

Article

Human-Centered Explainable AI for Improving Cybersecurity Fraud Investigation and Decision-Making

Jafrin Reza¹, Abdul Rahman Mohammad², Mst Murshida Khatun³, Fahmida Akter⁴

¹ Master of Science in Business Analytics, Trine University, USA

² Bachelor of Technology in Computer Science, Jawaharlal Nehru Technological University Hyderabad (JNTUH), India

³ Master of Business Administration, San Francisco Bay University, USA

⁴ Master of Science in Information Technology, Washington University of Science and Technology, USA

Abstract: Digital payment systems have come a long way, resulting in a challenge for cybersecurity experts and financial institutions to identify and trace fraudulent credit card transactions. While traditional machine-learning fraud detection systems can deliver high predictive accuracy, they tend to be black-box systems which offer little insight into the reasons behind a fraudulent transaction being classified as such. Lack of transparency can decrease investigator confidence, make the case more difficult to evaluate, and affect the utility of AI in fraud decision making. This study suggests a Human-Centered Explainable Artificial Intelligence (XAI) framework to enhance the process of cybersecurity fraud investigation and decision-making by leveraging a U.S.-based credit-card transaction dataset including transaction, merchant, geographic, temporal and customer-related attributes. The proposed framework combines machine-learning-based classification with explainable AI (XAI) techniques to translate fraud predictions into explainable and actionable evidence in investigations. For comparative fraud classification, Logistic Regression, Random Forest, XGBoost, and LightGBM are considered, and the performance of the models is measured by the precision, recall, F1-score, ROC-AUC, PR-AUC, confusion matrix, false positive rate and false negative rate. In the study, instead of focusing on accuracy only, class imbalance and cost-sensitive learning methods are discussed as they are appropriate. The study does not only focus on accuracy, but on appropriate class-imbalance and cost-sensitive learning methods, since fraudulent transactions make up a minority class. Incorporation of SHAP and LIME to understand the factors that affect fraud predictions, both globally and per transaction. Explanations generated are presented in a way that works for investigators by grouping them into categories of evidence: transaction amount, merchant category, temporal patterns, geographic patterns and customer related attributes, which allows for risk-based interpretation of each alert. The framework also provides a human centered decision layer that translates model outputs and model explanations into risk levels for fraud and priorities for investigations. The proposed approach aims to overcome the drawbacks of opaque fraud detection systems by leveraging interpretable evidence and human judgment to enhance transparency, consistency, and clarity in cybersecurity decision-making. This study offers a practical approach to combining predictive accuracy, explainability, and human oversight in credit-card fraud investigations and lays the groundwork for building reliable AI-powered fraud decision support systems in the financial contexts of the United States.

Keywords: Explainable Artificial Intelligence, Cybersecurity Fraud Detection, Human-Centered AI, Credit Card Fraud, Fraud Investigation and Decision Support



This is an open-access article under the [CC-BY 4.0](https://creativecommons.org/licenses/by/4.0/) license

1. Introduction

Background

Financial institutions handle a rising volume and complexity of transactions with credit cards, thanks to the growing digital financial services, electronic payment and online commerce. In tandem with these advances, fraudulent operations have evolved and intensified, posing significant risks to both financial and operational integrity and cyber security. Fraudulent transactions can lead to direct monetary loss, unhappy customers, damage to reputation and higher costs of investigation and recovery [1]. Financial institutions have become more eager to use machine-learning methods to detect unusual transactions and focus their investigation on possible fraud [2]. Machine-learning models can sift through a myriad of transactional information, uncover nonlinear relationships and behavior that may be hard to recognize using traditional rule-based systems. While these are useful, traditional machine-learning fraud detection systems tend to focus on predictive accuracy rather than on understanding how the decisions are made, and therefore do not necessarily yield high hit rates for human investigators. While more complex models like ensemble learning algorithms and gradient boosting can make more accurate predictions, they can also be less transparent in understanding what is driving other fraud alerts at an individual level [3]. The problem is that this is a major challenge in cybersecurity environments where investigators have to analyse and validate any suspicious activity, and determine whether any transaction should be approved, reviewed, escalated or even blocked [4]. Explainable Artificial Intelligence (XAI) provides an opportunity to address this transparency challenge by revealing the factors influencing machine-learning predictions [5]. Global and local explanations of model behavior can be obtained by techniques like SHAP and LIME. But, technical information does not necessarily translate into better investigation results. It is therefore very important to have an approach that is human-centered so that the outputs of the model can be translated into evidence that is understandable and can be acted upon [6]. This study aims to examine the ability of human-centric XAI to link fraud prediction to investigator interpretation and decision making by analyzing a U.S.-based dataset of credit-card transactions.

Human-Centered Explainable AI for Fraud Investigation

Human-centric Explainable Artificial Intelligence (XAI) is an approach that makes AI systems not only accurate prediction systems, but also ones which enable human understanding, judgment and decision making [7]. This is an essential way of thinking in a fraud investigation because the end goal is not always to determine that a transaction is fraudulent. Investigators need to know why an alert was triggered, consider supporting evidence, consider the trustworthiness of an alert, and decide on an appropriate response [8]. The properly functioning fraud-detection system would have to be a decision-support tool instead of a stand-alone solution. Typical fraud-detection systems typically produce a binary classification, or a probability score. Such outputs can be used to detect potentially suspicious transactions but only limited information regarding the rationale behind predicting the transaction. Human-centered XAI resolves this issue by developing interpretable representations of how the individual features in a prediction. LIME can provide locally interpretable approximations of complex model decisions, and SHAP can quantify the impact of the transaction characteristics on the overall model behavior and specific predictions [9]. The explanations in the selected credit-card data can include relevant attributes like transaction value, merchant category, transaction time, location, state, geographic coordinates, city population, and customer occupation. These factors can be converted to evidence that is investigator oriented, to help explain why a transaction has a specific risk score [10]. This information can then be sorted into the categories of investigation and the categories of risk within a human-centred framework. This can help to enhance transparency, assist analysts in verifying it and facilitate more uniform decision making. This study's main argument is that, effective fraud AI must be predictive, explainable, and supervised by humans, providing a clear linkage between automated fraud detection and cybersecurity investigation.

Problem Statement

Many models used in detecting credit-card fraud are effective, but state little about the basis for their conclusions. Machine-learning methods are increasingly vital to the problem of detecting

fraudulent credit-card transactions, but many of the best models are opaque, telling little of the reasons they make their predictions [11]. This makes it hard for fraud investigators to confirm alerts, process information and decide what to do. The problem of class imbalance also makes fraud detection more challenging because false negatives could mean that fraudulent transactions go undetected and over-investigations could occur if too many false positives are produced [12]. The techniques of XAI such as SHAP and LIME may be useful to discover the most influential features, but the resulting technical explanations may not be of use to human investigators for action [13]. Thus, there is a need for a human-centric approach to developing a framework that combines fraud prediction with explainable explanations, risk assessment, and investigator decision support. To fill this gap, this study introduces a novel Explainable AI (XAI) methodology to enhance transparency, investigation efficiency, and fraud-related decision-making.

Objectives of the Study

The objectives of this study are to:

1. Develop and evaluate machine-learning models for identifying fraudulent credit-card transactions using the selected U.S.-based dataset.
2. Evaluate four machine learning models (Logistic Regression, Random Forest, XGBoost, LightGBM) on suitable fraud-detection evaluation metrics.
3. Determine the most relevant transaction, merchant, temporal, geographical and customer factors relevant to fraud predictions.
4. Use SHAP and LIME to provide explanations for global and transaction-level prediction of machine learning fraud.
5. Create a framework for explanation from a human perspective that converts the model output into comprehensible and usable evidence for the fraud investigator.
6. Create a mechanism to support investigators in prioritizing transactions for approval, review, investigation, or escalation that is based on risk
7. Evaluate the potential of XAI for enhancing transparency, interpretability and consistency in cybersecurity fraud investigation and decision making.

Research Questions

The following large research questions are being tackled by the study:

1. What are the results of evaluating the ability of machine-learning models to identify fraudulent credit-card transactions in the chosen U.S.-based dataset?
2. Which are the most important transaction, merchant, temporal, geographic and customer-related features to predict fraud?
3. What is the ability of SHAP and LIME to explain the prediction of fraud in an easy-to-understand and actionable way for human investigators?
4. What value can be added by explainable AI that is human-centered, rather than just relying on AI model predictions?

Significance of the Study

This study is important due to a key challenge that artificial intelligence (AI) presents in cybersecurity fraud detection: the difference between predictive performance and human understanding. Machine learning systems are becoming more essential in the financial sector to process and review vast amounts of transaction data, as well as detect possible suspicious activity [14]. But having a very accurate prediction isn't the only thing an investigator needs to know when deciding whether or not to proceed with an alert, and what evidence backs up the alert [15]. This study brings together explainability and human support for decision making to create a framework for connecting automated fraud detection with practical investigation processes. The study brings the following advantages to the academic area: the use of machine-learning fraud classification with an

explanation layer based on SHAP and LIME; and a human-centered interpretation layer. The proposed framework doesn't only focus on explainability as a technical capability, but also on how explainability can facilitate human judgment, investigation prioritization, and risk-based decisions [15]. This outlook can help further studies on trustworthy, transparent, and accountable AI capabilities in cybersecurity and financial analytics. The proposed framework may make fraud alerts more understandable for financial institutions and enable investigators to prioritize their investigations on high-risk transactions [16]. It may also be useful to provide the analysts with the key elements that lead to a prediction, to lessen reliance on ambiguous model output and to boost a sense of trust in the model when dealing with problematic cases [17]. The framework can facilitate more uniform decision making by linking the likelihood of fraud and to more meaningful evidence and investigation priorities [18]. This study also has implications for customers and organizations: Fraud detection should not be too heavy-handed, as it leads to false-positive alerts that create a burden for the customer and the organization. Human centric approach can assist in creating a more clear-cut equilibrium between security and proper transaction processing [19]. This study highlights the potential of explainable AI to shift from prediction to actionable, transparent, and human-centric cybersecurity decision-making.

Detection of credit card fraud using machine learning

The machine learning approach is one that is emerging as crucial for the recognition of fraudulent credit-card transactions, due to its ability to process huge amounts of transactional data, and to detect complex patterns that can be related to fraudulent behaviour [20]. Old fraud detection systems are based on rules that have to be set beforehand, such as detecting a big purchase from an unusual location or a large amount for a small item. While they can still be effective, these systems can be ineffective at identifying new fraud tactics and can result in many false positives [21]. A more adaptive solution to this is the use of a machine-learning model that learns the relationships between transaction attributes and past fraud results. The broad range of supervised classification algorithms that have been widely applied to fraud detection, including Logistic Regression, Decision Trees, Random Forest, Support Vector Machines and gradient-boosting algorithms. Random Forest, XGBoost and LightGBM are ensemble models that are useful since they can model non-linear relationships and interactions between transaction variables. In the U.S. credit-card market, factors like transaction dollar amount, merchant type, transaction timing, geographic location, and customer-related features may be important clues for determining fraud from legitimate transactions [21]. The datasets involving fraud typically have a significant class imbalance as there are many more legitimate transactions than fraudulent ones. Thus, accuracy is not enough to arrive at accurate conclusions [22]. More meaningful evaluations would be provided by precision, recall, F1-score, area under the precision-recall curve (PR-AUC), ROC-AUC, and false-positive and false-negative rates. Imbalance can be overcome using techniques like class weighting, cost-sensitive learning and appropriate oversampling [23]. Even though high-performing machine-learning models are able to predict, they can be hard to interpret, especially in the case of complex ensemble models. This restriction is particularly applicable to fraud investigation as fraud analysts are required to know the reason for the flagged transaction [24]. Interpretability and human decision support along with the predictive performance should be integrated together for the development of reliable fraud-detection systems.

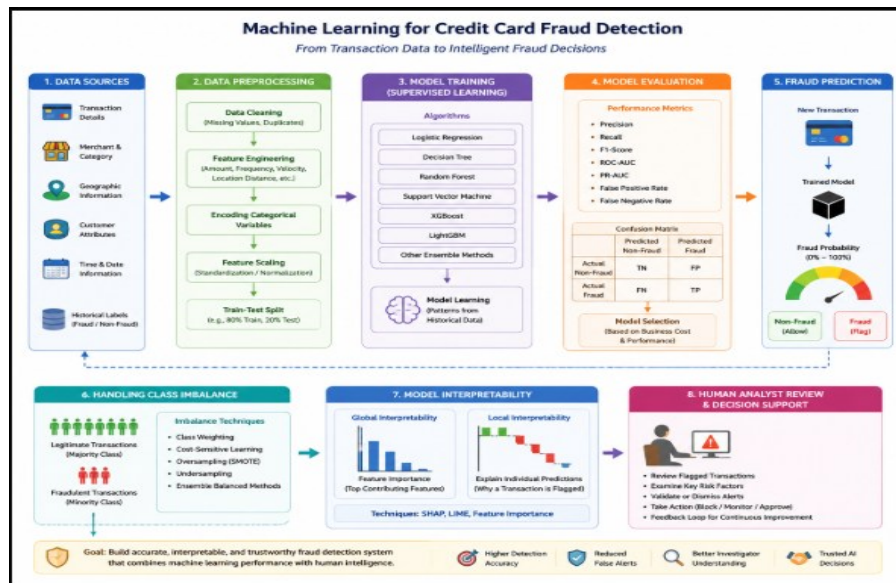


Figure 1. This flowchart represents a credit-card fraud detection and investigation system that is explained using machine learning methods

The flowchart shows an end-to-end machine-learning solution for credit-card fraud detection and investigation. It starts with the transaction data, merchant details, geographic details, customer attributes, time information and historical fraud labels. Before training the supervised learning models, the data is cleaned, feature engineered, encoded (categorical), scaled, and split into train and test sets. Various classifiers like Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, XGBoost, and LightGBM are considered for fraud classification. Precision, recall, F1-score, ROC-AUC, PR-AUC, false-positive rate, false-negative rate and confusion matrices are used to evaluate a model. Class imbalanced issues are solved with the class weighting, cost sensitive, SMOTE and undersampling methods. Explanations of the model provided by SHAP, LIME and feature importance [25]. Last but not least, human analysts go over flagged transactions, look at risk factors, review alerts, and make informed decisions, and then loop back for continual improvement.

Explainable Artificial Intelligence in Fraud Detection

Explainable Artificial Intelligence (XAI) is becoming a significant field of research in order to tackle the lack of transparency in complex machine learning systems. XAI's goal in fraud detection is to help understand why an algorithm deems a specific transaction fraudulent and how specific attributes or features of the transaction might be most influential in predicting fraud [26]. This information is useful since fraud investigators will want evidence to confirm an alert before proceeding to make a decision. This can, therefore, be a bridge between automated prediction and human interpretation by explaining [27]. One of the most popular XAI methods to explain machine-learning predictions is called SHAP (SHapley Additive explanations). SHAP provides a range of contribution values to individual features, and can give global explanations of the behavior of the model, as well as local explanations of individual transactions. SHAP can determine if the transaction amount, merchant category, timing, location, or customer characteristic factors are causing fraud to be more or less likely than expected for fraud investigation [28]. LIME (Local Interpretable Model-Agnostic Explanations) is another model-agnostic approach that builds an interpretable approximation to a prediction made by a given model to a particular instance. Complementary evaluation of explanation consistency and local interpretability can be achieved by employing both techniques. But technical explainability isn't necessarily useful in practice [29]. The feature importance ranking could be mathematically meaningful, but not easy for a non-technical investigator to understand. Furthermore, explanations may get entangled when there are multiple variables and/or correlated variables causing simultaneous effects [30]. Thus, there is a need to carefully present the

explanations, in a way that matches investigative reasoning, if effective XAI is to be achieved. Thus, beyond generating explanations, there is a growing focus on XAI research on transparency, trust, accountability and human oversight. For fraud detection, the explanation should focus on the key evidence, explain how it is affecting the transaction, and help an investigator determine if the transaction should be investigated further [31]. This work is further inspired by this view, and it proposes to integrate SHAP and LIME into a human-centered framework for fraud investigation.

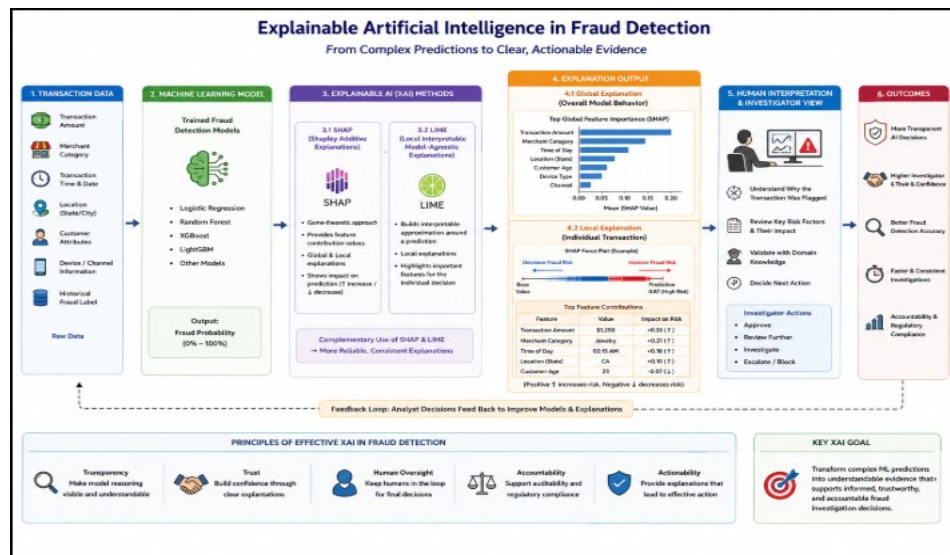


Figure 2. This flowchart illustrates an explainable AI framework for making decisions on fraud investigations in a human-centric and transparent manner

The flowchart is a representation of an explainable artificial intelligence (XAI) system for enhancing transparency and decision-making within fraud detection processes. It starts with transaction data such as the amount of the transaction, merchant type, time, location, customer data, device data and historical fraud labels. The machine-learning models used in this process are Logistic Regression, Random Forest, XGBoost and LightGBM, which calculate fraud probabilities [32]. Finally, global important features are identified by SHAP and individual predictions made for the transactions are explained by LIME. The explanation output is used to identify factors that make a transaction more or less risky and shows their impact on specific transactions [33]. Human Investigators review flagged transactions, review the key risks, validate alerts based on domain knowledge, and decide on the proper actions [34]. Transparency, trust, human oversight, responsibility and actionability all help to build a strong case for responsible fraud investigation while a feedback loop helps to improve models constantly.

Human-Centered AI for Fraud Investigation and Decision-Making

In a human-centred Artificial Intelligence, the design of AI systems is centred on human needs, capabilities, judgment and accountability [35]. This is especially important in cyber security fraud investigation as alerts typically come out of automated systems and need to be reviewed and acted on by human analysts. Human-centered AI is not intended to replace investigators, but rather to enhance their ability to manage complex information, prioritize cases and make consistent decisions [36]. During a fraud investigation, it's possible that human decision makers will receive thousands of alerts, some of which could be legitimate transactions. Too many 'false' alerts can lead to 'alert fatigue' and make investigations more ineffective; too few can mean fraud can go on in the background [37]. To help tackle this challenge, a human-centric decision-support system can incorporate fraud probabilities and easily explainable evidence and risk-based prioritization. For instance, an AI system could pick out a transaction as being a high risk and give the investigator the key factors that formed its prediction so that the analyst could decide if the evidence merits escalation. Another key element

of human-centred AI is trust. Investigators will be more inclined to trust and adopt AI recommendations when they can comprehend, ask questions about, and challenge the reasons behind the AI's recommendations [38]. But, over-dependence on automated recommendations can lead to automation bias and critical human oversights are necessary [39]. Therefore, good systems should provide explanations in a way which makes them seem like information rather than a given. There has been existing research that has laid foundations for machine-learning fraud detection and XAI, and a fundamental divide between the technical explanation and the real-world investigative decision-making process [40]. Much of the research concentrates on predicting performance or on feature importance, but neglects to consider how explanations can be turned into an investigative body of evidence that can be acted upon. In this study, we tackle this gap by presenting a framework that combines machine-learning predictions with human-understandable explanations thereof based on SHAP, and by translating them into human-understandable risk evidence and human investigation priorities. This integration can enhance transparency, aid analyst decision-making and help achieve more accurate AI-driven fraud investigation.

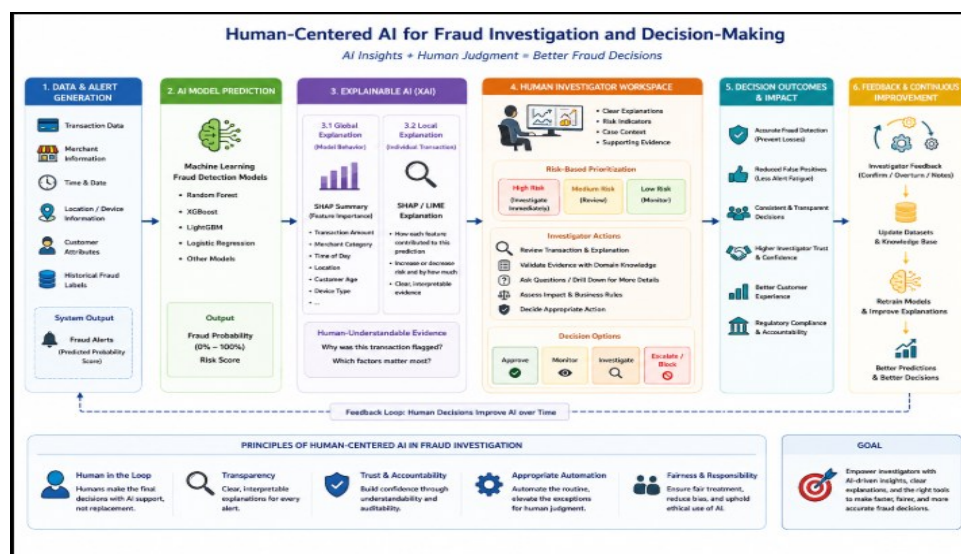


Figure 3. This flowchart displays a human-centered AI framework for explainable fraud investigation and decision-making

This flowchart shows an AI-based approach to fraud investigation and decision making in a human-centric way. It starts with transaction, merchant, temporal, location, device, customer and historical fraud data to create fraud alerts and probability scores. These inputs are then fed into machine-learning models such as Random Forest, XGBoost, LightGBM and Logistic Regression, which provide a prediction of the risks of fraud. Explainable AI techniques, in particular, SHAP and LIME, give explanations globally, and locally, that explain why, for particular predictions, certain factors were important [41]. These explanations are then turned into proof that is understood by humans to be used by investigators. Each transaction is analysed, evidence is checked against the domain knowledge, risk is evaluated and decisions are made on whether to approve, monitor, investigate or escalate/block the transaction [42]. Decision outcomes enable fraud prevention, minimize false positive alerts, provide transparency, build trust and ensure compliance. A feedback loop includes investigators' choices to enhance datasets, models, explanations, and predictions.

Empirical Study

In their article "Explainable AI in Insider Financial Fraud Detection Models: A Review of Transparency and Trust", Ofori et al. (2026) explored the role of AI in insider financial fraud detection and the need for explainability in this context. The authors analysed 107 empirically tested studies published from 2015 to 2025 that included machine-learning and deep-learning models, approaches

for XAI, evaluation procedures, and data characteristics. Their research suggests that machine-learning systems can be used to enhance automated fraud and insider-threat detection, but using too much machine-learning based modeling without transparency can reduce trust in the analysts, make forensic investigation difficult, and cause difficulty in regulatory accountability [1]. The review lists attention mechanisms and hybrid temporal models as emerging applications of post hoc explainability techniques, and highlights SHAP and LIME as important post hoc explainability techniques. Importantly, Ofori et al. suggest a Three-Pillar Framework for the explainability of AI systems, including Algorithmic Transparency, Evaluation Accountability, and Data Traceability, which places explainability at the core of system design, not as an afterthought. Another area of importance identified by the authors as research gaps is the limited use of temporal and behavioral modeling, as well as the narrow evaluation practices. These findings are robust and support the present study by illustrating the need for linking model implications to evidence that is comprehensible, for which analysts should have some means of accountability and cognizance, and for which humans must make decisions. To this, the current research enhances the above by incorporating XAI at the human-centered credit-card fraud investigation framework.

In the article “Explainable AI: Enhancing Decision-Making in the Detection of Cyber Threats,” Prasad et al. (2026) explore how to increase transparency, trust, and practitioner decision-making in cybersecurity through the lens of Explainable Artificial Intelligence. The authors review 53 peer-reviewed papers published between 2018 and 2025, focusing on the applications of XAI in cybersecurity, especially anomaly detection, threat identification and interpretable AI systems. Their study reveals that machine-learning and deep-learning models have the potential to be effective in detection, but can be opaque, making it difficult for cybersecurity professionals to validate and trust prediction capabilities [2]. The authors highlight the potential of XAI techniques to improve the ability of analysts to understand and act on complex model outputs, aiding in effective incident response and decision-making. Importantly, the review shows that there is still a gap between the technical development of XAI and the deployment in the cybersecurity arena. Much of the previous research uses benchmark data sets and performance assessment indices related to technical considerations, and very little work investigates the ability of explanations to assist real-world analysts under operational conditions. The authors also point to some of the compromise considerations involved in embedding explainability mechanisms, such as between prediction accuracy and computation efficiency, and real-time scalability. The results of this research align well with the current study, as it reinforces the significance of tying the predictions of AI, the evidence that can be understood, and human judgment together. The present study builds on this by combining machine-learning predictions with human-centered decision support and SHAP/LIME explanations to tackle the credit-card fraud investigation.

In the chapter “Human-Centered Artificial Intelligence in Cybersecurity” (Chapter 1), Chen et al. discuss the potential of human-centered AI to create trusted and successful cyber security systems. The authors believe traditional AI-based cybersecurity solutions may fail to account for all factors, especially regarding the human element, which can be a key factor in successful cybersecurity outcomes. They conclude that human-centred AI should complement human capabilities, which involves bringing in meaningfully human-controlled, transparent, ethical and effective human-AI teaming, and decision support. It outlines a number of challenges that comes along with AI in cybersecurity, such as data quality, algorithmic bias, privacy, explainability, transparency, and trust calibration [3]. The authors also stress the need for cybersecurity professionals to have clear information on AI recommendations to be able to check outputs, have situational awareness, and take proper decisions. They discuss the significance of explainable AI in making machine-learning systems more transparent and trustworthy. More importantly, the chapter introduces a more human-centered approach, where technical performance needs to be integrated with the needs of the users, the training, the accountability as well as appropriate human oversight. These findings directly support the present research since they provide the theoretical background for the use of AI as a decision-support tool as opposed to an autonomous investigator. On this basis, this study proposes to combine the machine learning prediction results and SHAP/LIME explanation mechanism, risk prioritization,

and human judgment to apply human-centered principles to credit-card fraud investigation.

2. Materials and Methods

This study is a quantitative machine-learning based approach that examines the ability of Explainable Artificial Intelligence (XAI) to enhance credit-card fraud investigation and decision-making from a human perspective. The approach combines data preparation, exploratory analysis, feature engineering, handling class imbalance, machine-learning classification, model evaluation, explainability and human-centered decision support [43]. The chosen Credit Card Fraud Detection for Machine Learning data consists of credit-card transactions in the western United States, with attributes of the transactions, merchants, geographic location, time, customers, and fraud [44]. The methodological framework is designed to be a series of steps aimed at transitioning from transaction data to prediction of fraud, explanation, risk interpretation, and decision making to assist investigators.

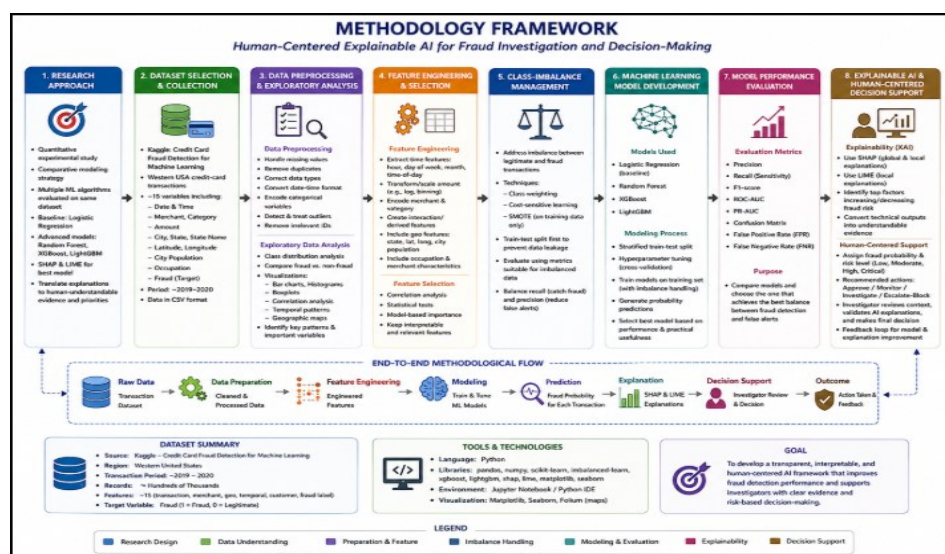


Figure 4. This flowchart represents the full methodology for explainable and human-centric credit-card fraud investigation and decision-making

This flowchart illustrates the full methodology framework of the proposed Human-Centered Explainable AI approach for the investigation and decision-making process of credit-card fraud. It begins with the quantitative research approach and selection of the U.S.-based credit-card fraud dataset. Data Preprocessing and Exploratory Analysis include handling missing data, removing duplicates, correcting data types, distribution of data, and patterns of fraud [45]. The process of feature engineering generates time, transaction, geographic features, merchant features, and customer features, while maintaining interpretability [46]. Class-imbalance management weights the classes, uses cost sensitive learning and SMOTE on training data. Four machine learning models are built and assessed based on accuracy, precision, recall, F1 score, and ROC-AUC, PR-AUC, and error rates. Precision, recall, F1 score, ROC-AUC, PR-AUC, and error rates are evaluated for four machine learning models such as logistic regression, random forest, XGBoost, and LightGBM. Explanation is provided by SHAP and LIME leading to the provision of evidence and investigator decisions based on risk, which in turn should lead to feedback for ongoing improvement.

Research Approach

This study employs a quantitative experimental research approach to investigate the effectiveness of machine-learning and Explainable Artificial Intelligence (XAI) techniques for credit-card fraud detection, investigation, and decision-making [47]. The quantitative approach is suitable since the study involves measurable data from the transactions and there are objective performance

measures to assess fraud-detection models. A comparative modelling approach is used where a number of supervised classification methods are trained, optimized and tested under the same experimental conditions [48]. Logistic Regression is used as a baseline model to be implemented as an interpretable model for comparison with more complex models. The random forest method is used to detect the nonlinear relationships and interactions among transaction characteristics, and gradient-boosting methods such as XGBoost and LightGBM are used as powerful methods to uncover complex fraud patterns [49]. The following measures are used to evaluate the model performance: precision, recall, F1-Score, ROC-AUC, and PR-AUC, confusion matrix, false positive rate and false negative rate with a focus on the detection of minority classes [50]. After the model comparison, the model that has the best performance will be chosen to perform detailed Explainability Analysis by applying SHAP and LIME. Global and local explanations are produced, to determine the factors about which predictions of fraud are made [51]. Then, these technical explanations are translated into evidence understood by humans, fraud-risk levels, and the priorities for investigation [52]. The research thus moves from prediction to explanation, ultimately to human-in-the-loop decision support to enhance the capabilities of the investigators.

Dataset Selection and Data Collection

The data set for the study is Kaggle's Credit Card Fraud Detection for Machine Learning. It contains credit-card transactions made in the western United States and offers a large level of transactions to build and test credit-card fraud-detection models [53]. The data has about 15 variables of the transactions, merchants, geographic, temporal and customer characteristics. They have key attributes such as transaction date and time, merchant, transaction category, transaction amount, city, state, state name, latitude, longitude, city population, customer occupation and the fraud indicator [54]. The transaction records span from around 2019 to 2020, offering historical insights for trends in fraudulent and legitimate credit-card transactions. The dataset is chosen because the variables are somewhat understandable and it can be used for predictive modeling and explanation to humans [55]. The data selected have meaningful contextual information that can be conveyed to fraud investigators, whereas datasets with only anonymous mathematical elements would not. Possible explanations for individual fraud predictions; factors that could provide meaningful explanations, such as the amount of money in the transaction, the merchant category, the time of the transaction, the geographic area, and the customer's occupation [56]. The data is retrieved in CSV format and uploaded into the analytical space. After collection, the data set is methodically examined to see how the data is organized, what type of data it contains, the number of times each data class appears, and the quality of the data. The fraud attribute will be the target attribute and the relevant transaction and contextual attributes will be the predictive attributes [57]. The resulting dataset is then used to preprocess the data, engineer the features, build machine-learning models, analyze the XAI, and investigate the frauds from the human side.

Data Pre-processing and Exploratory Analysis

Data quality, consistency, and analytical reliability are explored and verified in the raw data prior to building the machine learning models. The first step in pre-processing is to look at the data for any missing values, duplicate records, inconsistent data types, invalid observations, and any unusual numerical values [58]. The date and time of transactions are captured and then transformed into standard datetime data, for both temporal analysis and feature extraction. Some of the categorical variables like merchant, transaction category, state, occupation etc are transformed into machine readable variables using suitable encoding techniques. Extreme values, skewness and potential outliers are compared for numerical variables like transaction amount, latitude, longitude, and city population [59]. Where appropriate, variables that serve only as identifiers and do not lead to meaningful prediction information are omitted. After preprocessing, the next step is to perform Exploratory Data Analysis (EDA) to detect significant characteristics and patterns from fraudulent and legitimate transactions. Class imbalance is examined for the distribution of the fraud target [60]. Transaction amounts are compared between fraud classes using descriptive statistics and visualizations [1]. Merchant categories, transaction times, states, geographic location, population of

cities, and customer occupations are analysed and differences that could be seen between fraudulent and legitimate transactions are identified [2]. Relationships among variables are explored through bar charts, histograms, boxplots, correlation analysis, temporal visualisations and geographic plots [3]. The results of EDA feed into feature engineering and model building, and also help to identify variables that can be useful for meaningful explanations using SHAP and LIME. This guarantees that the predictive accuracy and interpretability of the processing are ensured.

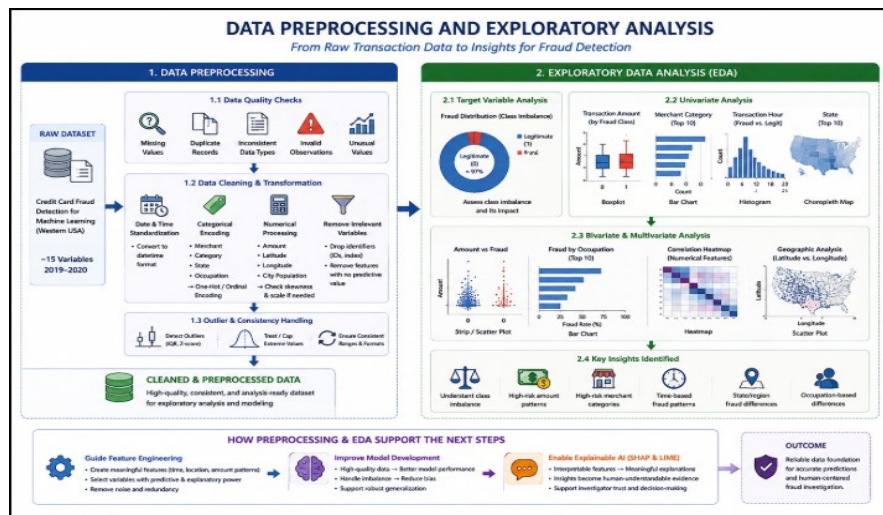


Figure 5. This flowchart illustrates the data preprocessing and exploratory analysis in order to ensure reliable and interpretable credit-card fraud detection

The flowchart shows the data preprocessing and exploratory analysis phases of the credit-card fraud detection study. It starts with the raw credit card transactions data set that was collected in the United States and conducts data-quality analysis for data gaps, duplicates, inconsistent data types, out-of-range values, and unusual values [4]. Data Cleaning and Transformation then normalize date and time data and encode categorical data, transform numerical attributes, and eliminate irrelevant identifiers. Next, the outliers and inconsistent values are studied and a clean data set is created for subsequent analysis. Exploratory Data Analysis examines distribution of the fraud class, explores transaction amounts, merchant categories, transaction hours, states, occupations, correlations, and geographic patterns employing boxplots, bar chart, histogram, heatmap, and maps [5]. The found patterns inform feature engineering and model building and also help to convey interpretable features for SHAP and LIME explanations.

Feature Engineering and Feature Selection

Feature engineering is done to convert raw transaction attributes into interpretable but important variables that can capture patterns that are associated with fraudulent activity [6]. Original transaction time is broken down into transaction hour, day of week, month and time-of-day categories to detect temporal irregularities. The amount of transactions is checked for skewness and outliers and potential transformations are explored to maximize model accuracy while preserving fraud signals [7]. Appropriate categorical encoding is used to transform merchant and transaction categories into a machine-readable format. Geographic information such as state, latitude and longitude is kept to capture potential location related fraud patterns, and city population data is kept to capture possible fraud patterns from the population of the specified city [8]. The merchant characteristics and the customer occupation are also added as they may contain some contextual details that can be used to determine the abnormal transactions. For a few merchants, relationships between amount, timing, merchant category and location are represented as interaction or derived features where appropriate. The feature selection is done by correlation analysis, statistical analysis, model based importance and domain relevance [9]. Some variables are removed to exclude model complexity, if they are highly redundant or irrelevant. Most important of all, feature selection emphasizes the semantic

interpretability [10]. The features should be understandable to investigators, to allow for the translation from SHAP and LIME outputs to meaningful evidence [11]. This can help to enhance the predictive power of the model, while also advancing the human-centric goal of the study, which is to enhance the understanding of the system.

Class-Imbalance Management

A credit card fraud dataset generally tends to have a lot more legitimate transactions than fraudulent transactions, which results in a significant class-imbalance problem. If not corrected, the machine learning model could end up with a high overall accuracy just by guessing that most transactions are good and missing out on fraud cases [12]. This study explicitly considers the class-imbalance management in the modeling process. Various approaches are considered such as class weighting, cost-sensitive learning and Synthetic Minority Over-sampling Technique (SMOTE). Class weighting gives extra weight to fraudulent transactions when training the model, to motivate the algorithms to pick up on the minority class [13]. Cost-sensitive learning also captures the varying penalties that can be expected from errors in fraud detection—higher penalties for more important misclassifications [14]. Another evaluation method for SMOTE is to compare it as a comparative approach to generate observations for the minority class. After the train-test split, any oversampling method is only used on the training set, to ensure that no data is leaked into the test set [15]. The test dataset is kept intact and has the same class distribution as the original dataset, allowing for the realistic evaluation of the results. Precision, recall, F1-score, PR-AUC, false-positive rate and false-negative rate are used to evaluate the effectiveness of each imbalance strategy [16]. Cost-sensitive learning is particularly important for the research since overfiring alerts can result in investigator overload and alert fatigue, while under firing alerts can result in financial loss and cybersecurity issues [17]. The aim of imbalance management is to achieve a balance between effective fraud detection and investigation requirements.

Machine-Learning Model Development

The four supervised classification models utilized for fraudulent credit-card detection are Logistic Regression, Random Forest, XGBoost and LightGBM. Because of its relatively simple mathematical form and interpretability, Logistic Regression is chosen as the baseline. It provides a standard that can be used to evaluate more complicated algorithms [18]. Random Forest is added due to the fact that the ensemble construction allows for capturing of nonlinear interactions and relations among transaction characteristics. These gradient boosting algorithms are chosen as they are advanced and can handle complex patterns in transactional data, specifically XGBoost and LightGBM. Stratified sampling is used to ensure the distribution of legitimate and fraudulent transactions among the training and testing subsets of the processed data set. In case of hyperparameter optimization and model selection, cross validation is performed within the training data set as appropriate [19]. Class-imbalance techniques are used during the training process only to avoid information leakage. Consistent experiments are conducted for all algorithms and consistent test sets are used for all algorithms' evaluations [20]. The final test results are not used to optimize the hyperparameters – rather, suitable validation procedures are used [21]. The models not only predict the classes, but also the probability of fraud which can be used for further risk scoring and explainability analysis [22]. Not only is the overall discrimination of the final predictive model important, but it also should be designed to achieve a balance between the number of fraudulent transactions captured and the number of false-positive transactions [23]. The chosen model is then subjected to detailed SHAP and LIME analysis to lay the groundwork for the human-centered investigation part of the proposed framework.

Model Performance Evaluation

Evaluated developed machine-learning models on metrics especially dedicated for imbalanced fraud classification. The main evaluation metrics are precision, recall, F1 Score, ROC-AUC and PR-AUC. Confusion matrices are created to find true positives, true negatives, false positives and false negatives [24]. Additionally, false positive rate (FPR) and false negative rate (FNR) are computed for assessing the implications of model errors in practice. Recall is especially significant as it represents

the percentage of fraudulent transactions that the model is able to successfully catch [25]. If the recall has a low value then it can mean that a lot of fraudulent activity is escaping detection. Important for unnecessary investigation alerts, precision is the number of transactions that are classified as fraudulent that are truly fraudulent. F1-score gives an overall judgement on precision and recall. PR-AUC is highlighted because it gives useful information on the problem of detecting minority classes in the case of a large class imbalance, whereas ROC-AUC assesses the overall performance of the model in discriminating between fraudulent and legitimate transactions [26]. The confusion matrices are also analyzed to find out the kind of errors produced by each model. The performance of the models is evaluated and compared among the different algorithms and imbalance-handling strategies to come up with the best predictive approach [27]. The best model isn't necessarily the one that is the most accurate, but rather the one that performs well and has an investigative value [28]. This guarantees that the chosen model can adequately contribute to the following stages of explainability and human-centered decision support.

Explainable AI Using SHAP and LIME

Two complementary Explainable Artificial Intelligence (XAI) methods are used: SHapley Additive explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) to interpret the selected machine-learning model [29]. At the global level, SHAP is applied on a global scale, whereas at the local level, it is applied on a local scale. Global SHAP analysis highlights the most significant features across all examples of fraud predictions and pinpoints which transaction, merchant, temporal, geographic and customer attributes are most critical. Local SHAP analysis provides insights into the factors influencing the prediction of individual transactions [30]. The predominant use of LIME is for transaction level interpretation. It creates an approximation of a complex machine learning model which can be understood locally and detects features that contribute to a certain fraud prediction [31]. The analyses from SHAP and LIME are compared to determine if important explanatory factors are reasonably consistent throughout explanation techniques [32]. The technical outputs are being converted to evidence for investigators. The technical output is then being presented as evidence to investigators [33]. Rather than only providing a numerical value for contribution, explanations can provide information on how a transaction's likely fraud risk is affected by things like an unusually high transaction amount, a specific merchant category, unusual transaction timing or geographic characteristics [34]. This translation is important for research purposes as it should not only expose the complexity of the model but it should be understandable by the people who review fraud alerts. SHAP and LIME can help bridge the gap between the predictions of machine learning and the interpretation of humans to ensure transparency, trust, and evidence-based fraud investigation.

Human-Centered Fraud Investigation and Decision Support

The final methodological step brings together predictions by the machine learning algorithms and the explanations provided by XAI into a framework of fraud-investigation and decision-support that will be human centered. Each transaction gets a predicted fraud probability, and SHAP and LIME determine the main factors that drive the prediction. These outputs are presented in a sensible and organized format for investigators to review with the transaction information [35]. The framework provides for transactions to be classified as Low, Moderate, High and Critical, and includes matching priorities for investigation [36]. The possible outcomes of a decision are Approve, Monitor, Investigate and Escalate/Block. The final decision is still in the hands of human investigators who must verify the evidence, take into account contextual information, scrutinize potentially misleading explanations, etc. The framework also includes key concepts of human oversight, transparency, accountability, trust and appropriate automation. By keeping investigators involved in the decision-making process, there is a lower likelihood of automation bias, where the AI system's suggestion is accepted without further independent analysis [37]. Conceptual integration can also be done in the refinement of future models and explanations with the feedback of investigators [38]. The human-centric layer thus converts the research into a framework for investigations in the help of AI. Its main goal is to bridge the gap between chances of fraud and accessible evidence and actionable investigation priorities [39]. The

integration will allow the proposed system to aid in more transparent, consistent and informed cybersecurity fraud decisions and keep the final investigative authority in the hands of the human decision-maker.

Conceptual Framework

This study's conceptual framework sets out to link machine-learning-based fraud detection with Explainable Artificial Intelligence (XAI) and human-centred cyber security decisions. The framework aims to go beyond traditional fraud prediction, leveraging both predictive analytics and the explanations that are provided while incorporating the judgment of the investigator [40]. It assumes that the success of fraud detection can be achieved not only by correctly categorizing the fraud alerts but also by providing clear evidence to enable the human investigator to process and respond to the alerts produced by the AI. The framework starts with the credit-card transaction data set that resides in the United States including transaction attributes, merchant attributes, temporal attributes, geographic attributes, customer attributes, and historical fraud attributes [41]. The data are pre-processed, explored, engineered, and addressed for class-imbalance to create reliable analytical inputs. Processed data is then fed to machine learning models such as Logistic Regression, Random Forest, XGBoost, and LightGBM which will produce fraud predictions and probability scores. The next layer is the Explainable AI layer, which includes SHAP and LIME. SHAP offers the feature contributions globally and per transaction, whereas LIME offers local explanations for individual predictions [42]. These explanations provide details about elements that affect or reduce the risk of fraud, including transaction size, merchant segment, transaction timing, location, and customer attributes [43]. The framework next adds the layer of investigation of humans with explanations being converted into understandable evidence and risk categories [44]. Investigators confirm AI predictions, validate evidence, take into account the context, and make decisions on actions like approving, monitoring, investigating, or escalating.

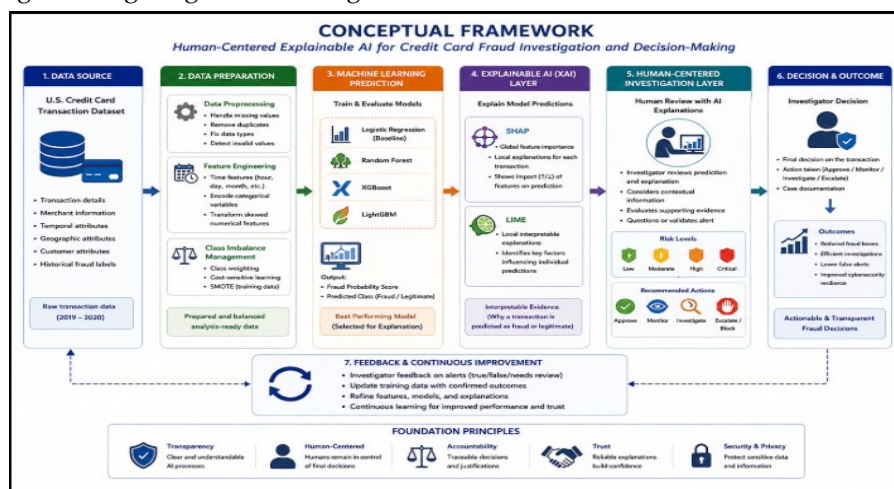


Figure 6. This flowchart illustrates the conceptual framework that combines XAI, human investigation, and fraud decision-making

The diagram shows the conceptual structure of the human-centered explainable AI for credit-card fraud investigation and decision-making. It starts with U.S. credit-card transaction data that includes transaction, merchant, temporal, geographic, customer and historical fraud information. The data are preprocessed, engineered, and handled for class imbalance before being analyzed via Logistic Regression, Random Forest, XGBoost and LightGBM. The model that performs best will provide fraud probabilities and fraud classifications. Lastly, SHAP and LIME provide explanations for global and individual predictions; they identify factors which enhance or diminish the likelihood of fraud [45]. These explanations are then checked by human investigators for the validity of supporting evidence, to evaluate the level of risk and to decide what action to take (to approve, monitor, investigate or escalate). Feedback from investigators to improve the model and explanation; measurable outcomes

from final decisions [46]. The framework is based on transparency, accountability, trust, human oversight and security.

Dataset

Screenshot of Dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	trans_date_time	merchant	category	amt	city	state	lat	long	city_pop	job	dob	trans_num	merch_lat	merch_long	is_fraud
1	1/1/2019 0:00	Heller, Gu	grocery_pos	107.2	Orient	WA	48.888	-118.2	149	Special	6/21/1978	1f76529f8	49.15905	-118.186	0
2	1/1/2019 0:00	Lind-Buck	entertainment	220.1	Malad City	ID	42.181	-112.3	4154	Nature	1/19/1962	a1a22d704	43.1507	-112.154	0
3	1/1/2019 0:07	Kiehn Inc	grocery_pos	96.29	Grenada	CA	41.613	-122.5	589	System	12/21/1945	413636e75	41.65752	-122.23	0
4	1/1/2019 0:09	Beier-Hya	shopping_pos	7.77	High Rolls	NM	32.94	-105.8	899	Naval	8/30/1967	8a6293af5	32.86326	-106.52	0
5	1/1/2019 0:21	Bruen-Yos	misc_pos	6.85	Freedom	WV	43.017	-111	471	Educator	8/2/1967	f3c43d336	43.75374	-111.455	0
6	1/1/2019 0:22	Kunze Inc	grocery_pos	90.22	Honokaa	HI	20.083	-155.5	4878	Physio	12/3/1966	95826e3ca	19.56001	-156.046	0
7	1/1/2019 0:22	Nitzsche,	shopping_pos	4.02	Valentine	NE	42.806	-100.6	4005	Netwo	3/15/1945	20490f3f0	42.47559	-101.266	0
8	1/1/2019 0:22	Kihn, Abe	shopping_net	3.66	Westfir	OR	43.758	-122.5	597	Forens	5/19/1961	870c92b28	44.27819	-121.815	0
9	1/1/2019 0:31	Ledner-Pf	gas_transport	102.1	Thompson	UT	39	-109.6	46	Survey	4/23/1987	47238da58	39.80731	-109.348	0
10	1/1/2019 0:34	Stracker-L	grocery_pos	83.07	Conway	WA	48.34	-122.3	85	Resear	9/1/1984	9b7a0619d	48.68211	-122.72	0
11	1/1/2019 0:40	Cummera	gas_transport	70.53	Athena	OR	45.829	-118.5	1302	Dealer	10/18/1976	7b91bd20f	46.04325	-117.962	0
12	1/1/2019 0:41	Huel-Lang	misc_net	177.6	Thompson	UT	39	-109.6	46	Survey	4/23/1987	048dc7fc7	38.41943	-109.38	0
13	1/1/2019 0:46	Ferry, Lyn	misc_net	2.76	San Jose	CA	37.33	-121.8	973849	Science	6/12/1955	70ca7fe41	37.54845	-121.536	0
14	1/1/2019 0:49	Little, Gut	shopping_net	83.52	Ravenna	NE	41.023	-98.9	2202	Solicitor	6/21/1974	5f7aafc80	40.99214	-98.8316	0
15	1/1/2019 0:56	Swaniawski	shopping_pos	317.1	Parks	AZ	35.256	-112	759	Geolog	3/28/1957	3fe3286b0	35.26812	-112.683	0
16	1/1/2019 0:56	Reichert,	shopping_net	113.4	Fort Wash	WY	43.005	-108.9	1645	Freight	2/26/1976	9d660a181	42.86897	-108.503	0
17	1/1/2019 1:00	Howe Lt	misc_pos	218.7	Littleton	CO	39.599	-105	320420	Water	7/31/1975	43204a453	39.33517	-104.103	0
18	1/1/2019 1:02	Wolf Inc	grocery_pos	89.11	Meadville	MO	39.78	-93.3	964	Tourist	12/23/1974	1cb03ad95	40.77331	-92.8213	0
19	1/1/2019 1:04	Vandervo	grocery_pos	50.68	Moab	UT	38.568	-109.5	9772	Locatio	11/24/1989	61861370f	38.48254	-108.854	0
20	1/1/2019 1:09	Ledner-Pf	gas_transport	90.54	Hawthorn	CA	33.914	-118.3	93193	Editor,	4/19/1995	295b782f9	32.99029	-117.837	0
21	1/1/2019 1:19	Schaefer,	gas_transport	51.33	Manville	WY	42.73	-104.7	241	Educator	12/16/1980	abaf4c635	42.50825	-105.345	0

(Source Link: <https://www.kaggle.com/datasets/dhruvb2028/credit-card-fraud-dataset>)

Dataset Overview

The data set used for the study is the publicly available Credit Card Fraud Detection for Machine Learning dataset from Kaggle, which consists of credit-card transactions made in the western United States. Such a combination of binary fraud outcome with transaction, merchant, temporal, geographic and customer characteristics makes the data set especially suitable for this research, enabling the use of machine-learning methods for fraud detection as well as the use of Explainable Artificial Intelligence (XAI) methods for giving meaningful explanations for individual predictions. The dataset is in CSV format and has a size of around 61.7 MB and is named `credit_card_frauds.csv`. The data spans a period of history from around 2019 to 2020 and, as per the dataset description, offers a comprehensive context for temporal and behavioral patterns of credit-card fraud. The dataset consists of around 15 variables such as transaction date and time, merchant, transaction category, transaction amount, city, state, state name, latitude, longitude, city population, customer occupation, and fraud indicator [47]. The transaction date and time variable gives details to consider temporal patterns such as transaction time, day, month, and possibly any unusual activity time [48]. Merchant and transaction-category variables describe the business context of each transaction and can be used to identify transaction categories with high fraud risk. The transaction amount is a numeric value that is used for analyzing very large or unusual transactions. Geographic variables such as city, state, latitude, longitude and city population, allow geographic analysis and identify location related fraud indicators [49]. The field occupation is a supplementary field that gives more context about the customer that can be analyzed for any relationship with fraud transactions. The fraud indicator is the main target variable and is used to distinguish fraudulent transactions from legitimate transactions, for supervised binary classification [50]. The data set is a class-imbalance problem, with legitimate transactions outnumbering fraudulent transactions significantly, and needs to be solved with class weighting, cost-sensitive learning, or careful use of SMOTE only on the training data set. The data can also be analyzed for exploration of such attributes as transaction distributions, merchant category comparisons, patterns over time, and state level fraud analysis, geographic visualization, and transaction to fraud outcome relations [51]. Human-centered XAI point the dataset has a significant benefit in the aspect of having many features that have understandable meanings for investigators [52]. The proposed framework can provide explanations of fraud predictions based on evidence rather than relying on anonymous or mathematically transformed variables such as transaction amount, merchant category, transaction timing, location, customer characteristics, etc. These can then be

explained with the help of SHAP and LIME and subsequently be converted into human-readable fraud-risk evidence [53]. Thus, the data set serves as a proper basis for building and testing the proposed framework, linking fraud prediction, explainability, human investigation and risk-based decision-making in a U.S.-based credit-card fraud setting.

3. Results

The results illustrate the ability of the proposed Human-Centered Explainable AI framework to enhance cybersecurity fraud investigation and decision-making using the dataset of U.S.-based credit card transactions used in this study [54]. This analysis starts by analyzing the characteristics and distribution of the fraudulent transactions, namely class imbalance, transaction categories, temporal patterns, transaction amount, and geographical variation [55]. The results are then compared to four machine learning models: Logistic Regression, Random Forest, XGBoost and LightGBM. The results in model comparison show that ensemble-based models can outperform the baseline Logistic Regression model in terms of fraud-detection performance, with LightGBM performing the best in terms of recall, ROC-AUC, and PR-AUC. The selected model is then explained using the Explainable Artificial Intelligence (XAI) methodology [56]. SHAP analysis highlights the most important factors that impact fraud predictions, such as transaction amount, category frequency, transaction hour, and customer age. Together, these findings suggest that predictive modeling and evidence-based decision making can be more robust and effective when the evidence is interpretable, transparent, and investigator-centered, as is the case in these results.

Distribute classes and determine the extent of fraud

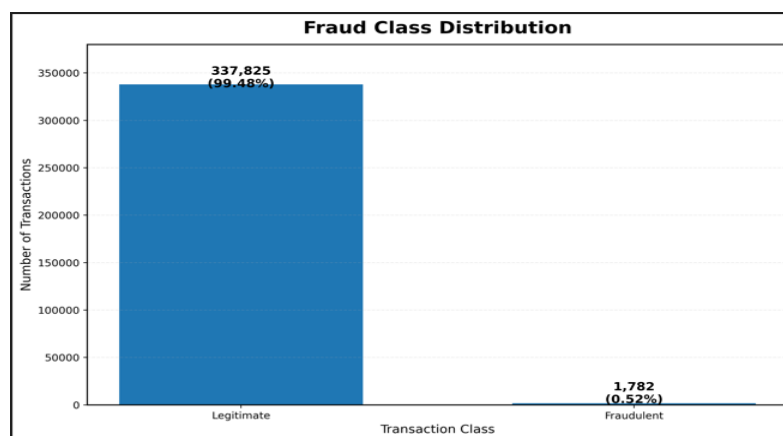


Figure 1. This image displays severe class imbalance between legitimate and fraudulent credit-card transactions

The distribution of legitimate and fraudulent transactions is shown in Figure 7 for the credit-card data set it is known to contain a subset of legitimate transactions and a subset of fraudulent transactions. The results indicate a highly imbalanced class distribution with 337,825 legitimate transactions accounting for 99.48% of all transactions and 1,782 transactions as fraud accounting for 0.52% of all transactions. Which implies that there is a substantial portion of the dataset that is composed of legitimate transactions, making it a crucial methodological problem in designing machine learning algorithms for fraud detection [57]. A model with a high overall accuracy with poor performance in the minority class of fraud transactions can be obtained. So, the accuracy alone would not give us a realistic measure of the fraud-detection ability. The imbalance observed confirms the need to employ class imbalance methods, such as Class Weighting, cost-sensitive learning and controlled oversampling when building models [58]. It also makes it important to focus on recall, precision, F1-score, PR-AUC, false-positive rate and false-negative rate in further model evaluation [59]. A human-centered approach to effectively prioritizing investigations and aiding decisions is the ability to correctly identify the relatively small number of fraud cases.

Transaction Category-Based Fraud Risk Analysis

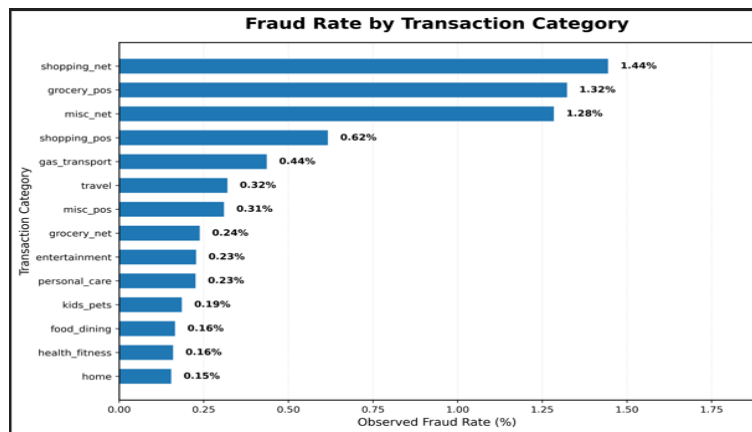


Figure 8. This image shows the difference in the fraud rate of the various types of credit-card transactions

Figure 8 shows the observed fraud rate for each one of the transaction types for the considered credit-card fraud dataset. The findings show there is a significant difference in the rates of fraud in each category. The top three fraud rates are seen on shopping_net (0.44%), grocery_pos (1.32%) and misc_net (1.28%). The categories show significantly higher rates of fraud than the other transaction types. Fraud rates for Shopping_pos and gas_transport are 0.62% and 0.44% respectively. The rates for travel and misc_pos are moderate (0.32% and 0.31%, respectively). On the other hand, home transactions are the ones that are observed to have the lowest rates of fraud at 0.15% followed by health_fitness and food_dining with rates observed to be around 0.16%. The results suggest that the transaction category could be a valuable contextual indicator for detecting fraud risk. But just because it's in a category doesn't mean that it can only be caused by that category. Rather, it can add to the machine-learning predictions along with the amount of transactions, time, location, and customer attributes [60]. These patterns also serve as potentially meaningful evidence for any subsequent SHAP based explanation and/or investigator decision support.

Analyzing fraudulent transactions over time

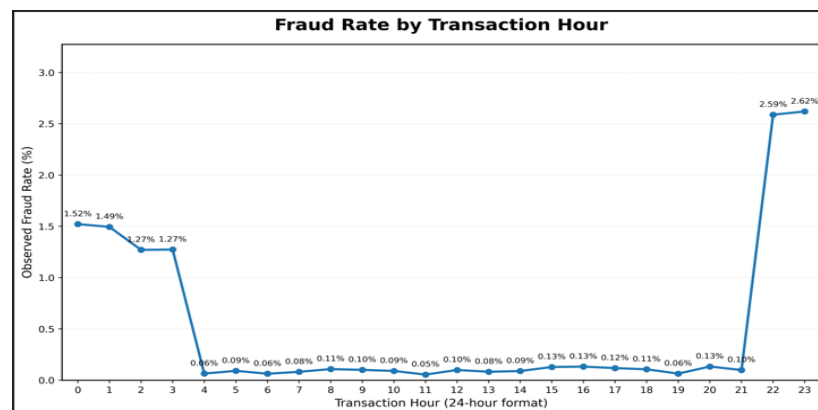


Figure 9. This graph depicts the time-based changes in the actual fraud rate by transaction hour

The observed number of frauds per transaction, shown in Figure 9, indicates a distinct time-of-day fraud trend. Fraud rates are relatively high during the early morning hours, reaching 1.52% at 00:00, 1.49% at 01:00, and approximately 1.27% at both 02:00 and 03:00. This rate drops significantly after 04:00 and stays below ~0.14% for much of the day and night. A key area which sees a sharp change is late evening when the fraud rate climbs to 2.59% at 22:00 and to a high of

2.62% at 23:00. The results indicate that transaction timing can be an important predictive criteria to identify potentially suspicious activity. Higher fraud rates occurring during the late night and early morning hours may be useful for prioritizing alerts resulting from this time frame [37]. Temporal patterns are not, sufficient evidence of fraud without taking into account contextual characteristics such as transaction amount, merchant category, location and so on.

Transaction Amount Differences between Fraudulent and legitimate transactions

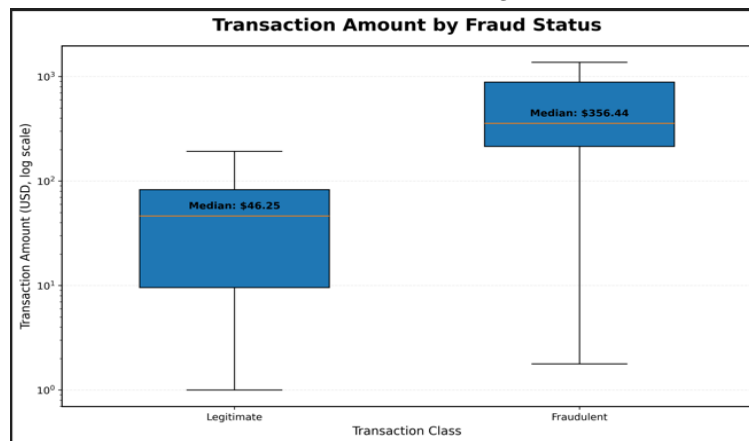


Figure 10. This image shows substantially higher transaction amounts among fraudulent transactions compared with legitimate transactions

Figure 10 plots a boxplot of the distribution of transaction amounts by legitimate and fraudulent transactions, and shows that amounts are distributed in a log-normal manner. The results indicate that there is a significant difference in the transaction value between the two classes. The median of the transaction amount for legitimate transactions is \$46.25, while the median for fraudulent transactions is much higher at \$356.44. This suggests that the amount of money associated with fraudulent transactions is significantly higher than that of legitimate transactions in the dataset. Fraudulent groups also exhibit greater distribution, starting at relatively low transaction values and going above \$1,000, suggesting a broad range of fraudulent transaction behavior. Since transaction amounts are very right skewed, the logarithmic scale is better [38]. The results indicate that transaction amount is a factor of significance in predicting fraud and could be a significant factor in machine-learning classification [39]. Unusually large transactions can also offer clear corroborating evidence for investigators to consider as they investigate suspicious transactions and review fraud alerts generated by AI.

Geographic Variation in Credit-Card Fraud Rates

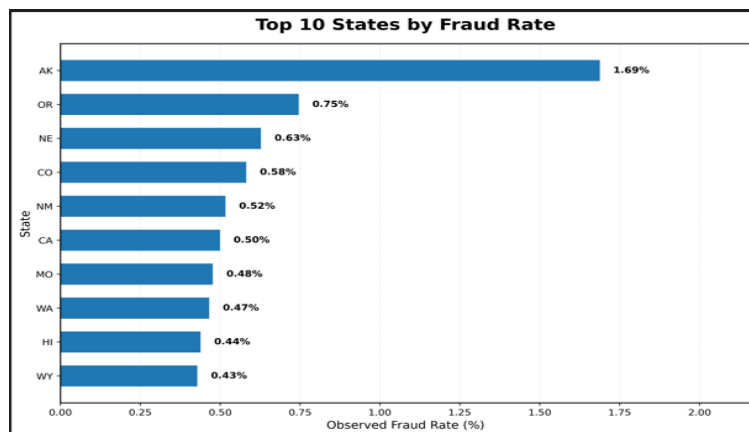


Figure 11. The image shows geographic variation in the observed credit card fraud rates in Ten U.S. States

The top 10 U.S. states ranked by observed credit-card fraud rates are shown in Figure 11. The findings show that there is a clear geographic variation in the prevalence of fraudulent transactions in the states sampled. Alaska (AK) has the highest reported rate of fraud at 1.69%, which is much higher than all other states. Oregon (OR) is second with a fraud rate of 0.75% and Nebraska (NE) is 0.63%. The rates for Colorado (CO) and New Mexico (NM) are 0.58% and 0.52%, respectively. California (CA) records 0.50%, followed by Missouri (MO) at 0.48% and Washington (WA) at 0.47%. Hawaii (HI) and Wyoming (WY) have relatively low rates of 0.44% and 0.43%, respectively. The results suggest that geography may be a good source of contextual clues for fraud-risk assessment [40]. That said, it's not necessarily an indication of direct location-to-fraud connections, as the number of transactions and the size of transactions also differ by state. Geographic features can then be used along with other predictive factors in the proposed XAI framework.

Comparing the performance of the machine-learning fraud detection models

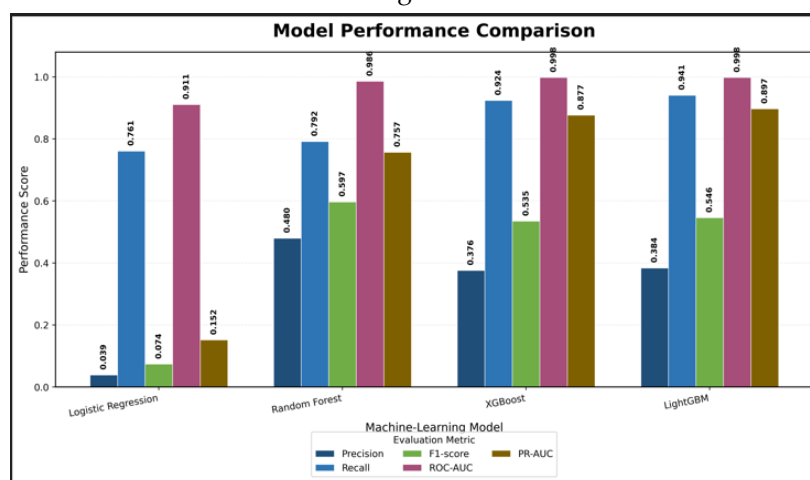


Figure 12. This image shows how four machine-learning models compare in terms of five fraud-detection metrics

In Figure 12, the performance of four machine-learning algorithms (Logistic Regression, Random Forest, XGBoost, and LightGBM) is compared with each other using precision, recall, F1 score, ROC, and PR. The results show that there is a clear difference in predictive performance between the models. The logistic regression model has a significant false-alert burden as it has a low precision score (0.039), a low F1 score (0.074), and a high recall score (0.761) and ROC-AUC (0.911). Random Forest provides a stronger balance, achieving 0.480 precision, 0.792 recall, 0.597 F1-score, 0.986 ROC-AUC, and 0.757 PR-AUC. The recall of XGBoost is significantly higher at 0.924, and the ROC-AUC and PR-AUC are 0.998 and 0.877, respectively. LightGBM is the best overall, with the best recall score (0.941), ROC-AUC score (0.998), and PR-AUC score (0.897). The results justify the choice of LightGBM for the next explainability analysis with SHAP and LIME, especially regarding the model's ability to detect fraud and discriminate against minorities.

Comparative Classification Discrimination Using ROC-AUC

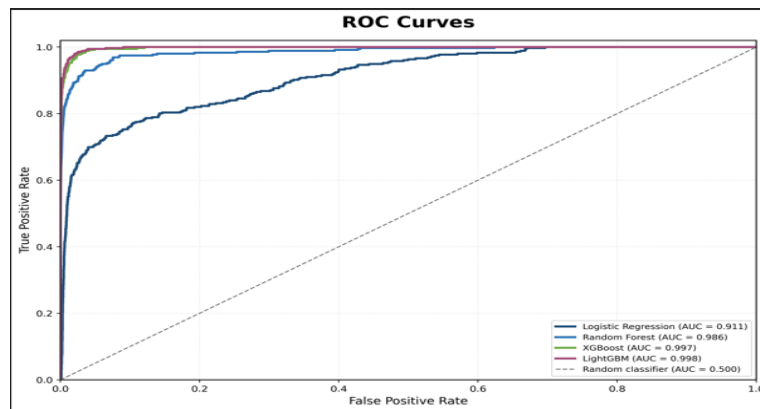


Figure 13. This image shows the classification discrimination of four machine-learning fraud detection models based on the ROC

The Receiver Operating Characteristic (ROC) curves of the four machine-learning models tested in this study are as shown in Figure 13. The ROC curve shows the relationship between the true positive rate (TPR) and false positive rate (FPR) at various classification thresholds, and the area under the curve (AUC) is a global evaluation of the model's discrimination. The results show good classification performance of all models, with significant differences being observed. Logistic Regression has an AUC of 0.911, which is a well discriminating model but not as good as the ensemble models. Random Forest significantly enhances discrimination and has an AUC of 0.986. XGBoost achieves an AUC of 0.997, while LightGBM produces the highest AUC of 0.998. The nature of the curves for XGBoost and LightGBM is similar, with them staying on the upper left part of the graph, which means they have significant separation of frauds from legitimate transactions with relatively low false-positive rates [41]. These results further validate the use of LightGBM for further post-explanatory analysis using SHAP and LIME.

Explainable AI-Based Feature Importance Analysis

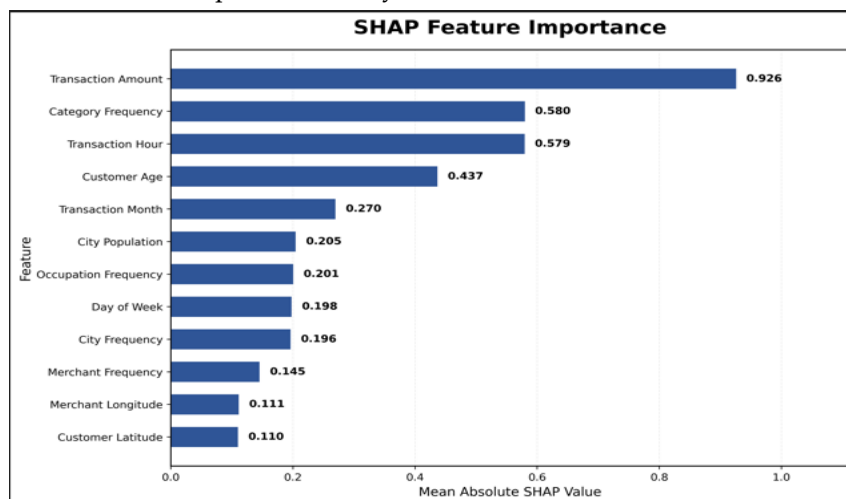


Figure 14. This image visualizes LightGBM's importance for interpreting fraud predictions using SHAP

In Figure 14, the global feature importance (SHAP) is shown for the selected LightGBM fraud-detection model. Among the results, the most influential predictor is the transaction amount, which has a mean absolute SHAP value of 0.926, significantly higher than the other predictors. The next

most influential features are category frequency (0.580) and transaction hour (0.579), showing that context of the transaction as well as temporal behavior are very important for prediction of fraud. Age of customers (0.437) is also informative about the predictive power, followed by month of transaction (0.270), city population (0.205), occupation frequency (0.201), day of week (0.198), and city frequency (0.196). The merchant frequency (0.145), merchant longitude (0.111) and customer latitude (0.110) contribute comparatively less [42]. The results show how SHAP can be used to make a model more interpretable and give a better decision support system. Investigators can also leverage these feature-importance results from a human viewpoint to gain insight into the features that are most predictive of AI-driven fraud scoring, helping them focus on key factors when investigating a transaction.

4. Discussion

The results show that the proposed Human-Centered Explainable AI framework is capable of integrating predictive fraud detection, interpretable evidence, and decision support for investigators [43]. The first one of the analyses of the dataset was a severe class imbalance where only 0.52% of observations are fraudulent and 99.48% are legitimate. This discovery has been reflected in a series of studies that showed that standard accuracy would not be optimal to measure fraud-detection performance since a model might be accurate, but might not detect the minority fraud class. Therefore, the precision, recall, F1 score, ROC-AUC and PR-AUC better reflect model effectiveness. Again, transaction-category analysis revealed meaningful differences in the observed frequency of fraud across various transaction categories; shopping_net, grocery_pos, and misc_net were comparatively higher. The patterns indicate that the context of a transaction can provide additional information that is useful in assessing fraud risks, but should not be read as a causal explanation or considered independently. The temporal analysis also revealed significant differences in the rates of fraud by hour of transaction. A high percentage of fraud (2.59%) occurred during the late night hours at 22:00 and 2.62% at 23:00, with most daytime hours showing much lower percentages [44]. The temporal variation can be an important contextual indicator for suspicious alert prioritization, if the time of the transactions matters. Likewise, the transaction-amount analysis showed there are significantly different amounts of legitimate and fraudulent transactions, with the median amounts of \$46.25 and \$356.44, respectively [45]. This is significant in the field of explainable fraud investigation as the amount of the transaction is a feature that investigators are easily able to decipher and assess. Also, geographic analysis revealed that the fraud rate was different among U.S. states, with Alaska having the highest observed fraud rate of the ten states examined. Geographic differences should be interpreted with caution, as the number of transactions and population characteristics differ from state to state [46]. The comparative analysis of machine learning gave more evidence of the effectiveness of the models. Logistic Regression had high recall, but moderate precision and F1-score, indicating that it can flag many of the actual fraudulent transactions but also many false alerts. Random Forest achieved the best precision and recall, XGBoost and LightGBM showed significantly better discrimination [46]. LightGBM obtained the top recall of 0.941, ROC-AUC of 0.998 and PR-AUC of 0.897, which helped in building the case for choosing LightGBM for the explainability stage. The SHAP analysis has additionally reinforced the human-centered point of view, with transaction amount being the most significant feature, followed by category frequency, transaction hour, and customer age. These results prove that XAI can transform the model's behavior into evidence that can be comprehended and assessed by investigators [47]. SHAP-based evidence can provide a useful way for analysts to understand the reasons behind an alert and whether further investigation is warranted, rather than accepting the AI prediction as a fact. The results showed that being predictive was not enough to ensure trustworthy fraud investigation [48]. This powerful minority-class detection, interpretable feature contributions and human oversight gives a far more practical base for transparent and responsible cybersecurity fraud decisioning.

Ethical Concerns and Limitations

There are several ethical issues and methodological constraints regarding the development and use of an XAI system for fraud investigation [49]. Ethically, fraud-detection models should ensure customer privacy as financial transaction records might include sensitive data such as financial

activity, geographic location, occupation, and behavioral patterns. Responsible and secure handling, anonymization and access control of the data are still crucial, even though the dataset presented here is provided for research use [50]. Algorithmic fairness is also a worry since the model predictions could produce different rates of error among geographic/ demographic groups [51]. Thus, fraud alerts should not be used as a definitive proof of fraud. Implementing human verification of AI recommendations to minimize automation bias and ensure accountability. Ensuring accountability by having human investigators review AI recommendations before making consequential decisions, minimizing automation bias [51]. Methodologically, the selected data set consists of credit-card transactions mostly from the western United States, and may not be representative of fraud activity in different areas or financial settings. The data set also captures historical transaction patterns, so new fraud patterns might not be well represented [52]. Another challenge is the existence of severe class imbalance, which can impact model stability and generalizability as the number of fraudulent observations is relatively small. SHAP explanations will describe the behavior of the model but do not necessarily indicate causality between features and fraud. LIME explanations may also vary depending on local sampling [53]. Lastly, the study does not perform a real-world investigator usability experiment, thus constraining conclusions about real human decision-making effectiveness.

Future Work

Future work can build on this study by creating a more extensive and operationally realistic human-centered Explainable AI framework for cybersecurity fraud investigation and decision-making. Future studies should test the proposed framework with more extensive, more recent and geographically spread financial transaction data across various financial institutions, payment channels and fraud types in the United States [54]. This would help to enhance the model generalizability and help researchers to investigate the consistency of fraud patterns across various populations and financial environments [55]. Second, future studies can explore novel machine-learning and deep-learning techniques such as neural networks, graph neural networks, transformer models, and ensemble architectures, to capture complex relationships between customers, merchants, devices, and locations, as well as between sequences of transactions [56]. These should be integrated with explainability techniques as well to prevent loss of transparency when predictive performance is enhanced. Third, research needs to be conducted on the topic of real-time fraud detection, where transactions are assessed in real-time, and investigators are given dynamic risk scores and explanations [57]. Architectures based on streaming data might be able to enable continuous monitoring and swift action on emerging patterns of fraud. Fourth, future study is needed to compare several XAI techniques, such as SHAP, LIME, counterfactual explanations, Partial Dependence Plots and Integrated Gradients, to test the consistency of explanations, stability of explanations, and comprehensibility of explanations [58]. Explanations should be highlighted that can be most helpful to nontechnical fraud investigators. Fifth, the importance of human-centered evaluation for research should be greatly increased. Professional or trained fraud analysts can be involved in controlled experiments to assess the effectiveness of explanations in enhancing investigation accuracy, shortening decision time, reducing workload of false positives, and improving trust in AI recommendations [59]. The algorithmic recommendations in future studies should also be examined for evidence of automation bias and when human intervention should override the algorithmic recommendation [60]. The principles of fairness and ethical assessment should be built in via subgroup performance analysis, which would seek to determine if there are differences between geographic, demographic or customer-based subgroups in terms of precision, recall, false positive rates and false negative rates. Future work could involve a continuous human feedback loop, where investigative actions are fed back to the model's learning, explanations, and threshold. These additions would make the proposed framework more effective in the future in the event the scheme evolves into a more effective and secure AI-assisted investigation tool [61], [62].

5. Conclusion

This study examined the potential of **Human-Centered Explainable Artificial Intelligence (XAI) for improving cybersecurity fraud investigation and decision-making using a U.S.-based credit-card transaction dataset. This study brought together three key elements of machine learning prediction, explainability, and human oversight in a single fraud-investigation model. The results

illustrate that the data set has meaningful patterns related to fraudulent transactions, such as variation in transaction category, transaction timing, transaction amount and geographic area. The large class imbalance in the data even with fraudulent transactions constituting only 0.52% of the observations further illustrates the need to focus on the right metrics for evaluation, which is normally classification accuracy, rather than being limited to that metric. Ensemble-based approaches gave a better fraud-detection performance than Logistic Regression according to the comparative modeling results. LightGBM in particular, outperforming the other models, with overall the recall of 0.941, ROC-AUC of 0.998 and PR-AUC of 0.897, succeeded in the recognition of the minority fraud class while also giving the best results in distinguishing the fraudulent transactions. To gain insights into the features that most strongly affect LightGBM's predictions, explainability was used, and the results showed that the amount of the transaction, the frequency of the category, the hour of the transaction, and customer age are important explanatory factors. They show how XAI can make those intricate machine learning predictions understandable as a resource that investigators can consult when evaluating suspicious transactions. This proposed framework also highlights that AI should not be used as a standalone agent of investigation, but rather as a tool to support human investigation, thereby boosting transparency, accountability, and trust in AI-driven cybersecurity investigations. In conclusion, the study underscores the potential of combining machine learning, SHAP/LIME explainability, and human-in-the-loop decision-making processes to create more transparent, accountable, and effective credit-card fraud detection systems.

REFERENCES

- [1] H. K. Ofori, W. L. Brown-Acquaye, F. Lempogo, K. Bell-Dzide, and I. E. Agbehadji, "Explainable AI in insider financial fraud detection models: A review of transparency and trust," in Proc. 21st Int. Conf. Cyber Warfare and Security, Mar. 2026, Academic Conferences and Publishing Limited.
- [2] P. W. C. Prasad, M. S. Sayeed, D. M. Nguyen, D. P. Hutabarat, and G. M. Mohiuddin, "Explainable AI: Enhancing decision-making in the detection of cyber threats," *Frontiers in Computer Science*, vol. 8, Art. no. 1762332, 2026.
- [3] J. Chen, K. R. Garcia, Y. E. Zhang, L. E. Dudley, and A. D. Warren, "Human-centered artificial intelligence in cybersecurity," in *Handbook of Human-Centered Artificial Intelligence*. Singapore: Springer Nature Singapore, 2026, pp. 2283–2317.
- [4] A. Bhatia and A. Kumar, "AI explainability and trust in cybersecurity operations," in *Deep Learning Innovations for Securing Critical Infrastructures*. IGI Global Scientific Publishing, 2025, pp. 57–74.
- [5] M. A. Rahman and B. T. Haque, *XAI-Driven Cybersecurity*. Essay Publication Research and Consultancy, 2026, 240 pp.
- [6] H. An, H. Salemi, Q. Zhang, H. Purohit, and J. H. Cho, "Explainable AI for cyberscam prevention: A survey of multimodal scam detection, evaluation, and human-centered training," Jun. 14, 2026.
- [7] A. Calvo Ibáñez, "User-oriented explainable AI (XAI) for decision-making in critical sectors," Ph.D. dissertation, Universitat Politècnica de Catalunya, 2025.
- [8] R. Damodaram, "Explainable and trustworthy artificial intelligence architectures for enterprise cybersecurity and decision intelligence," *International Journal of Science, Research and Technology*, vol. 8, no. 4, pp. 14615–14623, 2025.
- [9] G. Cecchi, F. Benelli, M. Caronna, G. Palma, and A. Rizzo, "Connecting the dots: A systematic literature review of explainable AI, cybersecurity, human-centered design and edge computing," *Journal of Cybersecurity and Privacy*, vol. 6, no. 3, Art. no. 91, 2026.
- [10] S. Rahman, "Explainable trust-centric artificial intelligence for integrated healthcare, financial security, and cyber-risk management," *Frontiers in Computer Science and Artificial Intelligence*, vol. 4, no. 5, pp. 1–6, 2025.
- [11] D. Mahida, B. Bhatt, and B. Gambhava, "Cyber frauds and the human mind: Behavioral vulnerabilities and adaptive defense mechanisms," in Proc. 2025 IEEE Int. Conf. Advanced Computing Technologies (ICACT), Sep. 2025, pp. 510–515.
- [12] D. S. Adeyemi, "Human-AI collaboration in cybersecurity decision-making: A systematic review of literature," *Communication in Physical Sciences*, vol. 13, no. 3, pp. 403–421, 2026.
- [13] T. C. Troublefield, "Human-centric artificial intelligence in cybersecurity: Integrating cyberpsychology for

the next generation defense mechanisms," *International Journal of Security (IJS)*, vol. 16, no. 1, pp. 1–25, 2025.

[14] B. A. Bagustari, F. S. Far, N. A. Ghani, S. Hamid, S. H. Almadi, and A. R. Alshehadeh, "Exploring the human-centric approach to artificial intelligence's role in FinTech security," in *Proc. Conf. Sustainability and Cutting-Edge Business Technologies*. Cham, Switzerland: Springer Nature Switzerland, Apr. 2025, pp. 129–136.

[15] L. Reinhardt, "Explainable artificial intelligence as a foundation for responsible, ethical, and human-centered decision-making systems," *Colombo Scientific Journal of Computer Science & Information System*, vol. 10, no. 9, pp. 40–44, 2025.

[16] M. Zalfani, E. Süle, and M. Bakar, "Human-centered AI for decision support systems: Enhancing usability and trustworthiness," *Systems*, vol. 14, no. 6, Art. no. 651, 2026.

[17] C. Luca, "Explainable artificial intelligence for enhanced human-AI interaction," 2026.

[18] S. A. Schilhabel, "Design and evaluation of a human-centered explainable AI dashboard for cybersecurity education."

[19] D. H. Nguyen, D. P. T. Nguyen, Q. C. Ta, K. D. Tran, and K. P. Tran, "Human-centered explainable anomaly detection in smart manufacturing: Bridging AI and human decision-making in Industry 5.0," in *Human-Centered Explainable Anomaly Detection for Smart Manufacturing in Industry 5.0: Theories, Applications and Case Studies*. Cham, Switzerland: Springer Nature Switzerland, 2025, pp. 93–108.

[20] F. M. Khan, A. Zeb, T. Rahman, M. A. Al-Khasawneh, Y. I. Daradkeh, I. F. Siddiqui, et al., "XAI-driven data mining for self-defending IoT systems: Enhancing cybersecurity transparency in the age of smart cities," *Cognitive Computation*, vol. 18, no. 1, Art. no. 16, 2026.

[21] C. Mohitkar and D. Lakshmi, "Explainable AI for transparent cyber-risk assessment and decision-making," in *Machine Intelligence Applications in Cyber-Risk Management*. IGI Global Scientific Publishing, 2025, pp. 219–246.

[22] P. Selvakumar, C. Singh, S. K. Basu, T. C. Manjunath, S. N. Shalini, and R. K. Mishra, "Challenges of explainable AI in security decision-making," in *Intelligent Cyber Defense for Critical Infrastructure*. IGI Global Scientific Publishing, 2026, pp. 87–114.

[23] N. S. Thalpage and E. Jayarathne, "Explainable AI approaches for detecting and mitigating phishing attacks: A review," 2025.

[24] V. Subramaniam, "Transforming enterprise cloud computing through explainable AI driven security analytics and intelligent automation," *International Journal of Technology, Management and Humanities*, vol. 11, no. 4, pp. 191–199, 2025.

[25] M. Z. Uddin, "Trustworthy AI and explainability," in *Trustworthy Multimodal Intelligent Systems for Independent Living*. Cham, Switzerland: Springer Nature Switzerland, 2025, pp. 111–137.

[26] G. Sarkar, A. Sachdeva, and S. K. Shukla, "From trust to lock-in: Human-centered AI for understanding investment fraud narratives in Indian digital communities," in *Proc. Workshop on Security and Privacy for Asian Internet Communities*, Jun. 2026, pp. 1–11.

[27] M. Prasanna Kumar, "Explainable and ethical cloud artificial intelligence frameworks for risk-sensitive financial systems management."

[28] F. Adedoyin, H. Dogan, D. Cetinkaya, and N. Jiang, "Human-centered AI in FinTech: A conceptual model and strategic research agenda," in *Proc. 2025 IEEE Conf. Artificial Intelligence (CAI)*, May 2025, pp. 180–187.

[29] U. I. Henry, G. I. Aimufua, S. I. Bassey, and U. Musa, "Artificial intelligence-driven cybersecurity: A review of modern techniques and future directions," 2025.

[30] Z. Fatima, R. Hussain, A. Dilshad, M. Shakir, and A. A. Laghari, "Explainable AI-driven firewall evaluation: Empowering cybersecurity decision-making for optimal network defense," *Security and Privacy*, vol. 8, no. 3, Art. no. e70026, 2025.

[31] J. Mate, "Explainable artificial intelligence in modern machine learning systems: Enhancing transparency, trustworthiness, and ethical decision-making," 2026.

[32] B. Pandiya and P. Kulkarni, "Managing complexity in cybersecurity: The necessity of human oversight in digital immune systems from behavioral forensic perspective," in *Digital Immune System: Principles and Practices*, 2026, pp. 145–166.

[33] D. B. Audretsch, "Explainable artificial intelligence (XAI) techniques for enhancing transparency in intelligent decision-making systems," *Journal of Smart Computing and Data Intelligence*, vol. 1, no. 1, pp. 42–48,

2025.

- [34] S. Sawant, A. Bhanu, B. G. Schelble, and K. C. Madathil, "Addressing AI vulnerabilities through human-centered approaches and risk frameworks," in *Advances in Human-AI Collaboration*, 2026, pp. 325–339.
- [35] M. R. Anderson, J. L. Carter, D. P. Thompson, E. S. Walker, and C. James, "Explainable data mining for financial and security decision systems."
- [36] M. A. Hussain, "Exploring the collaboration between AI and human expertise in risk management for improved decision-making in software product security: A qualitative study," Ph.D. dissertation, University of the Cumberlands, 2025.
- [37] S. A. Oyedotun, G. P. Oise, and C. E. Ozobialu, "Towards intelligent cybersecurity in SCADA and DCS environments: Anomaly detection using multimodal deep learning and explainable AI," *Journal of Science Research and Reviews*, vol. 2, no. 3, pp. 20–31, 2025.
- [38] D. S. Sachs, *Behavioral Insights in Cybersecurity: A Guide to Digital Human Factors*. CRC Press, 2025.
- [39] J. El-Masry, "Adaptive explainable AI framework for trustworthy decision support in high-stakes intelligent applications," *Research Journal of Computer Systems and Engineering*, pp. 50–56, 2026.
- [40] O. K. Oladele, "Human factors and insider threat detection in modern cybersecurity systems," 2026.
- [41] U. Zafar and F. Wu, "Methodological challenges in explainable AI for fraud detection: A systematic literature review," *Artificial Intelligence Review*, vol. 59, no. 4, Art. no. 115, 2026.
- [42] L. Gu and R. Huang, "From reactive to proactive: An AI-enabled human-centered framework for telecommunication fraud prevention," *Design for Augmented Humanity*, Art. no. 29776481261477523, 2026.
- [43] I. O. Wiratsin and C. Ragkhitwetsagul, "Effectiveness of explainable artificial intelligence (XAI) techniques for improving human trust in machine learning models: A systematic literature review," *IEEE Access*, vol. 13, pp. 121326–121350, 2025.
- [44] C. S. Shwetha, "Explainable autonomous intelligence frameworks for next-generation cloud and cyber defense systems," *International Journal of Humanities and Information Technology*, vol. 8, no. 2, pp. 1–10, 2026.
- [45] G. Bindu Madhavi, P. Rajeshwari, P. Raja Sekhar Reddy, and P. Srilatha, "Integrating explainable AI into credit card fraud detection for transparent decision-making," 2025.
- [46] R. Madhumith, M. Ouaisa, P. Hemashree, and S. B. Mahalakshmi, "Advancing cybersecurity with generative adversarial networks and explainable AI: A comprehensive exploration," in *Generative Adversarial Networks for Cybersecurity*. Auerbach Publications, 2026, pp. 155–179.
- [47] R. Sugumar, "Explainable AI-driven secure multi-modal analytics for financial fraud detection and cyber-enabled pharmaceutical network analysis," *International Journal of Advanced Research in Computer Science & Technology (IJARCST)*, vol. 8, no. 6, pp. 13239–13249, 2025.
- [48] R. Almarwani, M. Almarwani, and F. Almarwani, "AI-driven synthetic threats in cybersecurity: A user-centered framework for awareness, detection, and protective behavior," *IEEE Access*, 2026.
- [49] L. Sun, W. Xu, and Z. Gao, "A human-centered privacy (HCP) approach to AI," in *Handbook of Human-Centered Artificial Intelligence*. Singapore: Springer Nature Singapore, 2026, pp. 1–47.
- [50] P. Jelinić, "Artificial intelligence and the future of cooperative banking: Human-centered innovation, responsible governance, and sustainable member value," May 12, 2026.
- [51] Y. Harrath, O. Adohinzin, J. Kaabi, and M. Saathoff, "Bridging domains: Advances in explainable, automated, and privacy-preserving AI for computer science and cybersecurity," *Computers*, vol. 14, no. 9, Art. no. 374, 2025.
- [52] M. J. Haque, "Explainable artificial intelligence for biometric authentication: Enhancing trust, security, and transparency in deep models."
- [53] S. A. Shreha, R. N. Ravikumar, and S. Aarthi, "Neuro-symbolic AI and behavioral intelligence: Towards explainable cybersecurity mechanisms," in *Neuroscience-Enhanced Artificial Intelligence for Cybersecurity*. IGI Global Scientific Publishing, 2026, pp. 121–152.
- [54] P. Ye and F. Y. Wang, "Social computing: A human-centered AI perspective," in *Human-Centered AI: Advancing Research and Applications in Human-AI Interaction*. Singapore: Springer Nature Singapore, 2026, pp. 523–554.
- [55] O. Kaya, A. S. Shah, M. A. Karabulut, S. N. Karahan, M. S. Osmanca, and N. Acir, "Explainable artificial intelligence (XAI): Concepts, applications, challenges, and future perspectives," *IEEE Access*, 2026.
- [56] H. Karthik and A. Kore, "Designing human-centered AI experiences," in *Handbook of Human-Centered*

Artificial Intelligence. Singapore: Springer Nature Singapore, 2026, pp. 769–831.

[57] D. Morgan and S. Lambert, "Adaptive XAI-driven cybersecurity architecture for real-time threat correlation, governance automation, and risk decision intelligence in US enterprise systems," 2026.

[58] M. A. Rahman and B. T. Haque, "Adaptive cybersecurity threat intelligence using explainable artificial intelligence for resilient protection of US critical information systems," *American Journal of Advanced Technology and Engineering Solutions*, vol. 1, no. 2, pp. 216–261, 2025.

[59] M. Anderson, J. Collins, R. Mitchell, E. Thompson, and C. James, "Explainable security analytics for financial institutions."

[60] J. Moss, J. Gordon, W. Duclos, Y. Liu, Q. Wang, and J. Wang, "Explainable AI in IoT: A survey of challenges, advancements, and pathways to trustworthy automation," *Electronics*, vol. 14, no. 23, Art. no. 4622, 2025.

[61] J. Lee, "Trustworthy, adaptive, and human-centered artificial intelligence," *Spectrum of Research*, vol. 6, no. 1, 2026.

[62] D. B., "Credit Card Fraud Dataset," Kaggle. [Online]. Available: <https://www.kaggle.com/datasets/dhruvb2028/credit-card-fraud-dataset>.