Refining Credit Card Fraud Detection with Human In The Loop Approach

Aditya Kulkarni
Simon Fraser University
Burnaby, BC, Canada
aka163@sfu.ca

Johann Tin Tsin Fong
Simon Fraser University
Burnaby, BC, Canada
jtintsin@sfu.ca

ABSTRACT

Credit card fraud remains a pervasive challenge in today's digital economy, necessitating innovative solutions that combine advanced technology with human expertise. In this report, we propose a collaborative Human-in-the-Loop (HITL) approach to credit card fraud detection, aiming to improve fraud detection systems. Through our exploration of related work, we identify key limitations in existing automated systems and highlight the potential benefits of integrating human intelligence into the detection process. Our solution leverages machine learning algorithms for automated analysis of transaction data and provides a human element to review flagged transactions and provide feedback. By intertwining machine learning models with human insights, our proposed HITL framework aims to improve fraud detection accuracy and adaptability.

1 INTRODUCTION

In the landscape of credit card fraud detection, prevailing approaches heavily rely on automated systems predominantly driven by machine learning algorithms [5]. These systems have showcased effectiveness in specific scenarios, leveraging the power of algorithms to identify patterns indicative of fraudulent activities. However, a common assumption underlying these automated approaches is that they can comprehensively adapt to the dynamic nature of credit card fraud. This assumption, while valid in some instances, becomes a critical point of contention as fraudsters continually refine and evolve their tactics.

The limitations of current automated systems [1–12] become apparent when confronted with the intricate nuances of evolving fraudulent behaviour. While these systems excel at identifying straightforward patterns, they may fall short in capturing the subtle, nuanced strategies employed by fraudsters. As fraud techniques become more sophisticated, the automated systems may struggle to keep pace, raising concerns about their adaptability to the ever-changing landscape of credit card fraud. This prompts a critical assessment of the existing solutions, recognizing the need for a more adaptive and context-aware framework that can discern the intricacies of fraudulent activities.

Amidst the evolving challenges in credit card fraud detection, there is a growing recognition of the necessity for innovative approaches that go beyond the limitations of purely automated systems. Existing fraud detection systems heavily rely on machine learning algorithms. While effective in certain scenarios, they struggle to keep pace with evolving fraud techniques due to their inability to capture nuanced patterns. Our solution introduces a collaborative framework where machine learning models and human analysts work together to detect credit card fraud. This HITL approach aims to leverage the strengths of both automated systems and human intuition to create a more adaptive and context-aware fraud detection system.

In this report, we provide a comprehensive overview of the current state-of-the-art methodologies in credit card fraud detection and compare them with our proposed HITL approach. We discuss the shortcomings of existing techniques and emphasize the need for a dynamic feedback loop between automated systems and human analysts to enhance detection accuracy and adaptability.

Furthermore, we outline the core implementations of our HITL framework: the Feedback Loop and Continuous Improvement and Adaptation strategies. In the Feedback Loop, automated systems identify potential fraud cases, which are then reviewed by human analysts to provide nuanced insights and validate suspicions. The Continuous Improvement and Adaptation strategy focuses on the ongoing evolution of the system through the integration of new data and emerging fraud patterns identified by human analysts.

Through rigorous evaluations, we substantiate the efficacy of our proposed approach by demonstrating higher accuracy rates in identifying fraudulent transactions compared to existing works. Additionally, we highlight the added value of integrating human expertise in enhancing the detection capabilities of credit card fraud detection systems.

2 RELATED WORK

The landscape of credit card fraud detection is rapidly evolving, necessitating the integration of sophisticated technologies and human expertise to combat fraudulent activities effectively. This section compares the current state-of-the-art methodologies, with our proposed Human-in-the-Loop (HITL) approach.

Machine Learning Techniques and Graph-based Models
Recent advancements in fraud detection have leveraged machine learning (ML) and graph-based models to analyze transaction data and identify fraudulent patterns. Our approach employs a traditional ML approach, involving data prepossessing, model training and valuation to detect the fraudulent transactions. The use of machine learning techniques for fraud detection [4] and the work in the papers [6, 7] utilize deep learning and graph neural networks to capture complex transaction relationships and patterns indicative of fraud. While these approaches demonstrate significant improvements in detection accuracy, they often lack the adaptability to swiftly respond to evolving fraud techniques due to their reliance on historical data patterns. Our HITL approach reflects this challenge and introduces a HITL approach to address predictions based a threshold, enhancing model reliability through manual corrections. This enhances these models by incorporating human insights

into the analysis process, allowing for real-time adaptability and contextual interpretation that automated systems may overlook.

Semi-supervised and Weakly Supervised Learning
The studies in A Framework for Detecting Frauds from Extremely Few Labels [1] and the frameworks discussed in [2, 8] explore semi-supervised and weakly supervised learning techniques to overcome the challenge of limited labeled data in fraud detection. These methodologies are innovative in their approach to leveraging unlabeled data for model training. However, they may still encounter difficulties in capturing new fraud patterns without explicit labeling. Our framework addresses this limitation by utilizing human analysts to identify and label new fraud patterns, thus enriching the training dataset and enhancing model performance.

Human-in-the-Loop (HITL) Approaches
Although HITL frameworks have been explored in various domains, their application in fraud detection remains underutilized. We implemented a HITL system where uncertain cases are flagged for manual review by domain experts, demonstrating an initial first step towards integrating human expertise with automated systems. Our research fills this gap by systematically integrating human expertise. Unlike the solely automated systems explored in [9–12], our HITL framework facilitates a dynamic feedback loop where human analysts review, verify, and provide nuanced feedback on flagged transactions. This collaborative effort not only improves the accuracy of fraud detection but also ensures the system remains agile and responsive to novel fraud techniques.

Improving Accuracy and Adaptability
The works by [10, 12] and similar studies have focused on enhancing the precision and recall of fraud detection systems through advanced algorithmic solutions. While these efforts have yielded improved detection rates, they often do not account for the rapidly changing tactics employed by fraudsters. Our solution bridges this gap by empowering human analysts to directly influence the detection process, allowing for immediate adjustments based on emerging trends and ensuring a higher degree of adaptability.

In summary, while existing research has significantly advanced credit card fraud detection through the use of machine learning, graph-based models, and semi-supervised learning techniques, our proposed HITL framework stands out by effectively marrying the strengths of automated analysis with the critical thinking and adaptability of human intelligence. This collaborative approach promises not only to enhance the immediate accuracy of fraud detection systems but also to ensure their sustained effectiveness in the face of evolving fraud strategies.

3 METHOD

This section deals with how we implemented our solution.

3.1 Finding the Dataset

We obtained our dataset from Kaggle, selecting it based on its popularity and suitability for our Human-in-the-Loop (HITL) approach. This dataset encompasses credit card transactions conducted by European cardholders throughout September 2013, offering a comprehensive view of real-world financial activities. Despite its temporal scope of just two days, the dataset records a significant number of transactions, totaling 284,807 instances. Within this timeframe, 492 cases of fraud were identified, illustrating the rarity of fraudulent activities within the dataset, which account for a mere 0.172% of all transactions.

This dataset primarily comprises numerical input variables, with features derived from a Principal Component Analysis (PCA) transformation. However, it's important to note that due to confidentiality concerns, the original features and additional background information are not disclosed. The transformed features, labeled as V1 through V28, encapsulate the principal components resulting from the PCA process. The dataset also includes two non-transformed features: 'Time' and 'Amount'. 'Time' denotes the elapsed seconds between each transaction and the first recorded transaction in the dataset, providing temporal context. Meanwhile, 'Amount' signifies the monetary value of each transaction, offering crucial information for further analysis and model development.

The 'Class' feature serves as the response variable, indicating the fraudulent status of each transaction. It takes a binary value of 1 for fraudulent transactions and 0 for legitimate ones, enabling the development of predictive models to discern between the two classes effectively. Despite the unavailability of original features and detailed background information, this dataset provides a valuable resource for exploring credit card fraud detection techniques and refining our HITL approach.

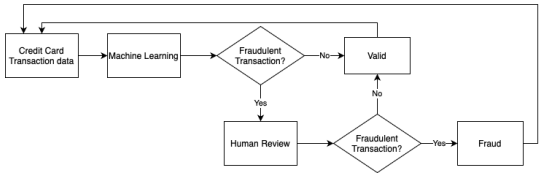

Figure 1: Architecture

3.2 Data Gathering and Assessment

We initially imported the dataset using Pandas, reading it from the provided CSV file. This allowed us to inspect the structure and contents of the data to gain insights into its distribution and characteristics. We observed an uneven class distribution, with a significantly higher number of non-fraudulent transactions compared to fraudulent ones.

3.3 Class Balancing

To address the class imbalance issue, we performed class balancing by randomly sampling a subset of non-fraudulent transactions to match the number of fraudulent transactions. This step ensured a balanced representation of both classes in the dataset, thereby preventing the model from being biased towards the majority class.

3.4 Feature Extraction

We separated the dataset into feature variables (X) and the target variable indicating the transaction's fraudulent status (y). The 'Class' column served as the target variable, while the remaining columns were considered as features. This facilitated the extraction of relevant information for model training and evaluation.

3.5 Train-Test Split

We split the data into training and testing sets using the train_test_split function from scikit-learn. The data was partitioned into training data (X_train, y_train) used for model training and testing data (X_test, y_test) used for model evaluation. Stratification was applied to ensure an equal distribution of both classes in the training and testing sets.

3.6 Feature Scaling

To ensure that all features were on a similar scale and to prevent any feature from dominating the model training process, we applied StandardScaler to the feature variables. This transformation normalized the feature values, bringing them within a similar range and improving the stability and convergence of the model during training.

3.7 Training the Model

We used a Convolutional Neural Network (CNN) model for credit card fraud detection. The input data, comprising features related to credit card transactions, is reshaped to accommodate the requirements of the CNN architecture. Utilizing the Keras library, the model is constructed using the Sequential API, allowing for a linear stack of layers. The architecture includes two Conv1D layers with ReLU activation, followed by batch normalization and dropout layers to enhance training stability and prevent overfitting. Subsequently, the flattened output is fed into fully connected layers with ReLU activation, further regularized by dropout layers. Finally, a sigmoid activation function in the output layer provides the binary classification prediction, indicating the likelihood of a transaction being fraudulent. This approach aims to leverage the hierarchical feature extraction capabilities of CNNs to effectively detect fraudulent activities in credit card transactions.

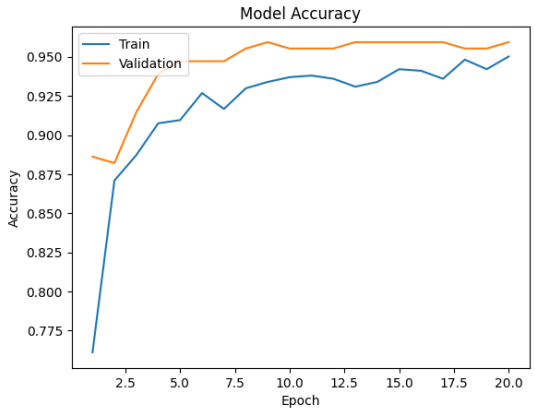

Figure 2: Model Accuracy

3.8 Adding a Human Element

In our HITL approach to credit card fraud detection, the "human review" function serves as a pivotal component in integrating human judgment into the detection process. This function facilitates

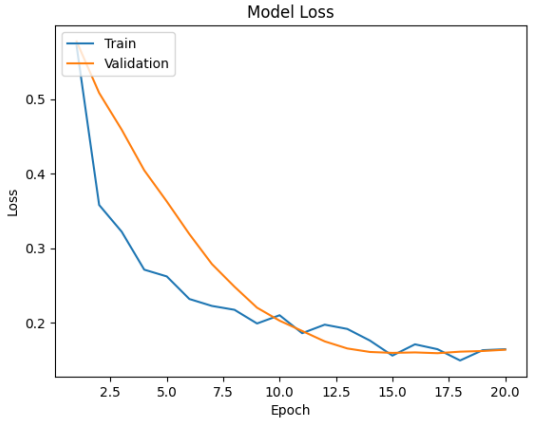

Figure 3: Model Loss

human intervention in reviewing and correcting model predictions near the decision boundary. It accepts the trained machine learning model and the feature matrix of test data. By iteratively examining cases where the model's prediction probability falls within a predefined uncertainty threshold, the function prompts the user to input the correct label (0 for non-fraud, 1 for fraud) for each case. If the human input differs from the model's prediction, the function records the correction details (index, predicted probability, correct label). These corrections are then saved to a CSV file named "corrections.csv" for further analysis. Through this collaborative process, our HITL approach harnesses both machine learning algorithms and human expertise to refine and enhance credit card fraud detection.

4 EVALUATION

We suggest that incorporating a HITL approach will enhance the accuracy of detecting fraud detection when using a ML model for fraud detection.

4.1 Results

In assessing the performance of our ML model, both quantitative and qualitative evaluations were conducted. The model achieved an overall accuracy rate exceeding 90% over 280,000 transactions. Transactions predicted to score below a 50% threshold were flagged for human review, which initially resulted in only 3% accuracy. The 50% threshold was chosen because it strikes a balance between detecting fraud and managing false alarms. When considering all the transactions below 50% predicted score as fraudulent, the model's accuracy improved from 3% to 63%. This variability in accuracy highlights the critical role that the threshold settings and the nature of transactions play in determining the model's effectiveness in detecting fraud. Such findings underscore the importance of continuously refining the model's predictive capabilities to enhance its accuracy and reliability in different scenarios. Additionally, incorporating an ongoing feedback loop and real-time analytics into the model could help in dynamically adjust thresholds to better respond to new fraud trends.

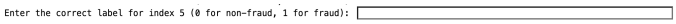

Figure 4: UI of Human Review

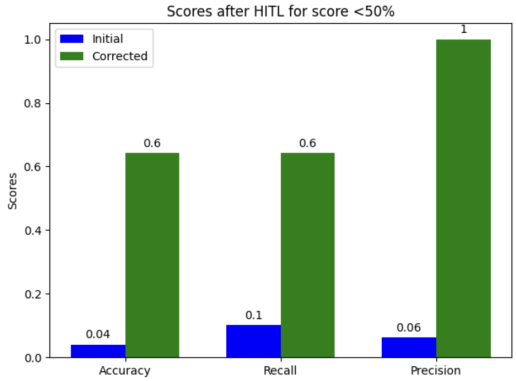

Figure 5: Model Accuracy, recall and precision after HITL if all transactions below 50% are fraud

The recall of the model is 10%, highlighting a poor performance to detect fraudulent transactions. The low recall rate implies that the majority of fraudulent activities would pass undetected, potentially leading to a fraudulent transactions going undetected. After the HITL, the recall improved to 60%, suggesting that that the model become more effective at identifying fraudulent transactions.

On the other hand, the precision of the model was also low at 6%. This indicates that a high number of legitimate transactions were incorrectly flagged as fraudulent. This could likely lead to customer frustrations due to wrongful blocks. After the HITL, the precision increased to 100%. This perfect precision score assumes that there is no bias in human judgement and that the human evaluator's decisions are accurate.

The results of the evaluation are in line with our initial expectations. Implementing a HITL approach was anticipated to improve both precision and recall of the ML model in fraud detection.The increase in recall from 10% to 60% shows the model's enhanced ability to identify fraudulent transactions with human oversight. Overall, the results validate our approach and support the application of HITL strategies within fraud detection processes.

5 LIMITATIONS

The Human-in-the-Loop (HITL) approach to credit card fraud detection offers promising avenues for improved accuracy and adaptability. However, several inherent limitations need to be addressed to improve its effectiveness.

5.1 Time Cost of Human Review

The incorporation of human reviewers introduces a time-intensive element to the fraud detection workflow. Each flagged transaction necessitates human scrutiny, potentially leading to delays in the detection process. Moreover, the need for skilled analysts to ensure thorough examination further compounds this time cost. Mitigating this limitation requires streamlining review procedures and possibly investing in automation technologies to expedite the process without sacrificing accuracy.

5.2 Requirement of Qualified Reviewers

A fundamental assumption in the HITL framework is the availability of qualified personnel to review flagged transactions. However, identifying and training individuals proficient in fraud detection techniques poses a considerable challenge. Training programs must cover a diverse range of fraud scenarios and evolving tactics, demanding ongoing investment of time and resources. Overcoming this limitation may involve developing specialized training modules and fostering collaborations with industry experts to enhance the skill set of reviewers, thereby ensuring the effectiveness of the human component in the detection process.

5.3 Potential for Human Bias

Human reviewers are susceptible to biases that can influence their assessment of flagged transactions, potentially skewing detection outcomes. These biases may stem from subjective interpretations, prior experiences, or contextual factors, leading to inconsistencies and unfair treatment of certain cases. Addressing this limitation requires implementing robust protocols to mitigate bias, such as blind review procedures, continuous training on bias awareness, and regular audits of reviewer decisions. Additionally, leveraging diverse teams of reviewers with varied backgrounds and perspectives can help counteract individual biases and promote impartiality in the review process.

5.4 Scalability Challenges

As transaction volumes escalate, the scalability of the HITL framework becomes a pressing concern. The manual review process may struggle to keep pace with the influx of transactions, leading to bottlenecks and potential backlogs. Scaling up human resources to match transaction volumes presents logistical and financial challenges, necessitating innovative solutions to optimize workflow efficiency. Implementing automated triaging systems to prioritize high-risk transactions for human review, deploying scalable infrastructure, and adopting agile workforce management strategies are potential approaches to address scalability concerns effectively.

6 FUTURE WORK

The Human-in-the-Loop (HITL) approach for detecting credit card fraud shows potential for better accuracy and adaptability. But, there's still room to make it even better. Here are the following ways we plan on enhancing the proposed approach.

6.1 Integration of Unstructured Data

Future research could focus on incorporating unstructured data sources, such as social media activity and geolocation information, into the fraud detection framework. This integration could provide valuable contextual insights that enhance the accuracy of fraud detection algorithms. Potential solutions may involve natural language processing techniques to extract relevant information from textual data and advanced data fusion methods to integrate unstructured data with transactional data effectively.

6.2 Real-Time Detection

Exploring real-time detection capabilities would enable prompt identification and response to fraudulent activities as they occur. This could involve the development of predictive models that analyze transaction data in real-time and trigger alerts for suspicious transactions. Implementing high-speed data processing systems and leveraging machine learning algorithms optimized for real-time analysis could facilitate this objective.

6.3 Blind Review to Combat Bias

Implementing a blind review process could help mitigate biases in human review by concealing certain transaction information from reviewers. Determining which information to hide requires careful consideration to ensure that essential details for fraud detection remain accessible. Potential solutions may involve hiding personal identifiable information (PII) or transaction amounts during the initial review stage, allowing reviewers to focus solely on transaction patterns and behavior. Additionally, incorporating randomized sampling techniques and regular audits can further enhance the effectiveness of blind reviews in combating bias.

7 CONCLUSION

In our paper the application of a Human-in-the-Loop approach to our machine learning model for fraud detection has shown advantages in terms of improved accuracy and reliability. By implementing human insights, results show significant improvements in recall and precision. The results from this paper confirms the vital role of human collaboration together with systems to building more resilient and accurate fraud detection systems.

REFERENCES

[1] Ayesha Aslam and Adil Hussain. 2024. A Performance Analysis of Machine Learning Techniques for Credit Card Fraud Detection. *Journal on Artificial Intelligence* 6 (2024), 1–21. https://doi.org/10.32604/jai.2024.047226

[2] F. Braun, O. Caelen, E. N. Smirnov, S. M. Kelk, and B. Lebichot. 2017. Improving Card Fraud Detection Through Suspicious Pattern Discovery. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*.

[3] Chengliang Chai, Lei Cao, Guoliang Li, Jian Li, Yuyu Luo, and Samuel Madden. [n. d.]. *Human-in-the-loop Outlier Detection.* Retrieved 2024-03-09 from https://dl-acm-org.proxy.lib.sfu.ca/doi/10.1145/3318464.3389772

[4] A. Correa Bahnsen, A. Stojanovic, D. Aouada, and B. Ottersten. 2014. Improving Credit Card Fraud Detection with Calibrated Probabilities. (04 2014).

[5] Abdul Rehman Khalid, Nsikak Owoh, Omair Uthmani, Moses Ashawa, Jude Osamor, and John Adejoh. [n. d.]. *Enhancing Credit Card Fraud Detection: An Ensemble Machine Learning Approach.* Retrieved 2024-03-10 from https://www.mdpi.com/25042289/8/1/6#:~:text=These%20approaches%20involve%20combining%20multiple,overall%20predictive%20power%20%5B7%5D.

[6] A. R. Khalid, N. Owoh, O. Uthmani, M. Ashawa, J. Osamor, and J. Adejoh. 2024. Enhancing credit card fraud detection: An ensemble machine learning approach. *Big Data Cogn. Comput.* 8, 1 (Jan 2024), 6.

[7] B. Lebichot, F. Braun, O. Caelen, and M. Saerens. 2017. A graph-based, semi-supervised, credit card fraud detection system. 693 (11 2017), 721–733.

[8] Q. Li, Y. He, C. Xu, F. Wu, J. Gao, and Z. Li. 2022. Dual-Augment Graph Neural Network for Fraud Detection. In *Proceedings of the 31st ACM International Conference on Information Knowledge Management.* Atlanta, GA, USA, 4188–4192.

[9] C. Liu, Y. Gao, L. Sun, J. Feng, H. Yang, and X. Ao. 2022. User Behavior Pretraining for Online Fraud Detection. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* Washington DC, USA, 3357–3365.

[10] S. Xiang. 2023. Semi-supervised Credit Card Fraud Detection via Attribute-Driven Graph Representation. *AAAI* 37, 12 (Jun 2023), 14557–14565.

[11] C. Zhang et al. 2021. Fraud Detection under Multi-Sourced Extremely Noisy Annotations. In *Proceedings of the 30th ACM International Conference on Information Knowledge Management.* Virtual Event, Queensland, Australia, 2497–2506.

[12] Y.-L. Zhang et al. 2023. A Framework for Detecting Frauds from Extremely Few Labels. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, WSDM 2023.* Singapore, 27 February 2023 - 3 March 2023, 1124–1127.