UNIOFFA JOURNAL OF INTERDISCIPLINARY STUDIES (OJIS)
A Multidisciplinary Journal of Science, Management and Social Sciences
Website: www.unioffajis.com

FRAUD DETECTION SYSTEM USING DEEP LEARNING WITH SHAPLEY ADDITIVE EXPLANATIONS MODELS

Asaju-Gbolagade, Ayisat Wuraola
University of Ilorin
Saka Kamil
Al-hikmah University
Seriki Aliu Adebayo
, , Abdulrasaq Abubakar Toyin University
Correspondence Email ; **gbolagade.aaw@unilorin.edu.ng**

Volume 1.1

ABSTRACT

Financial fraud poses a significant and evolving threat in the landscape of electronic transactions. This study presents the design, implementation, and evaluation of an explainable AI (XAI) system for credit card fraud detection. The system integrates a deep neural network (DNN) autoencoder for fraud detection, SHAP (SHapley Additive exPlanations) for model interpretation. Results demonstrate the effectiveness of the DNN in detecting fraudulent transactions and the ability of SHAP to provide insights into influential features. Notably, user feedback revealed a strong preference for LLM-generated narrative explanations over traditional SHAP visualizations. This research contributes to the field of XAI in finance by showcasing a practical approach to enhancing the interpretability and user-friendliness of fraud detection systems, emphasizing the importance of accessible explanations in building user trust, and understanding in AI driven financial decision-making.

Keywords: SHAP (SHapley Additive exPlanations), Financial fraud, Large Language Model (LLM), Fraud Detection System

INTRODUCTION

Financial fraud poses a significant and evolving threat in the landscape of electronic transactions. The surge in digital payment systems has expanded the attack surface for fraudsters, leading to increasingly sophisticated methods of perpetrating fraud. Traditional rule-based fraud detection systems have been partially effective, but they often struggle to keep pace with the dynamic nature of fraudulent activities Swamy et al., (2023). In our world today, payments can made through varying devices anywhere, with real-time fulfilment a basic requirement, human analysis alone cannot catch fraud. Rules-based detection methods which superseded humans in speed now struggle to find new patterns for fraud in this dynamic, demanding environment. Mill et al., (2023).

In response to this challenge, companies now employ machine learning (ML) models for real-time identification of patterns in transaction data indicative of fraud Mill et al., (2023). These models have shown promise for analysing transaction data. But their decision-making is opaque, and their output is unexplainable Swamy et al., (2023).

Machine learning models, such as random forests and neural networks, have introduced a change in thinking about fraud detection by finding complex patterns and relationships within transaction data. However, the inherent 4 black-box nature of these models hinders their explainability. Financial institutions are left without a clear understanding of the factors influencing the models' decisions. Secondly, regulatory bodies and compliance frameworks increasingly demand accountability and transparency in the decision-making process, especially in the sensitive domain of financial transactions Dwivedi et al., (2019). Moreover, end-users and customers are rightfully concerned about the opacity of automated systems making decisions about their financial activities. A lack of explanations for why a particular transaction is fraudulent can lead to mistrust and dissatisfaction Ijeoma et al., (2013). Therefore, the challenge at hand is not only to improve the accuracy of fraud detection but also to ensure that the decision-making process is transparent and explained to specific stakeholders Ribera et al., (2019).

The study recognizes that existing fraud detection models, while powerful in their predictive capabilities Chen et al., (2021), fall short of supplying adequate explanations for their decisions. The black-

box nature of these models exacerbates the challenges associated with building trust, understanding model behaviour, and ensuring compliance with regulatory requirements. This study focuses on developing an Explainable Fraud Detection System using a machine learning model and an LLM (Large Language Model) wrapper. The study aims to address the limitations of existing fraud detection models that, while powerful in their predictive capabilities, fail to provide adequate explanations for their decisions. The study's primary objective is to integrate machine learning models with explainability methods and LLMs to produce suitable explanations about why a transaction is deemed fraudulent to an end user.

LITERATURE REVIEW

Kurshan et al., (2020) present an enhanced threshold selection policy to improve fraud alert systems in retail banking. Current fraud detection systems face operational inefficiencies caused by fixed thresholds, leading to inaccurate alerts, missed fraud cases, and customer dissatisfaction. The authors formulate the problem as a sequential decision-making process and apply a Deep Q network-based reinforcement learning approach. They suggest dynamic thresholds that adapt hourly to replace static thresholds. Their method includes a state, action, and reward cycle which includes an hour of the day, fraud losses, capacity utilization, and score threshold. Actions are selecting one of several score thresholds. Rewards are based on maximizing fraud savings while minimizing losses.

Olaoye et al., (2024) investigates the integration of (ML) with behavioural analytics to bolster fraud detection for the fintech industry. The paper delves into supervised and unsupervised ML algorithms and the significance of behavioural analytics for identifying fraudulent patterns within vast datasets and real-time user interactions.

Olaoye et al., (2024) state, that ML can identify complex relationships and anomalies, which helps it adapt to evolving fraud tactics. But by pairing ML with behavioural analytics, deviations from normal user patterns are monitored & used to uncover potential fraud. Combining ML and behavioural analytics should build adaptable systems, reducing false positives. Furthermore, it examines real-world successes where this tech mix has improved fraud detection.

Fan et al., (2019) discuss Ghana's Struggle with Cyber Fraud and its Impact on Online Shopping. The study attributes low online shopping participation (39% of the interviewed populace) to two key factors: Prevalence of Cybercrime & the Erosion of Trust. The paper suggests strengthening Cybersecurity & for E-commerce businesses to actively build trust by implementing robust security measures to promote safe online transactions to rebuild consumer confidence.

Ijeoma et al., (2013) investigated the impact of fraud management on organizational survival in Nigeria. The researchers aimed to determine if businesses with a comprehensive approach to fraud management reduce fraud and improve overall business performance. Per the results from a questionnaire distributed to 44 staff in small and medium-sized businesses, adopting a holistic approach to fraud management did not lead to improved fraud prevention or performance in Nigerian businesses. However, the study found that effective coordination of fraud management activities did help in reducing fraudulent activities.

Somanchi et al. (2017) investigated how customer behaviour changes after experiencing a fraudulent transaction. Analysing data from a large US bank, the researchers focused on customers who reported unauthorized transactions. Using propensity score matching, they compared customers who experienced fraud with those who did not. The study found that customers were more likely to terminate their banking relationship after a fraudulent transaction, especially if unauthorized. However, they were less likely to churn when fraud was attributed to a specific third party. Customers denied fraud claims by the bank were also more likely to churn. The study highlights the importance of clear communication and addressing customer concerns to maintain loyalty.

METHODOLOGY

This explores the development of an intelligent fraud detection system, focusing on the selection, evaluation, and explanation of machine learning (ML) techniques. It begins by discussing the importance of rigorous performance metrics such as F1-score, precision, and recall in the context of imbalanced datasets like those found in fraud detection. The chapter then evaluates several ML algorithms, selecting a Deep Neural Network (DNN) based on its superior performance. Next, the chapter delves into eXplainable AI (XAI) methods examining their necessity for understanding complex ML models and justifies the choice of SHAP (SHapley Additive exPlanations) due to its theoretical foundation and ability to provide both global and local insights. To make SHAP results accessible to non-technical users, the chapter proposes integrating a Large Language Model (LLM) capable of translating SHAP outputs

into clear narratives. Finally, the chapter outlines a user study designed to evaluate the effectiveness of these LLM-generated explanations. The study will assess metrics such as understandability, satisfaction, and trustworthiness, aiming to demonstrate the system's ability to provide clear insights into its decision-making process. Figure 1 display the process of the proposed study

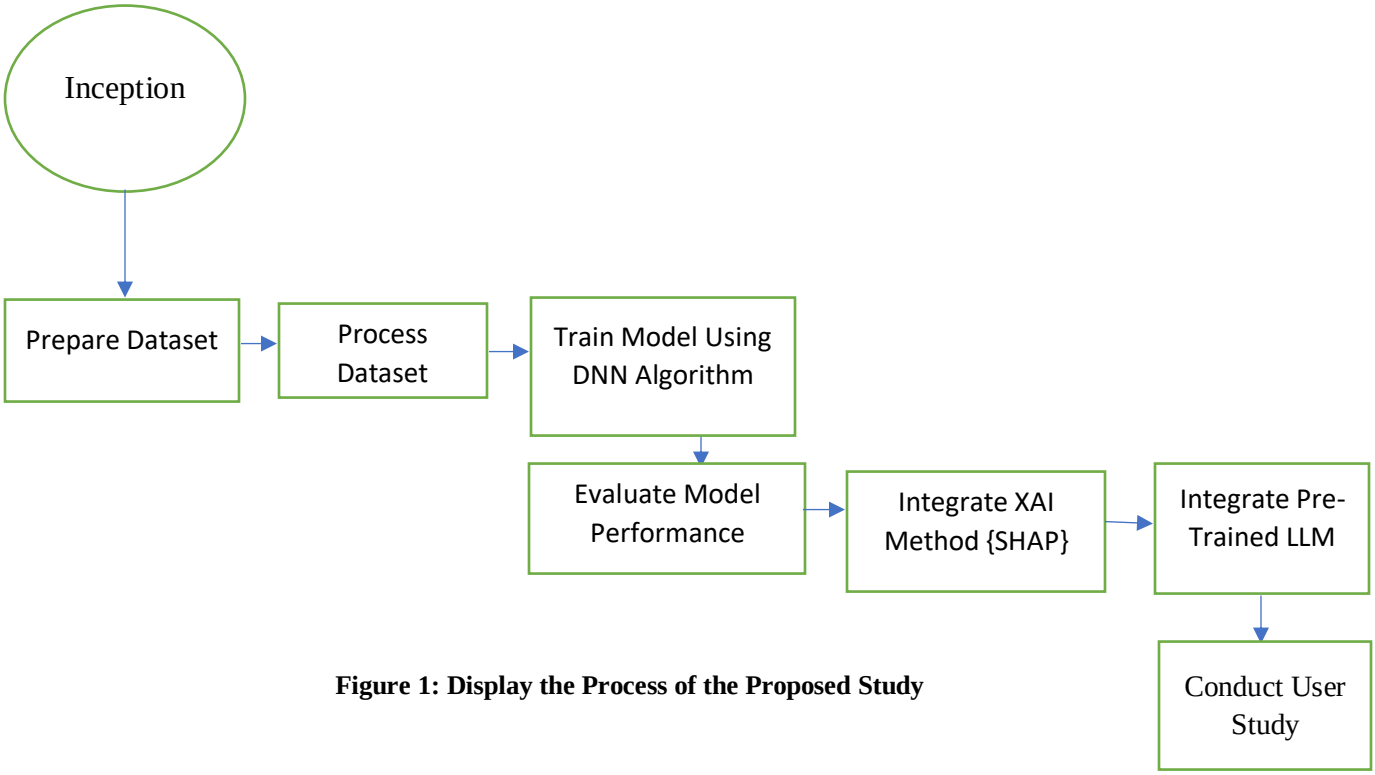


Figure 1: Display the Process of the Proposed Study

Data Processing

Dataset Loading and Trimming

The initial step involved loading the credit card transaction dataset (credit_card_transactions-ibm_v2.csv) and the card information dataset (sd254_cards.csv) into Pandas Data Frames. These datasets were then merged to associate card types with each transaction, ensuring that only credit card transactions were included in the analysis. Following the data preparation, a temporal analysis of fraudulent transactions over multiple years is conducted. This analysis revealed a fluctuating trend in fraud occurrence. To manage computational resources efficiently and to focus on a period with a significant fraud rate, transactions from the year 2008 were selected for further in-depth analysis. Figure 2: Shows the Bar chart of fraudulent transactions changes over the years in the utilized dataset.

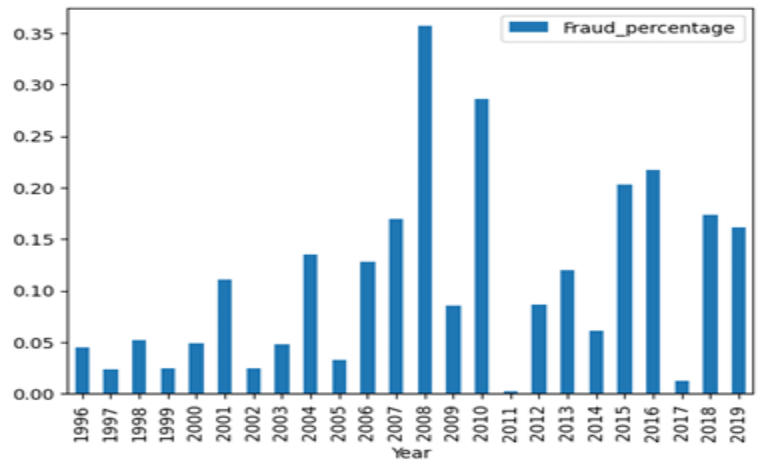


Figure 2: Bar chart of fraudulent transactions changes over the years in the utilized dataset

Data Pre-processing

The pre-processed dataset is loaded from a CSV file for further analysis. During this stage, specific features that were deemed irrelevant, such as Latitude, Longitude, and hist_fraud_trans_24h, were removed from the dataset. This refinement is crucial for ensuring that the dataset is appropriately prepared for subsequent machine learning tasks, particularly focusing on oversampling to address class imbalance, a common issue in fraud detection scenarios. To tackle the class imbalance, the Synthetic Minority Over-Sampling Technique (SMOTE) is applied to the training data. SMOTE generates synthetic samples of the minority class, in this case, fraudulent transactions thereby balancing the class distribution. This technique enhances the model's ability to learn and generalize from patterns of fraudulent behaviour, which is essential for effective fraud detection. Following the application of SMOTE, the dataset is split into training and testing sets using a 70/30 ratio. This division is crucial as it ensures that the model is trained on a substantial portion of the data while being evaluated on a separate, unseen portion. Such a split provides a more reliable assessment of the model's performance and its ability to generalize to new, unseen data.

Model Training

The dataset is split into training and testing sets in a 70/30 ratio, ensuring the model is trained on a substantial portion of the data while reserving a separate portion for evaluation. The deep neural network (DNN) model, incorporating the autoencoder architecture, is trained on the training set and evaluated on the test set. It returned the probability of fraud for each transaction, which is subsequently converted into a binary outcome. The model is trained for 100 epochs with a batch size of 32. To track the model's progress and ensure robustness, checkpoints were saved at each epoch. Tensor Board is employed to visualize training metrics such as loss and accuracy, facilitating better monitoring and interpretation of the training process. To monitor overfitting and assess model generalization, 10% of the training data is designated as a validation set during the training phase. This validation set provided a continuous assessment of the model's performance on unseen data, ensuring it did not merely memorize the training data but generalized well to new, unseen transactions.

SHAP (SHapley Additive exPlanations)

SHAP (SHapley Additive exPlanations) is a model interpretability technique used to elucidate the decision-making process of machine learning models. The central idea of SHAP is to treat features as players in a cooperative game where the "payout" is the model's output. By systematically adding and removing features from various combinations, SHAP assesses the marginal impact of each feature on the prediction. This process yields a Shapley value for each feature, reflecting its importance in determining the model's outcome.

Performance Evaluation Metrics

Performance evaluation metrics are measurable values used to determine the efficiency, effectiveness, and quality of the prediction or defection model. The metrics that will be used in the developed study are as follows:

Accuracy

This is the percentage of test instances that were correctly and incorrectly classified. The correctly classified instances give the accuracy, while the incorrectly classified instances can be computed by

subtracting the correctly classified instances from 100, as shown in Equation (1).

$$\textbf{Accuracy} = \frac{\textbf{TP+TN}}{\textbf{TP+FP+TN+FN}} \times \textbf{100}\% \quad \text{(equa…1)}$$

Where TP = True positive, FP = False positive, TN = True negative, FN = False negative

Recall

The proportion of actual positives (e.g., people with a disease) that were correctly identified. It is also referred to as the true positive rate or recall, as shown in Equation (2).

$$\textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP+FN}} \times \textbf{100}\% \quad \text{(equa….2)}$$

Precision

The ratio of correctly predicted positive observations to the total predicted positive observations. It is computed by dividing the number of true positives by the sum of true positives and false positives, as shown in Equation 3.

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP+FP}} \times \textbf{100}\% \quad \text{(equa….3)}$$

F1-Measure

The harmonic mean of precision and recall. It provides a balance between precision and recall, especially when both false positives and false negatives are important. The formula is shown in Equation 4.

$$\textbf{F1} = \textbf{2} * \frac{\textbf{precision*Recall}}{\textbf{Precision +Recall}} \quad \text{(equa…4)}$$

RESULT

The experimental setup was simulated in the Python environment. Python boasts a rich collection of libraries specifically designed for data science tasks. Libraries like NumPy for numerical operations, Pandas for data manipulation and analysis, Matplotlib and Seaborn for data visualization, and Scikit-learn for machine learning algorithms provide powerful tools that streamline complex tasks. The results were provided using the following subsections.

Performance metrics of Credit card fraud Detection Models

This section provides results for the credit card fraud detection model. Also, the results were evaluated on different metrics which as shown in Table 1

Table 1: Performance metrics of Credit card fraud Detection Models

Algorithm	Accuracy	Precision	Recall	F1-Score
DNN	0.9684	0.950	0.989	0.969
RF	0.9642	0.982	0.946	0.965
LGBMC	0.9995	0.957	0.757	0.845
LR (balance Data)	0.9723	0.900	0.060	0.120
LR (unbalance Data)	0.9991	0.570	0.930	0.710
Autoencoder [*] =5	0.9870	0.640	0.011	0.190

Table 1: shows the result of the six algorithms implemented on fraud detection datasets. Algorithms 1 through 8 have good accuracy. The accuracy of LGBMC is the highest but has poor recall against RF and DNN. RF is the most precise, however across all models, the F1-score for DNN was the highest. Although LR is an option, it performs poorly against DNN, but furthermore Chen et al., (2021) points to LRs' vulnerability to over-fitting and RF being able to handle only small volumes of data which isn't ideal for an application in financial fraud detection.

CONCLUSION

This paper presents the design and implementation of an explainable AI (XAI) system for detecting credit card fraud using machine learning algorithms and interpretability techniques. The key objectives were to investigate ML algorithms and XAI methods for fraud detection, design and implement an XAI system using an ML model, XAI methods, and large language models (LLMs), and evaluate the developed system. The study demonstrated the effectiveness of the DNN model in detecting fraudulent transactions and the ability of SHAP analysis to provide insights into the most influential features for fraud detection. The LLM successfully generated narrative explanations from SHAP outputs, translating technical data into understandable text. User feedback collected through the web application

showed a strong preference for narrative explanations over SHAP plots or no explanation, with 30 instances preferring narratives compared to 2 instances each for SHAP plots and neither option

RECOMMENDATION

On the outcomes & limitations of this work, here are some future research opportunities:

Temporal Adaptation: Future research could explore how the model and explanations can be dynamically adjusted to accommodate evolving fraud patterns over time, ensuring sustained effectiveness and relevance.

Balancing Explainability and Performance: Investigating the trade-offs between explainability and predictive performance could yield insights into how to optimize both aspects without significantly compromising either.

REFRENCES

Aboze, B. J. (2024, March 30). A Comprehensive Guide into SHAP (SHapley Additive exPlanations) Values | Deepchecks. Deepchecks. https://deepchecks.com/a comprehensive-guide-into-shap-shapley-additive explanations-values

Ali, S., Abuhmed, T., El–Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion, 99, 101805. https://doi.org/10.1016/j.inffus.2023.101805

Al-Shabi, M. (2019). Credit card fraud detection using autoencoder model in unbalanced datasets. Journal of Advances in Mathematics and Computer Science, 1–16. https://doi.org/10.9734/jamcs/2019/v33i530192

Chen, J. I., & Lai, K. (2021b). Deep Convolution Neural Network Model for Credit-Card Fraud Detection and Alert. Journal of Artificial Intelligence and Copsule Networks, 3(2), 101–112. https://doi.org/10.36548/jaicn.2021.2.003

Cirqueira, D., Helfert, M., & Bezbradica, M. (2021). Towards design principles for User-Centric Explainable AI in fraud detection. In Lecture notes in computer science (pp. 21–40). https://doi.org/10.1007/978-3-030-77772-2_2

Göbel, K., Niessen, C., Seufert, S., & Schmid, U. (2022). Explanatory machine learning for justified trust in human-AI collaboration: Experiments on file deletion recommendations. Frontiers in Artificial Intelligence, 5.

Goebel, R., Chander, A., Holzinger, K., Lécué, F., Akata, Z., Stumpf, S., Kieseberg, P., & Holzinger, A. (2018). Explainable AI: the new 42? In Lecture notes in computer science (pp. 295–303). https://doi.org/10.1007/978-3-319-99740-7_21