The Investigation of Real-Time Credit Card Fraud Detection (RTCCFD) Based on Machine Learning and Apache Spark

Jiacheng He[a]
Big Data Management and Application, China University of Mining and Technology, Xuzhou, China

Keywords: Apache Spark, Credit Card Fraud Detection, Machine Learning.

Abstract: This paper presents a comprehensive review of the advancements in Real-time Credit Card Fraud Detection (RTCCFD), leveraging machine learning algorithms and Apache Spark. With financial fraud, particularly credit card fraud, posing significant losses to society and becoming increasingly prevalent due to the advent of new internet technologies, there is a pressing need for efficient detection systems. This study highlights the critical role of machine learning algorithms, such as Random Forest and Neural Networks, which, when integrated with Apache Spark, offer substantial improvements in processing speed and detection accuracy. Through an analysis of various research efforts, including the use of ensemble models and real-time processing frameworks, this paper demonstrates the effectiveness of these technologies in identifying fraudulent transactions. However, it also addresses the significant challenges that remain, including the lack of model interpretability, the need for models to generalize across evolving fraud tactics, and the imperative of ensuring data privacy and security in sensitive financial contexts. By discussing potential solutions like Federated Learning for enhancing privacy and suggesting directions for future research, this review aims to outline the progress made in the field while acknowledging the hurdles that lie ahead in the quest for more secure and trustworthy financial transactions.

1 INTRODUCTION

Financial fraud which could cause much loss to society has troubled people significantly especially obtaining property from people easily through the Internet. According to the China Judicial Big Data Research Institute report, credit card fraud cases accounted for 50.4% of all financial fraud cases in China from 2019 to 2021, totalling 3,375 instances, making it the most prevalent type (Zeng, 2022). Additionally, companies reliant on credit card transactions spend a substantial amount annually to prevent such incidents. However, as new Internet technologies develop, new fraud means are getting more day by day and harder to be detected. Typically, there are two main categories of credit card fraud: the first is off-line fraud, which involves the unauthorized use of a purloined card in physical locations such as retail stores, and the second is on-line fraud, which transpires via digital channels like the web, telephonic transactions, e-commerce platforms, or any situation where the card owner is not physically present to authorize the transaction (Chaudhary, 2012).

Considering the instantaneous nature and terrifying consequences of credit card fraud cases, the detection system must receive and process all the real-time data then send an alarm to cardholder immediately. Apache Spark is a distributed, general-purpose computing platform which is similar to Hadoop. However, Spark differentiates itself by allowing large amounts of data to be kept in memory, which offers significant performance improvements, making it up to 100 times faster than MapReduce in some scenarios. MLlib is a module of Apache Spark frame, enabling users to run machine learning algorithms on big data with higher efficiency to speed up all the process.

Training credit card fraud detection models with Spark and machine learning has become the mainstream method. Many scholars utilized various machine learning algorithms which are realized by Mllib to predict credit card fraud and proved many benefits with Spark. Madhavi et al. used random

[a] https://orcid.org/0009-0008-8827-4618

He, J.
The Investigation of Real-Time Credit Card Fraud Detection (RTCCFD) Based on Machine Learning and Apache Spark.
DOI: 10.5220/0012900800004508
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence (EMITI 2024)*, pages 52-56
ISBN: 978-989-758-713-9

forest ensemble model with Mllib and real-time processing using Kafka and Spark streaming jobs delivered the optimal results (Madhavi, 2021). Armel et al. compared the performance of simple anomaly detection algorithm, the naïve bayes algorithm, decision trees classifier algorithm and random forest algorithm further showing that random forest is a strong algorithm to handle credit card fraud (Armel, 2019). Ananthu et al. also provided a comparison of machine learning techniques and focused on recognizing fraudulent transactions through the analysis of previous transaction records (Ananthu, 2021). On the other hand, Cornelius et al. investigated prior spending habits to detect fraud in real-time transactions, utilizing distributed frameworks like Spark and Kafka and Cassandra for scalability (Nwankwo, 2023). Preprocessing with Spark Machine Learning Pipeline Stages for efficient fraud detection was emphasized. Alshammari and colleagues explored the swift proliferation and development of digital transactions, which have propelled credit cards to become the predominant method of payment (Alshammari, 2022). Contemporary research in the realm of credit card fraud detection has extensively examined the selection of machine learning algorithms, aiming to identify which could offer enhanced privacy and a heightened sense of security for businesses, as well as methods to increase the promptness of fraud surveillance. This paper seeks to summarize the progress made in this field, identify areas of weakness that require enhancement, and explore new potential fraud scenarios that need to be addressed. This article is organized into four main sections. Following the introduction, Section 2 delves into the framework of machine learning-based credit card fraud detection, detailing the methodology and various algorithms employed. Section 3 discusses the current challenges and limitations in the field and suggests potential directions for future research. The final section, Section 4, summarizes the key findings and underscores the importance of advancing these technologies for more effective fraud detection.

2 METHOD

2.1 Framework of Machine Learning-Based Credit Card Detection

The usual steps to build a system to detect credit card detection include data collection, preprocessing, model building etc shown in Figure 1. The first module is data collection which is the most basic step. For instance, Varmedja et al found some real transactions datasets which are helpful to model training and prediction but cause much error because of its imbalance (Alshammari, 2022). ARMEL et al. simulated the datasets by some kinds of distribution to test their models without prediction (Madhavi, 2021). Preprocessing includes data cleaning, feature selection, balancing the dataset etc. Various techniques have been implemented. For instance, D. Varmedja et al. used Synthetic Minority Oversampling Technique (SMOTE) technique for oversampling (Varmedja, 2019). Many Internet factories have published their ways to develop models to detect anomalies. Google, Twitter, and Netflix employ machine learning and statistical methods for anomaly detection in data streams. Google uses TensorFlow to train models like Deep Neural Networks (DNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memories (LSTMs) for regression and anomaly detection, combining two methods for effective anomaly identification. Twitter's approach utilizes the Seasonal Hybrid Extreme Studentized Deviate test (S-H-ESD) for detecting both global and local anomalies, considering seasonality. Netflix's Robust Anomaly Detection (RAD) leverages Robust Principal Component Analysis (RPCA) to handle high cardinality datasets, enabling quick anomaly identification and response, enhancing customer experience.

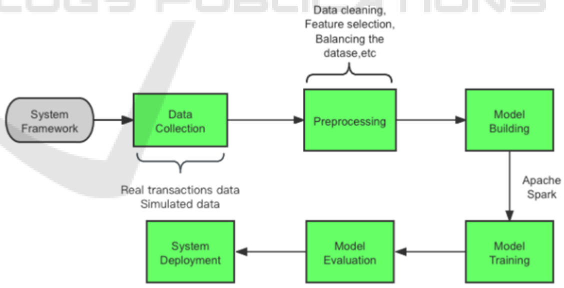

Figure 1: The architecture of machine learning-based credit card detection (Photo/Picture credit: Original).

Apache Spark is an open-sourced distributed engine which can accelerate machine learning algorithms to handle real-time big data. Figure 2 provides its components and principle. At its core, Spark uses Resilient Distributed Datasets (RDDs), an immutable collection of objects that can be processed in parallel. Each RDD can be partitioned across the cluster, allowing Spark to execute operations on each partition in parallel, significantly speeding up processing. The structural design of Spark incorporates a driver program responsible for initiating the user's primary function and performing a range of concurrent operations across the computing cluster. The cluster manager allocates resources across applications, while worker nodes execute tasks assigned to them. This model enables efficient data processing and analysis at scale. Deploying the detection system with Spark after passing the model evaluation is the last step which is the hardest and needs huge resources.

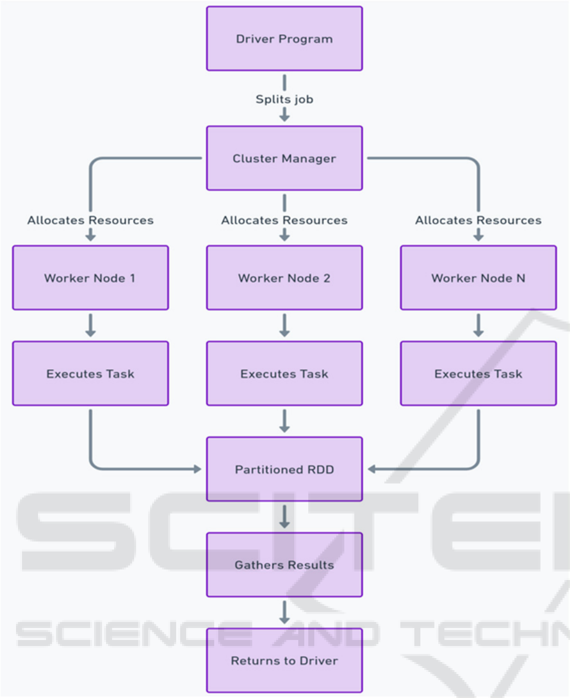

Figure 2: Components and principle of Apache Spark (Photo/Picture credit: Original).

2.2 Random Forest-Based Detection

The Random Forest approach employs an ensemble of learning models, specifically for tasks involving classification and regression. It operates through a collection of decision trees at the training phase and delivers the class median or the average prediction from the trees. It has been widely used to predict credit card fraud, almost the most popular, because this algorithm has been proved that it is very suitable to handle fraud detection. Armel et al. and Ananthu et al. gave their results of comparative of experiments proving random forest indeed a great choice to handle this (Armel, 2019; Ananthu, 2021). Furthermore, Mehvish proposed a strategy for detecting credit card transactions by employing a hybrid approach that integrates Random Forest with Extreme Learning Machine algorithms (Mehvish, 2023). They provided a foundation for the development for algorithms that could perform better. Rajesh PK et al. used Bayesian optimized random forest classifier on big dataset before feature engineering gaining 99.545% such high accuracy (PK, 2023). Contemporary research indicates that while Random Forest algorithms yield positive results in identifying credit card fraud, integrating them with other foundational algorithms could further improve detection effectiveness.

2.3 Neural Network-Based Detection

Drawing inspiration from the human brain's architecture and operational principles, neural networks serve as sophisticated computational constructs that excel at pattern recognition and problem-solving. These networks are comprised of multiple neuron-like units arranged in layers, with each linkage reflecting a synaptic weight. Diverse forms of neural networks exist, distinguished by their specialized features. Convolutional Neural Networks, for instance, are adept at analyzing visual information through their layered pattern identification capabilities, whereas Recurrent Neural Networks are tailored for time-sequenced data such as time series analysis or language processing, owing to their ability to retain information from earlier inputs. Deep Learning involves networks with many layers, enabling the extraction of high-level features from raw input. Atchaya used Artificial Neural Network (ANN), Support Vector Machine (SVM) and K Nearest Neighbors (KNNs) in predicting showing that ANN has the best accuracy 99.92% (Atchaya, 2024). Karthika et al. introduced a smart fraud detection utilizing a dilated CNN combined with a sampling technique, focusing on enhancing the detection accuracy of fraudulent transactions (Karthika, 2023). Berhane et al. also tried to develop a hybrid model which combines CNN and SVM called CNN-SVM (Berhane, 2023). As he said, CNN-SVM is more capable of handling fraud detection.

3 DISCUSSIONS

Based on the research progresses mentioned above limitations and challenges like model interpretability, model generalizability and data privacy and security can be identified and will be discussed in this part. This study has demonstrated the efficacy of machine learning algorithms, especially Random Forest and Neural Networks, in real-time detecting credit card fraud, leveraging the computational power of Apache Spark. However, one significant limitation of employing machine learning models like them is their lack of interpretability. They always be used to handle some actual and complex problems as “black boxes”, making it challenging to understand the rationale behind their predictions (Qiu, 2024). Molnar emphasizes the importance of model interpretability for ensuring transparency and accountability in machine learning applications, suggesting that interpretable models are crucial for gaining stakeholder trust and facilitating wider adoption (Molnar, 2020).

The generalizability of machine learning models, particularly in the context of fraud detection, is a critical concern. Models trained on historical data may not perform well on unseen data or adapt to evolving fraud patterns, leading to decreased detection accuracy over time. Domingos highlights the importance of creating models that not only learn from past data but also adapt to new patterns dynamically (Domingos, 2012). Furthermore, Goodfellow et al. discuss the concept of adversarial examples that can exploit model vulnerabilities, underscoring the need for robust machine learning models capable of generalizing across a broad spectrum of fraud tactics (Goodfellow, 2016). Addressing these concerns requires continuous model evaluation and updating, alongside the development of algorithms that can learn and adapt in real-time to maintain effectiveness in fraud detection.

The integration of machine learning in sensitive domains, such as financial fraud detection, raises significant privacy and security concerns. Traditional machine learning approaches often require centralized data collection, posing risks to user privacy and data security. To mitigate these issues, some machine learning algorithms, take Federated Learning (FL) for example. It emerges as a promising solution by enabling model training on decentralized data sources without needing to share the data itself. McMahan and colleagues pioneered the use of Federated Learning, a technique that allows for model training across several devices without centralizing data, thereby bolstering data privacy and system security (McMahan, 2017). Besides, Bonawitz et al. discuss advancements in secure aggregation protocols within FL, ensuring that individual updates cannot be inspected by the server, thus offering an additional layer of privacy (Bonawitz, 2019). These developments in Federated Learning not only address privacy concerns but also open new avenues for secure, decentralized machine learning applications. However, challenges remain in ensuring robustness against adversarial attacks and maintaining model performance with non-Independently and Identically Distributed (IID) data across devices. Addressing these challenges is crucial for the widespread adoption of FL in privacy-sensitive applications.

4 CONCLUSIONS

In conclusion, this paper has explored the application of machine learning algorithms, particularly Random Forest and Neural Networks, in conjunction with Apache Spark for RTCCFD. This investigation highlights the significant potential of these technologies to enhance the speed and accuracy of fraud detection systems, thereby offering a more secure transaction environment for both companies and consumers. However, this study also acknowledges the inherent challenges associated with these technologies, including issues of model interpretability, generalizability, and data privacy and security.

Future research should focus on addressing these challenges by developing more interpretable machine learning models, enhancing their adaptability to new fraud patterns, and ensuring the privacy and security of sensitive data. Collaborative efforts between academia, industry, and regulatory bodies will be essential in advancing these technologies and ensuring their effective and ethical application in combating credit card fraud.

REFERENCES

Alshammari, A., Alshammari, R., Altalak, M., & Alshammari, K. 2022. Credit-card Fraud Detection System using Big Data Analytics. In 2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME) (pp. 1-7). IEEE.

Ananthu, S., Sethumadhavan, N., & AG, H. N. 2021. Credit card fraud detection using Apache Spark analysis. In 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 998-1002). IEEE.

Armel, A., & Zaidouni, D. 2019. Fraud detection using apache spark. In 2019 5th International Conference on Optimization and Applications (ICOA) (pp. 1-6). IEEE.

Berhane, T., Melese, T., Walelign, A., & Mohammed, A. 2023. A Hybrid Convolutional Neural Network and Support Vector Machine-Based Credit Card Fraud Detection Model. Mathematical Problems in Engineering, 2023.

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., ... & Roselander, J. 2019.

Towards federated learning at scale: System design. Proceedings of Machine Learning and Systems, 1, 374-388.

Chaudhary, K., Yadav, J., & Mallick, B. 2012. A review of fraud detection techniques: Credit card. International Journal of Computer Applications, 45(1), 39-44.

Domingos, P. 2012. A few useful things to know about machine learning. Communications of the ACM, 55(10), 78-87.

Goodfellow, I., Bengio, Y., & Courville, A. 2016. Deep learning. MIT Press.

Karthika, J., & Senthilselvi, A. 2023. Smart credit card fraud detection system based on dilated convolutional neural network with sampling technique. Multimedia Tools and Applications, 1-18.

Madhavi, A., & Sivaramireddy, T. 2021. Real-Time Credit Card Fraud Detection Using Spark Framework. In Machine Learning Technologies and Applications: Proceedings of ICACECS 2020 (pp. 287-298). Springer Singapore.

McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In Artificial Intelligence and Statistics (pp. 1273-1282). PMLR.

Mehvish. 2023. Random Forest and Extreme Learning Machine Algorithms for High Accuracy Credit Card Fraud Detection. https://www.ijraset.com/best-journal/credit-card-fraud-detection-using-ann

Molnar, C. 2020. Interpretable machine learning. Lulu.com.

Muhammad, K., Ullah, A., Lloret, J., et al. 2020. Deep learning for safe autonomous driving: Current challenges and future directions. IEEE Transactions on Intelligent Transportation Systems, 22(7): 4316-4336.

Nwankwo, U. C., Onuora, J. N., Obi, J. N., Obiukwu, E. N., & Okore, U. E. 2023. Detection of Credit Card Fraud in Real Time Using Spark ML. International Journal of Computer Science and Mobile Computing, 12(12). https://dx.doi.org/10.47760/ijcsmc.2023.v12i12.006

PK, R. 2023. Enhanced Credit Card Fraud Detection: A Novel Approach Integrating Bayesian Optimized Random Forest Classifier with Advanced Feature Analysis and Real-time Data Adaptation. International Journal for Innovative Engineering & Management Research, Forthcoming.

Qiu, Y., Hui, Y., Zhao, P., Cai, C. H., Dai, B., Dou, J., ... & Yu, J. 2024. A novel image expression-driven modeling strategy for coke quality prediction in the smart cokemaking process. Energy, 130866.