervised and unsupervised.

- Supervised learning: refers to a group of structures and algorithms that use sets of labeled data with known outcomes to create a predictive model. The model is learned through a series of trials with the learning algorithm. Several studies have focused on building supervised learning algorithms for fraud detection. In fact, (Adepoju, O. et al., 2019) have used supervised ML algorithms, namely Support Vector Machine (SVM), Naive Bayes (NB), Logistic Regression (LR), and K-Nearest Neighbor (KNN), to detect fraud in a CC context. The comparative results show that logistic regression outperforms other algorithms in terms of accuracy. Also, (Dhankhad et al., 2018) have compared eight supervised algorithms, namely SVM, NB, LR and KNN, Random Forest (RF), stacking classifier, decision tree (DT), and Gradient Boosted Tree (GBT), to evaluate the most accurate model for a detect fraudulent CC transaction. The outcomes show that the stacking classifier gives the best accuracy.
- Unsupervised learning investigates how systems may deduce a function from unlabeled data to explain a hidden framework. Unsupervised algorithms are used when the training data is not classified or labeled. Large bodies of literature have tackled unsupervised learning-based fraud detection models. Such as (Vaishali, 2014) has used an unsupervised algorithm named optimized K-Means clustering to detect fraud in credit card transactions. K-means showed a good result. (Subudhi & Panigrahi, 2020), have used an unsupervised algorithm named optimized Fuzzy C-Means clustering for automobile insurance fraud detection. The C-Means achieved 81.87% accuracy with the balanced dataset.

Data Mining: is a method for identifying potentially valuable trends in large data sets. It is multidisciplinary expertise that integrates ML, statistics, and AI to extract data. These techniques are used to detect fraud. Data mining is a significant advantage of building a new category of models to spot further attacks before they are seen (Yufeng Kou et al., 2004). We present here some standard methods used in fraud detection:

- Imbalanced Classification: As a pre-processing stage, data balancing techniques rebalance skewed data sets or eliminate noise. Several studies employ data-level balancing techniques. In (Benchaji, I. et al., 2018), a new approach for generating data for the minority class of an unbalanced data collection is proposed as an oversampling technique to boost fraud detection in e-banking. Also, (Dornadula & Geetha, 2019) have used SMOTE (Synthetic Minority Over-Sampling Technique) to balance the credit card dataset. The outcomes show that the classifiers (SVM, LR) perform better using SMOTE method.
- Model Combination: This approach consists of building a new model by combining multiple algorithms to increase efficiency and accuracy rates. Researchers have combined several models. (Raghavendra Patidar & Lokesh Sharma, 2011) have suggested the combination of neural networks and genetic algorithms to identify fraudulent CC transactions successfully. Also, (Song 2020) has proposed a hybrid algorithm based on both lightguns and boost to detect fraud in electronic banks. The proposed model gave an excellent Auc-Roc score and precision. In the same vein, (Tiwari et al., 2021) have proposed a hybrid algorithm based on DT, NN, and K-Nearest Neighbor algorithm to detect fraudulent CC transactions. The proposed model gives a good result.
- Outlier detection: It includes statistical, and data mining approaches. It uses behavioral profiling techniques to model each individual's behavior and track any deviations from the standard. These are essentially unsupervised learning techniques. (Richard J. Bolton & David Hand, 2001) have used Peer group analysis and Breakpoint analysis methods for behavioral fraud detection, analyzing longitudinal data.

Deep Learning is a class of ML that uses many hidden layers. It generally achieves better results than other ML algorithms. Several researchers have adopted DL to deal with fraudulent acts. As an example, (Raghavan & Gayar, 2019) have compared different DL algorithms like Convolutional Neural Networks (CNN), autoencoders, restricted Boltzmann machine (RBM), and deep belief networks (DBN) with ML algorithms. The result reveals that CNN outperforms other DL algorithms. Also, (H. Gomi et al., 2016) have suggested a model based on trends in network access logs, using the Recurrent Neural Networks (RNN) model to detect fraudulent behaviors. The outcomes demonstrate that RNN outperforms ML methods SVMs when it comes to learning the habits of a non-genuine person. For network anomaly detection systems, (Z. Chen et al., 2018) have proposed a model based on Autoencoder (AE) and Convolutional Autoencoder (CAE). The result shows that CAE outperforms the AE.

All the mentioned above studies have shown promising results for small datasets. In the table below, we provide a comparison between these kinds of literature

Table 1: comparative literature.

Detection Method	Paper	Technique	Performance
Supervised learning	(Adepoju, O. et al., 2019)	logistic regression, k- nearest neighbor, and SVM	LR archived the best accuracy 99.07%,
	(Dhankhad et al., 2018)	SVM, Naive Bayes, LR and K-Nearest Neighbour, RF, stacking classifier, decision tree, and Gradient Boosted Tree	High accuracy: stacking classifier 95.27% High rank: Support Vector Machine 53.6%
unsupervised	(Vaishali, 2014)	K-Means	Good results
	(Subudhi & Panigrahi, 2020)	C-Means	81.87% accuracy
Classification Emblanced	(Benchaji, I. et al., 2018)	Oversampling	Performs well
	(Dornadula & Geetha, 2019)	SMOTE	Give good results
Model Combination	(Raghavendra Patidar & Lokesh Sharma, 2011)	neural networks with genetic algorithms	Good results
	(Song, 2020)	lightgun and boost	98.5% accuracy
	(Tiwari et al., 2021)	Trees, Neural Network, and K-Nearest Neighbor	good results
Outlier Detection	(Richard J. Bolton & David Hand, 2001)	Peer group analysis Breakpoint analysis	Peer group analysis outperforms Breakpoint
Deep Learning	(Raghavan & Gayar, 2019)	AE, CNN, RBM, DBN	CNN outperforms other models
	(H. Gomi et al., 2016)	RNN, ML models	RNN outperforms ML algorithms
	(Z. Chen et al., 2018)	AE, CAE	CAE outperforms AE

Although these methods look promising and give good results, they are unsuccessful in a complex environment, massive data. It is difficult to apply traditional fraud detection algorithms on vast amounts of data within a reasonable time. Furthermore, the performance of these conventional models could drastically decrease by giving false alarms (Sathyapriya & Thiagarasu, 2015). It is also important to mention that the cost of upgrading and maintaining these methods is high (Allan & Zhan, 2010). Researchers have proposed new models to detect fraud in a significant data context to deal with these limits. We review these models in the following section.

4 BIG DATA ANALYTICS FOR FRAUD DETECTION

Big Data analytics is the de facto method for solving various modeling and decision-making issues (Faroukhi et al., 2021). This is due to their capability to process many data and generate information in real-time, which ultimately reduces costs by ensuring high precision (Melo-Acosta et al., 2017) (Faroukhi et al., 2020). In this section, we review fraud detection studies using Big Data analytics in different fields.

Credit Card: Credit card companies have started to use big data technologies to identify fraudulent transactions in real-time, using a range of big data analytic tools such as Apache Hadoop, MapReduce, Apache Spark, and APACHE FLINK. In (Sathyapriya & Thiagarasu, 2015), researchers have compared these tools in credit card fraud detection context, basing on factors such as efficiency, scalability, fault tolerance, processing speed, and latency. They have found that Apache Spark outperforms other techniques. Combining big data with ML techniques can also give better results. In fact, (Melo-Acosta et al., 2017) have proposed a fraud detection system (FDS) for CC transactions based on the Spark RF model and Balanced Random Forest (BRF). The technology was developed to overcome the three difficulties related to fraud detection data sets: a heavyclass imbalance, the incorporation of unlabeled and labeled samples, and handling many transactions. The result reveals that the proposed model successfully solved all the problems. As well, (Armel & Zaidouni, 2019) have used Apache Spark's (MLlib) to detect credit card fraud in banking transaction data, using four supervised algorithms as Simple Anomaly detection, DT, RF, and Naïve Bayes. The random Forest algorithm performed best in terms of accuracy and running time, while a simple anomaly algorithm performed worst. In the same context, (Hormozi et al., 2013) have run the Artificial Immune System's algorithm to credit card FDS on Hadoop and MapReduce to cope with the AIS long training time. For this, they have parallelized the negative selection algorithm on the cloud with Apache Hadoop and MapReduce. The results reveal that the algorithm's training time is greatly reduced compared to the basic algorithm. Other studies have used hybrid methods to improve the performance of FDS. (Kamaruddin & Ravi, 2016) have used a model (POSAANN) over a credit card fraud dataset to obtain the one-class classification (OCC) solution within a SAPRK cluster, including the particle swarm optimization (PSO) and the auto-associative neural network (AANN). They have also parallelized AANN in a hybrid architecture. The proposed model achieves good performance.

Healthcare: Big data from outlets such as Medicare is being used to protect fraud and patient treatment. Healthcare fraud significantly impacts insurance schemes' ability to deliver reliable and affordable care, such as Medicare (R. Bauder & Khoshgoftaar, 2018). (Georgakopoulos et al., 2020) have presented the HNOPH (Hellenic National Organization for the Provision of HeHealthService's methodological approach to identify medical fraud in claims using the Local Correlation Integral algorithm. This study aims to detect any outliers (fraudulent cases) on the dataset. (Herland, M. et al., 2018) have examined four Medicare datasets: Part B, Part D, DMEPOS, and Combined, to assess the fraud detection capabilities of Medicare datasets individually and in combination. Each dataset was trained and evaluated using three different learners: RF, Boosted GradientTrees, and LR. The examination is run and validated using Spark on top of a Hadoop YARN cluster. The combined dataset had the best overall fraud detection results using LR. Using the same dataset, (R. A. Bauder et al., 2018), have provided experimental results using six data sampling methods (RUS, ROS, SMOTE, ADASYN, SMOTEb1, SMOTEb2). They have also used three machine learning models (RF, LR, and GBT) with Apache Spark to compare the performance of Medicare fraud detection through data sampling methods. The outcomes show that RUS performs well across all learners.

Financial Statements: Big Data analytics are commonly used in the financial sector. They can help banks to understand their customers' actions better and to detect fraud efficiently. (Y.-J. Chen & Wu, 2017) have proposed a method for fraud detection in financial statements based on big data. They have used QGA-SVM as a clustering model to enter established datasets to improve fraud detection accuracy. The proposed method allows minimizing losses and investment risks. It is essential to mention that combining big data tools with DL and ML gives good results in detecting fraud. In fact, (Purushe, P., & Woo, J, 2020) have used Amazon AWS with Spark ML and DL to detect fraudulent transactions on a financial transaction dataset. They have compared three ML algorithms: LR, DT, and random forest with feed-forward (FF) as a DL model. The results reveal that random forest had the best accuracy, 95.9%, and FF had the best recall, with minor false negatives. (Zhou et al., 2021) have suggested an intelligent and distributed Big Data solution to detect Internet financial fraud. They have implemented the graph embedding algorithm Node2Vec on Spark GraphX and Hadoop to learn and represent each vertex's topological characteristics into a dense low-dimensional vector. The proposed method aims to predict the fraudulent samples of the dataset. The experimental outcomes show that the proposed method increases the efficiency of Internet financial fraud detections by improving accuracy and recall.

Network Intrusion: Big Data Analytics can anticipate and track intrusions and threats in real-time. Big Data technologies such as the Hadoop ecosystem and flow computing can store and interpret massive datasets quickly. They also transform network security detection by extracting large-scale data for various internal and external sources, such as vulnerability databases. (Terzi et al., 2017) have presented an unsupervised anomaly detection approach based on Apache Spark cluster in Azure HDInsight platform. The proposed approach system accuracy of 96%. Also (Kato & Klyuev, 2017) have presented an anomaly-based intrusion detection method using Apache Hadoop and Spark. They have used a real IDS dataset provided by the UNBISCE (University of New Brunswick's Information Security Centre of Excellence) in the framework. They dealt directly with the packet capture files (pcap) and evaluated 90.9 GB of data on Hadoop clusters. They also used principal component analysis (PCA) to reduce feature dimension and the Gaussian mixture model (GMM) to divide network behaviors into regular and attack classes. The results showed that the method could create intelligent IDS, with a detection accuracy of 86.2 % and a false positive rate of 13%. Other studies have adopted DL for intrusion detection. In fact, (Haggag et al., 2020) have presented an intrusion detection system (IDS), named Deep Learning Spark Intrusion Detection System (DLS-IDS), to address the issue of dataset class imbalance and to improve accuracy and speed. IDS is implemented on Apache Spark using three DL algorithms: Multilayer Perceptron (MLP), Recurrent Neural Network (RNN), and Long-Short Term Memory (LSTM).

Furthermore, they have compared these models to ML algorithms. The results have shown that the proposed model performs well in terms of accuracy and time. The combination of LSTM and SMOTE increases the detection accuracy by 83.57%.

All the presented above studies are exciting and deal with fraud detection in a significant data context. They provide reliable and promising predictive models to prevent fraud. In the following table, we provide a comparison between all these models.

Table 2: Compared Fraud researches.

Area	Paper	Used data	Techniques & technologies	Performance
Credit Card	(Sathya priya & Thiagarasu, 2015)	Credit card transactions	Apache Hadoop, MapReduce, Apache Spark,	Apache Spark performs better (speed, low latency, high fault tolerance, fast performance, and high scalability).
	(Melo-Acosta et al., 2017)	Colombian payment gateway company	Spark and Hadoop, Balanced Random Forest	BRF based on the Spark RF attains an improvement of around 24% in terms of geometric compared to a standard RF.
	(Armel & Zaidouni, 2019)	Large Brazilian bank dataset	Apache Hadoop, MapReduce Artificial Immune System, the adverse selection algorithm	The training time was significantly reduced.
	(Hormozi et al., 2013)	Banking transaction data: Price and Distance	MLlib of Apache Spark Simple Anomaly detection algorithm, Decision Tree, Random Forest, and Naïve Bayes	Accuracy: Random Forest: 98,18%, Decision tree: 96,96%, Naïve Bayes: 91,24% and simple anomaly algorithm: 77,04%
	(Kamaruddin & Ravi, 2016)	Credit card fraud dataset	SAPRK, hybrid model PSOAANN ((PSO) and (AANN))	The model achieves an average of 89% accurate classification of the CC fraud transactions.
	(Dai, Y. et al., 2016)	Synthetic data	hybrid framework: Hadoop, Spark, Storm, Hbase	Achieves a scalable, fault-tolerant, and high-performance system.
	(Patil et al., 2018)	Credit card fraud dataset	LR, DT, Random Forest Decision Tree	Accuracy of: LR: 70%, DT: 72% and RFDT: 76%.
Healthcare	(R. Bauder & Khoshgoftaar, 2018)	Physician and Other Supplier Data calendar 2012 to 2015	Random Forest, random undersampling (RUS)	The best class distribution is 90:10 with an AUC of 87.302%

	(Georgakopoulos et al., 2020)	Prescription data	Local Correlation Integral algorithm (LOCI), Angle-based outlier detection (ABOD) algorithm	The model detects 7 out of 879 items as fraudulent
	(Herland, M. et al., 2018)	CMS datasets t (Part B, Part D, and DMEPOS)	Spark on top of a Hadoop YARN cluster, Random Forest, Gradient Boosted Trees, and Logistic Regression	The Combined dataset had the best overall fraud detection performance with an AUC of 0.816 using LR
	(R. A. Bauder et al., 2018)	Medicare datasets, from 2013 to 2015 by (CMS)	Data sampling (RUS, ROS, SMOTE, ADASYN, SMOTEb1, SMOTEb2) Apache Spark, Mllib LR, GBT, RF	Full dataset (with no sampling) performed well for all (GBT, LR, RF) respectively, with an accuracy of 0.79047, 0.81554, 0.79383
	(Ana-Ramona Bologa et al., 2010)	The National House for Health Insurance	Business rules, anomaly detection, text mining, database searches, social network analysis, and Predictive modeling	Applying BDA techniques can lead to rapid detection of abnormal claims
Financial statements	(Y.-J. Chen & Wu, 2017)	financial statements of business groups i	Decision tree, logistic, neural network, KNN, GA-SVM, POS-SVM, and QGA-SVM.	QGA-SVM has the higher (90.5%)
	(Purushe, P., & Woo, J, 2020)	financial transaction dataset	Amazon AWS, Spark ML, and DL. ML algorithms: logistic regression, decision tree, and random forest, DL: FF	Random forest had the best accuracy of 95.9%, and FF had the best recall, with minor false negatives.
	(Zhou et al., 2021)	Internet financial service provider in China	Spark GraphX, Hadoop, Node2Vec, DeepWalk, SVM, SMOTE	The proposed solution will increase the reliability of Internet financial fraud detections.
Network Intrusion	(Terzi et al., 2017)	Public NetFlow data	unsupervised anomaly detection approach on Apache Spark cluster in Azure HDInsight	96% accuracy rate was obtained
	(Kato & Klyuev, 2017)	Network packet dataset	Apache Hadoop and Spark, Hive SQL, GMM, OCSVM, K-Means, LR, SVM, RF, GB tree, and Naive Bayes	The system achieves a detection accuracy of 86.2% and 13% of false-positive rate
	(Haggag et al., 2020)	NSL-KDD	Apache Spark, Multilayer Perceptron, RNN, and LSTM	The combination of LSTM and SMOTE increases detection accuracy to 83.57%.

5 DISCUSSION

The emergence of Big Data has created a new area of research known as data-driven fraud detection. We have tackled studies that focus on using Big Data analytics (BDA) to examine various forms of criminal activities in multiple fields, such as healthcare, network intrusion, and credit card fraud. Basing on these studies, we discuss some significant data advantages and challenges in the context of fraud detection.

Big data technologies provide several advantages in detecting and preventing fraudulent activities. We present here some of the most important benefits:

- Extensive data processing: Big data tools, as Apache-Hadoop, enable the collection, process, and analyze a variety of data from various sources such as finance, messages, and social media similarly. It also allows storing several data types (structured, semi-structured, and unstructured data) (Jha et al., 2020).
- Accurate time detection: big data analytics allow to detection of malicious activities in real-time by using several techniques and technologies. Among these methods, the Deep analytics (DA) approaches present advancement from discrete analysis of structured data to connected analysis of unstructured data in real-time. The DA systems can spot trends and warn of potentially suspicious activity in real-time by observing each customer's behavioral trends.

Furthermore, real-time processing allows gathering data from a variety of sources and quickly create contextual baselines. This allows reducing the number of false alarms (Bharath Krishnappa, 2015) (Singla & Jangir, 2020)

- Fraud Prediction: BDA algorithms are one of the most effective solutions to predict security concerns. These predictive algorithms help to enhance the reliability of predictions and analysis models. They are generally based on ML methods and allow to expect and avoid malicious activity by evaluating conventional security incidents, results, and user data. (Fatima-Zahra Benjelloun & Ayoub Ait Lahcen, 2015) (Singla & Jangir, 2020).
- Reduce sampling: reducing samples is a significant advantage of big data technology. Data analytics sampling methods can examine a subset of all data to discover meaningful information in a broader data set and giving the best results.

Although big data technology had solved conventional techniques limits, it faces some challenges:

- Skewed distribution (Imbalance class): the classification algorithms used within a Big Data framework are more vulnerable to this issue. Big Data tools as Hadoop generally split data into chunks. Hence, the tiny amount of data in the samples shrinks dramatically. It is important to note that using extremely skewed big data does not yield effective fraud detection results (Georgakopoulos et al., 2020; Makki et al., 2017).
- Performance: In big data, it is challenging to carry out input validation or data filters in the incoming data due to the massive number of terabytes of continuous data flow. Therefore this issue has a significant impact on performance. (Bhandari et al., 2016)
- Data privacy: Analysts may correlate several separate data sets from different organizations to expose confidential or critical data even with anonymization methods. Such correlation could allow the identification of persons or the discovery of sensitive information (Yadav et al., 2019)(Bhandari et al., 2016)(Jensen, 2013) (Gahi et al., 2016).

6 CONCLUSION

Cases of fraud have caused substantial damage and losses to financial statements, the government, and individuals. For this reason, several researchers have proposed several fraud detections and prevention models based on various techniques. In this paper, we have presented a state of the art of fraud detection metamethods. We have introduced traditional fraud detection methods such as data mining, ML, and DL, identifying their limitations. Then, we have given big data analytics techniques that allow us to cope with these limitations. We have also presented the advantages and challenges of big data technologies.

In future works, we aim to address these challenges such as imbalance class and data privacy.

REFERENCES

Adepoju, O., Wosowei, J, lawte, S, & Jaiman, H. (2019). Comparative Evaluation of Credit Card Fraud Detection Using Machine Learning Techniques. 2019 Global Conference for Advancement in Technology (GCAT).

Allan, T., & Zhan, J. (2010). Towards Fraud Detection Methodologies. 2010 5th International Conference on Future Information Technology, 1–6. https://doi.org/10.1109/FUTURETECH.2010.5482631

Ana-Ramona Bologa, Razvan Bologa, & Alexandra Maria Ioana Florea. (2010). Big Data and Specific Analysis Methods for Insurance Fraud Detection. Database Systems Journal Vol. I, No. 1/2010. https://www.researchgate.net/publication/292980386_Big_Data_and_Specific_Analysis_Methods_for_Insurance_Fraud_Detection

Armel, A., & Zaidouni, D. (2019). Fraud Detection Using Apache Spark. 2019 5th International Conference on Optimization and Applications (ICOA), 1–6. https://doi.org/10.1109/ICOA.2019.8727610

Bauder, R. A., Khoshgoftaar, T. M., & Hasanin, T. (2018). Data Sampling Approaches with Severely Imbalanced Big Data for Medicare Fraud Detection. 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI), 137–142. https://doi.org/10.1109/ICTAI.2018.00030

Bauder, R., & Khoshgoftaar, T. (2018). Medicare Fraud Detection Using Random Forest with Class Imbalanced Big Data. 2018 IEEE International Conference on Information Reuse and Integration (IRI), 80–87. https://doi.org/10.1109/IRI.2018.00019

Benchaji, I., Douzi, S, & ElOuahidi, B. (2018). Using Genetic Algorithm to Improve Classification of Imbalanced Datasets for Credit Card Fraud Detection. Using Genetic Algorithm to Improve Classification of Imbalanced Datasets for Credit Card Fraud Detection.

2018 2nd Cyber Security in Networking Conference (CSNet).

Bhandari, R., Hans, V., & Ahuja, N. J. (2016). Big Data Security – Challenges and Recommendations. International Journal of Computer Sciences and Engineering, 4, 7.

Bharath Krishnappa. (2015). BIG DATA ANALYTICS FOR CYBERSECURITY. EMC, India Center of Excellence. https://education.dellemc.com/content/dam/dell-emc/documents/en-us/2015KS_Krishnappa-Big_Data_Analytics_for_Cyber_Security.pdf

Chen, Y.-J., & Wu, C.-H. (2017). On Big Data-Based Fraud Detection Method for Financial Statements of Business Groups. 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), 986–987. https://doi.org/10.1109/IIAI-AAI.2017.13

Chen, Z., Yeo, C. K., Lee, B. S., & Lau, C. T. (2018). Autoencoder-based network anomaly detection. 2018 Wireless Telecommunications Symposium (WTS), 1–5. https://doi.org/10.1109/WTS.2018.8363930

Dai, Y., Yan, J., Tang, X., Zhao, H., & Guo, M. (2016). Online Credit Card Fraud Detection: A Hybrid Framework with Big Data Technologies. IEEE Trustcom/BigDataSE/ISPA.

Dhankhad, S., Mohammed, E., & Far, B. (2018). Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study. 2018 IEEE International Conference on Information Reuse and Integration (IRI), 122–125. https://doi.org/10.1109/IRI.2018.00025

Dornadula, V. N., & Geetha, S. (2019). Credit Card Fraud Detection using Machine Learning Algorithms. Procedia Computer Science, 165, 631–641. https://doi.org/10.1016/j.procs.2020.01.057

El Alaoui, I., & Gahi, Y. (2019). The Impact of Big Data Quality on Sentiment Analysis Approaches. Procedia Computer Science, 160, 803–810. https://doi.org/10.1016/j.procs.2019.11.007

El Alaoui, I., Gahi, Y., & Messoussi, R. (2019). Big Data Quality Metrics for Sentiment Analysis Approaches. Proceedings of the 2019 International Conference on Big Data Engineering, 36–43. https://doi.org/10.1145/3341620.3341629

Faroukhi, A. Z., El Alaoui, I., Gahi, Y., & Amine, A. (2020). Big data monetization throughout Big Data Value Chain: A comprehensive review. Journal of Big Data, 7(1), 3. https://doi.org/10.1186/s40537-019-0281-5

Faroukhi, A. Z., El Alaoui, I., Gahi, Y., & Amine, A. (2021). A Novel Approach for Big Data Monetization as a Service. In F. Saeed, T. Al-Hadhrami, F. Mohammed, & E. Mohammed (Eds.), Advances on Smart and Soft Computing (pp. 153–165). Springer. https://doi.org/10.1007/978-981-15-6048-4_14

Fatima-Zahra Benjelloun & Ayoub Ait Lahcen. (2015). Big Data Security: Challenges, Recommendations, and Solutions. Handbook of Research on Security Considerations in Cloud Computing. https://www.researchgate.net/publication/278962714_Big_Data_Security_Challenges_Recommendations_and_Solutions

Gahi, Y., Guennoun, M., & Mouftah, H. T. (2016). Big Data Analytics: Security and privacy challenges. 2016 IEEE Symposium on Computers and Communication (ISCC), 952–957. https://doi.org/10.1109/ISCC.2016.7543859

Georgakopoulos, S. V., Gallos, P., & Plagianakos, V. P. (2020). Using Big Data Analytics to Detect Fraud in Healthcare Provision. 2020 IEEE 5th Middle East and Africa Conference on Biomedical Engineering (MECBME), 1–3. https://doi.org/10.1109/MECBME47393.2020.9265118

H. Gomi, H. Tanaka, & Y. Ando,. (2016). Detecting Fraudulent Behavior Using Recurrent Neural Networks. Computer Security Symposium, 11-13 October 2016 Pdfs.Semanticscholar.Org, 2016.

Haggag, M., Tantawy, M. M., & El-Soudani, M. M. S. (2020). Implementing a Deep Learning Model for Intrusion Detection on Apache Spark Platform. IEEE Access, 8, 163660–163672. https://doi.org/10.1109/ACCESS.2020.3019931

Herland, M., Khoshgoftaar, T. M, & Bauder, R. A. (2018). Big Data fraud detection using multiple medicare data sources. Journal of Big Data, 5(1).

Hormozi, H., Akbari, M. K., Hormozi, E., & Javan, M. S. (2013). Credit cards fraud detection by negative selection algorithm on hadoop (To reduce the training time). The 5th Conference on Information and Knowledge Technology, 40–43. https://doi.org/10.1109/IKT.2013.6620035

Jensen, M. (2013). Challenges of Privacy Protection in Big Data Analytics. 2013 IEEE International Congress on Big Data, 235–238. https://doi.org/10.1109/BigData.Congress.2013.39

Jha, B. K., Sivasankari, G. G., & Venugopal, K. R. (2020). Fraud Detection and Prevention by using Big Data Analytics. 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 267–274. https://doi.org/10.1109/ICCMC48092.2020.ICCMC-00050

Kamaruddin, Sk., & Ravi, V. (2016). Credit Card Fraud Detection using Big Data Analytics: Use of PSOAANN based One-Class Classification. Proceedings of the International Conference on Informatics and Analytics, 1–8. https://doi.org/10.1145/2980258.2980319

Kato, K., & Klyuev, V. (2017). Development of a network intrusion detection system using Apache Hadoop and Spark. 2017 IEEE Conference on Dependable and Secure Computing, 416–423. https://doi.org/10.1109/DESEC.2017.8073860

Makki, S., Haque, R., Taher, Y., Assaghir, Z., Ditzler, G., Hacid, M.-S., & Zeineddine, H. (2017). Fraud Analysis Approaches in the Age of Big Data—A Review of State of the Art. 2017 IEEE 2nd International Workshops on Foundations and Applications of Self* Systems (FAS*W), 243–250. https://doi.org/10.1109/FAS-W.2017.154

Melo-Acosta, G. E., Duitama-Munoz, F., & Arias-Londono, J. D. (2017). Fraud detection in big data using supervised and semi-supervised learning techniques. 2017 IEEE Colombian Conference on Communications and Computing (COLCOM), 1–6. https://doi.org/10.1109/ColComCon.2017.8088206

Patil, S., Nemade, V., & Soni, P. K. (2018). Predictive Modelling For Credit Card Fraud Detection Using Data Analytics. Procedia Computer Science, 132, 385–395. https://doi.org/10.1016/j.procs.2018.05.199

Purushe, P., & Woo, J. (2020). Financial Fraud Detection adopting Distributed Deep Learning in Big Data. KSII The 15th Asia Pacific International Conference on Information Science and Technology(APIC-IST) 2020.

Raghavan, P., & Gayar, N. E. (2019). Fraud Detection using Machine Learning and Deep Learning. 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), 334–339. https://doi.org/10.1109/ICCIKE47802.2019.9004231

Raghavendra Patidar & Lokesh Sharma. (2011). Credit Card Fraud Detection Using Neural Network. International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-NCAI2011, June 2011.

Richard J. Bolton, & David Hand. (2001). Unsupervised Profiling Methods for Fraud Detection. Credit Scoring and Credit Control, VII. https://www.researchgate.net/publication/2407747_Unsupervised_Profiling_Methods_for_Fraud_Detection

Sathyapriya, M., & Thiagarasu, D. V. (2015). Big Data Analytics Techniques for Credit Card Fraud Detection: A Review. International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391, 6(5), 6.

Singla, A., & Jangir, H. (2020). A Comparative Approach to Predictive Analytics with Machine Learning for Fraud Detection of Realtime Financial Data. 2020 International Conference on Emerging Trends in Communication, Control and Computing (ICONC3), 1–4. https://doi.org/10.1109/ICONC345789.2020.9117435

Song, Z. (2020). A Data Mining Based Fraud Detection Hybrid Algorithm in E-bank. 2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), 44–47. https://doi.org/10.1109/ICBAIE49996.2020.00016

Subudhi, S., & Panigrahi, S. (2020). Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection. Journal of King Saud University - Computer and Information Sciences, 32(5), 568–575. https://doi.org/10.1016/j.jksuci.2017.09.010

Terzi, D. S., Terzi, R., & Sagiroglu, S. (2017). Big data analytics for network anomaly detection from netflow data. 2017 International Conference on Computer Science and Engineering (UBMK), 592–597. https://doi.org/10.1109/UBMK.2017.8093473

Tiwari, P., Mehta, S., Sakhuja, N., Gupta, I., & Singh, A. K. (2021). Hybrid Method in Identifying the Fraud Detection in the Credit Card. In V. Suma, N. Bouhmala, & H. Wang (Eds.), Evolutionary Computing and Mobile Sustainable Networks (Vol. 53, pp. 27–35). Springer Singapore. https://doi.org/10.1007/978-981-15-5258-8_3

Vaishali, V. (2014). Fraud Detection in Credit Card by Clustering Approach. International Journal of Computer Applications, 98(3), 29–32. https://doi.org/10.5120/17164-7225

Yadav, D., Maheshwari, Dr. H., & Chandra, Dr. U. (2019). Big Data Hadoop: Security and Privacy. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3350308

Yufeng Kou, Chang-Tien Lu, Sirwongwattana, S., & Yo-Ping Huang. (2004). Survey of fraud detection techniques. IEEE International Conference on Networking, Sensing and Control, 2004, 2, 749–754. https://doi.org/10.1109/ICNSC.2004.1297040

Zhou, H., Sun, G., Fu, S., Wang, L., Hu, J., & Gao, Y. (2021). Internet Financial Fraud Detection Based on a Distributed Big Data Approach With Node2vec. IEEE Access, 9, 43378–43386. https://doi.org/10.1109/ACCESS.2021.3062467