

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL

THE IMPACT OF HUMAN AI COLLABORATION
ON AI APPLICATIONS

BURHANETTİN BURUN

A THESIS OF MASTER OF SCIENCE
DEPARTMENT OF COMPUTER ENGINEERING

ADVISOR: ASSOC. PROF. DR. MEHMET GÖKTÜRK

SEPTEMBER 2025

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL

THE IMPACT OF HUMAN AI COLLABORATION
ON AI APPLICATIONS

BURHANETTİN BURUN

A THESIS OF MASTER OF SCIENCE
DEPARTMENT OF COMPUTER ENGINEERING

ADVISOR: ASSOC. PROF. DR. MEHMET GÖKTÜRK

SEPTEMBER 2025

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

İNSAN-YZ İŞBİRLİĞİNİN
YZ UYGULAMALARI ÜZERİNDEKİ ETKİSİ

BURHANETTİN BURUN

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMAN: DOÇ. DR. MEHMET GÖKTÜRK

EYLÜL 2025



MASTER of SCIENCE JURY APPROVAL FORM

A thesis submitted by Burhanettin BURUN, defended on 26/09/2025 before the jury formed with the 30/06/2025 date and 2025/34 numbered decision of the GTU Graduate Administration Board, has been accepted as a MASTER of SCIENCE thesis in the Department of Computer Engineering.

JURY

MEMBER

(THESIS ADVISOR) : Assoc. Prof. Dr. Mehmet GÖKTÜRK

MEMBER : Asst. Prof. Dr. Burcu YILMAZ

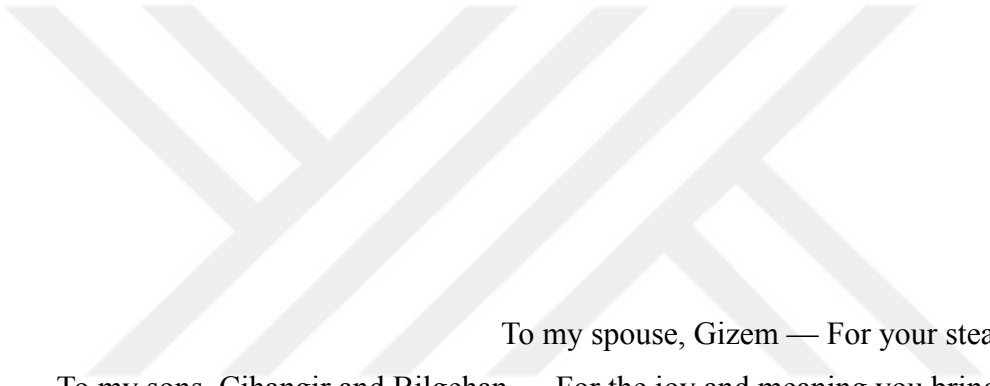
MEMBER : Prof. Dr. Mehmet Fatih AMASYALI

APPROVAL

Gebze Technical University Graduate Administration Board

...../...../..... date and/..... numbered decision.

SIGNATURE/SEAL



To my spouse, Gizem — For your steady support.

To my sons, Cihangir and Bilgehan — For the joy and meaning you bring to my life.

ABSTRACT

This thesis explores an innovative and interactive approach to credit card fraud detection by leveraging the strengths of Human-AI collaboration. In recent years, the increasing sophistication of fraudulent techniques and the massive volume of online transactions have posed serious challenges to conventional fraud detection systems. Traditional rule-based or static machine learning models often fail to adapt to the dynamic and evolving nature of fraud patterns. To address these limitations, this study proposes a dynamic fraud detection framework that integrates Long Short-Term Memory (LSTM) neural networks with the SHapley Additive exPlanations (SHAP) interpretability framework, supported by continuous user feedback.

The core of the system lies in the time-series modeling capability of LSTM networks, which are particularly well-suited for analyzing sequential transaction data and identifying subtle temporal patterns that may indicate fraud. However, the effectiveness of deep learning models is often limited by their “black box” nature, which makes it difficult for users or domain experts to understand how and why a specific prediction was made. To tackle this issue, the SHAP framework is employed to provide interpretable explanations for each model output, offering a breakdown of how individual features contribute to the prediction. This not only fosters transparency but also enables users—such as fraud analysts—to trust and critically assess the model’s decisions.

A key innovation in this thesis is the integration of user feedback into the learning loop. Users can evaluate the system’s predictions and provide corrective input when a transaction is misclassified. This feedback is then incorporated into the model through an iterative retraining process, allowing it to adapt and improve over time. Such a Human-in-the-Loop approach not only enhances model accuracy but also aligns the system with real-world operational needs, where expert insight is invaluable.

The experimental results obtained from real-world transaction datasets show that the proposed system achieves a significant improvement in performance metrics—especially in the F1 score—after incorporating human feedback. This indicates that continuous learning from user interactions enables the model to become more robust and responsive to evolving fraud strategies.

Overall, this study demonstrates the practical benefits of combining explainable artificial intelligence with interactive human feedback in fraud detection applications. The proposed framework not only increases the detection accuracy of suspicious transactions but also enhances end-user confidence through transparent and interpretable decision-making processes. As a result, the system is not merely a predictive tool, but a collaborative partner for human experts in the ongoing fight against financial fraud. This approach contributes to a broader understanding of Human-AI collaboration in critical domains and provides a foundation for future systems that blend machine learning capabilities with domain expertise in real-time decision-making environments.

Keywords: Credit Card Fraud, LSTM, SHAP, Human-AI Collaboration, Continuous Learning, Explainable Artificial Intelligence

ÖZET

Bu tez çalışması, günümüz dijital dünyasında giderek artan kredi kartı dolandırıcılığına karşı etkili bir çözüm sunmayı amaçlayan, İnsan-Yapay Zeka (YZ) işbirliğine dayalı yenilikçi bir yaklaşımı ortaya koymaktadır. Finansal işlemlerin hacmindeki hızlı artış, işlem verilerinin karmaşık yapısı ve dolandırıcılık yöntemlerinin sürekli değişim göstermesi, mevcut geleneksel sistemlerin yetersiz kalmasına neden olmaktadır. Kural tabanlı yaklaşımlar ve statik makine öğrenmesi modelleri, dolandırıcıların yaratıcı yöntemlerine karşı yeterince esnek ve adaptif olamamaktadır. Bu bağlamda, bu tezde önerilen sistem, derin öğrenme tekniklerinden Uzun-Kısa Süreli Bellek (Long Short-Term Memory – LSTM) ağırları ile açıklanabilir yapay zeka yaklaşımlarından SHapley Additive exPlanations (SHAP) çerçevesini bir araya getirerek, kullanıcı geri bildirimini ile sürekli öğrenen bir yapay zeka tabanlı dolandırıcılık tespit sistemi geliştirmeyi hedeflemektedir.

Önerilen sistemin temel bileşenlerinden biri olan LSTM ağırları, zaman serisi şeklinde kaydedilen kredi kartı işlemleri arasındaki bağıntıları ve örüntüleri öğrenmekte oldukça etkilidir. Geleneksel sinir ağlarından farklı olarak, LSTM modelleri, uzun vadeli bağımlılıkları da dikkate alabilmekte, bu sayede dolandırıcılığa işaret eden daha karmaşık davranış kalıplarını ortaya çıkarabilmektedir. Kredi kartı dolandırıcılığı çoğu zaman küçük ve dikkat çekmeyen işlemlerle başlar, dolayısıyla bu tür desenlerin yakalanması oldukça önemlidir. Ancak, derin öğrenme modelleri genellikle “kara kutu” (black box) yapılar olarak değerlendirilmekte ve neden belirli bir tahminde bulunduklarını açıklamakta yetersiz kalmaktadır.

İşte bu noktada SHAP çerçevesi devreye girmektedir. SHAP, her bir özelliğin model kararına olan katkısını sayısal olarak ifade eden, güçlü ve kuramsal temellere dayanan bir açıklanabilirlik yöntemidir. SHAP sayesinde sistemin verdiği her karar, kullanıcıya neden o şekilde sınıflandırma yaptığına dair anlaşılır ve güvenilir açıklamalarla sunulmaktadır. Özellikle finans sektöründe çalışan risk analistleri ve denetim uzmanları için bu tür açıklanabilirlik oldukça değerlidir, çünkü bu sayede kararların sadece yapay zekaya bırakılması yerine, insan denetimiyle birlikte ilerlenmesi mümkün olmaktadır.

Bu çalışmanın en önemli katkılarından biri de İnsan-Yapay Zeka etkileşimini merkeze almasıdır. Geliştirilen sistem, kullanıcıların modelin yaptığı tahminlere geri bildirimde bulunmasına olanak tanımaktadır. Örneğin, bir işlem model tarafından “şüpheli” olarak sınıflandırılmış ancak kullanıcı bunu yanlış bulmuşsa, bu bilgi sisteme geri beslenerek modelin bir sonraki eğitim sürecinde bu geri bildirim dikkate alınmaktadır. Bu sürekli öğrenme döngüsü, sistemin zamanla daha doğru ve kullanıcı ihtiyaçlarına daha uygun hale gelmesini sağlamaktadır.

Sonuç olarak bu tez, kredi kartı dolandırıcılığı gibi gerçek zamanlı ve kritik bir problem alanında, İnsan-Yapay Zeka işbirliğine dayanan bütüncül bir çözüm yaklaşımı sunmaktadır. Derin öğrenme, açıklanabilir yapay zeka ve insan etkileşiminin entegre edildiği bu sistem, yalnızca tespit performansını artırmakla kalmamakta, aynı zamanda kullanıcı güvenini güçlendirmekte ve sistemin sürdürülebilirliğini sağlamaktadır. Gelecekte benzer sistemlerin finans dışındaki alanlarda da uygulanabilirliğinin araştırılması, İnsan-Yapay Zeka işbirliğinin potansiyelinden daha geniş ölçekte yararlanılmasına olanak sağlayacaktır.

Anahtar Kelimeler:Kredi Kartı Dolandırıcılığı, LSTM, SHAP, İnsan-Yapay Zeka İşbirliği, Sürekli Öğrenme, Açıklanabilir Yapay Zeka

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Assoc. Prof. Mehmet Göktürk, for their invaluable guidance, continuous support, and insightful feedback throughout this research journey. Their expertise and encouragement have been instrumental in shaping this thesis and my academic development.

I am grateful to the Department of Computer Engineering at Gebze Technical University for providing an excellent academic environment and resources that facilitated this research. Special thanks to the faculty members and administrative staff for their assistance throughout my graduate studies.

I am deeply indebted to my family for their unwavering support, patience, and encouragement throughout my academic journey. Their belief in me has been a constant source of motivation. Finally, I would like to thank all those who directly or indirectly contributed to the completion of this thesis.



TABLE OF CONTENTS

	Page
ABSTRACT	vi
ÖZET	vii
ACKNOWLEDGEMENT	viii
TABLE of CONTENTS	ix
LIST OF SYMBOLS AND ABBREVIATIONS	xiii
LIST of FIGURES	xiv
LIST of TABLES	xv
1. INTRODUCTION	1
1.1. Research Background	1
1.2. Problem Statement	2
1.3. Research Objectives and Significance	2
1.4. Research Questions and Hypotheses	3
1.5. Scope and Limitations of the Research	4
1.6. Thesis Structure	5
2. LITERATURE REVIEW	7
2.1. Credit Card Fraud: An Overview	7
2.1.1. Definition and Types of Credit Card Fraud	7
2.1.2. Global Impact of Credit Card Fraud	8
2.2. Traditional Fraud Detection Methods	9
2.2.1. Rule-Based Systems	9
2.2.2. Statistical Methods	9
2.2.3. Expert Systems	10
2.3. Machine Learning-Based Approaches	10
2.3.1. Supervised Learning Methods	10
2.3.2. Unsupervised Learning Techniques	11
2.3.3. Semi-Supervised and Hybrid Approaches	11
2.4. Deep Learning and LSTM Networks	12
2.4.1. Neural Networks in Fraud Detection	12
2.4.2. Recurrent Neural Networks (RNNs)	12
2.4.3. LSTM Networks: Architecture and Advantages	12
2.4.4. Applications of LSTM in Fraud Detection	13
2.5. Explainable Artificial Intelligence (XAI)	14
2.5.1. The Need for Explainability in Fraud Detection	14
2.5.2. XAI Techniques and Frameworks	14
2.5.3. SHAP Framework	15
2.5.4. Challenges in Applying XAI to Deep Learning Models	15
2.6. Human-AI Collaboration	16
2.6.1. Human-in-the-Loop (HITL) Methods	16
2.6.2. Benefits of Human-AI Collaboration in Financial Services	16
2.6.3. Challenges in Implementing Human-AI Collaboration	17
2.6.4. Existing Applications in Fraud Detection	18
2.7. Open Research Questions And Opportunities	18
3. METHODOLOGY	20
3.1. Research Design	20

3.2.	Dataset Description and Preprocessing	21
3.2.1.	Dataset Overview	21
3.2.2.	Feature Description	21
3.2.3.	Data Exploration	21
3.2.4.	Feature Engineering	23
3.2.5.	Data Preprocessing	23
3.3.	LSTM Model Architecture	24
3.3.1.	Model Design Considerations	24
3.3.2.	LSTM Layer Configuration	24
3.3.3.	Model Training Strategy	26
3.3.4.	Hyperparameter Tuning	26
3.4.	SHAP Integration for Explainability	27
3.4.1.	SHAP Framework Overview	27
3.4.1.1.	Deep Explainer for LSTM Models	27
3.4.1.2.	Deep Explainer Technical Details	27
3.4.1.3.	Generating Explanations	29
3.4.1.4.	Visualization Implementation	30
3.4.2.	Comparison with Alternative XAI Methods	30
3.4.2.1.	Alternative Methods Considered	30
3.4.2.2.	Comparison Criteria	31
3.4.2.3.	Comparison Results	31
3.4.2.4.	Selection Rationale	31
3.5.	Human-AI Collaboration Mechanism	32
3.5.1.	Feedback Loop Design	32
3.5.2.	User Interface for Interaction	32
3.5.3.	Continuous Learning Implementation	33
3.6.	System Architecture and Implementation	34
3.6.1.	Overall System Architecture	34
3.6.2.	Technology Stack	35
3.6.3.	Implementation Details	35
3.6.4.	Experimental Setup and Evaluation Metrics	36
3.6.4.1.	Experimental Design	36
3.6.4.2.	Evaluation Metrics	36
3.6.4.3.	User Study Design	37
4.	EXPERIMENTAL RESULTS	38
4.1.	Model Performance	38
4.1.1.	Training Results	38
4.1.2.	Test Set Evaluation	39
4.1.3.	Confusion Matrix Analysis	40
4.2.	SHAP Analysis	41
4.2.1.	Global Feature Importance	41
4.2.2.	Local Explanations for Individual Transactions	42
4.2.2.1.	Example 1: Fraud Transaction	42
4.2.2.2.	Example 2: Legitimate Transaction	45
4.2.2.3.	Comparison of Fraud vs. Legitimate	46
4.3.	User Study Results	47
4.3.1.	Participant Demographics	47
4.3.2.	Task Performance	47

4.3.3.	Qualitative Feedback	49
4.3.4.	System Usability	50
4.4.	Performance Improvement After Feedback	50
4.4.1.	Model Retraining with Feedback	50
4.4.2.	Performance Improvement Analysis	51
4.4.3.	Continuous Learning Benefits	52
5.	DISCUSSION	54
5.1.	Contribution of Human-AI Collaboration	54
5.1.1.	Synergistic Relationship Between Human Expertise and AI Capabilities	54
5.1.2.	Impact on Different Types of Fraud	55
5.1.3.	Continuous Learning and Adaptation	55
5.2.	Role of Explainability	56
5.2.1.	Enhancing Trust and Transparency	56
5.2.2.	Challenges in Generating Meaningful Explanations	56
5.3.	Model Performance Comparisons	57
5.3.1.	LSTM vs. Traditional Machine Learning Methods	57
5.3.2.	Comparison with Other Deep Learning Architectures	58
5.3.3.	Impact of Human Feedback Across Models	58
5.4.	System Limitations and Challenges	58
5.4.1.	Computational Complexity	58
5.4.2.	Data Privacy and Security Concerns	59
5.4.3.	Integration with Existing Systems	59
5.4.4.	Human Factors and Scalability	60
5.5.	Practical Application Assessment	60
5.5.1.	Cost-Benefit Analysis	60
5.5.2.	Implementation Strategies	61
5.5.3.	Generalizability to Other Financial Fraud Types	61
5.5.4.	Ethical Challenges	62
6.	CONCLUSION AND FUTURE WORK	64
6.1.	Research Summary	64
6.2.	Main Findings	65
6.2.1.	Effectiveness of LSTM Networks	65
6.2.2.	Impact of SHAP Explanations	65
6.2.3.	Effect of Human Feedback	65
6.2.4.	Human-AI Collaboration for Rare Patterns	66
6.2.5.	Balancing Explainability and Performance	66
6.3.	Research Contributions	66
6.3.1.	Theoretical Contributions	67
6.3.2.	Practical Contributions	67
6.4.	Future Research Directions	68
6.4.1.	Advanced Model Architectures	68
6.4.2.	Enhanced Explainability	68
6.4.3.	Human-AI Interaction	69
6.4.4.	Real-world Deployment and Evaluation	69
6.4.5.	Extension to Other Domains	70
6.5.	Final Assessment	70

REFERENCES	72
BIOGRAPHY	75
PUBLICATIONS AND PRESENTATIONS FROM THE THESIS	76
APPENDICES	77



LIST OF SYMBOLS AND ABBREVIATIONS

AI	:	Artificial Intelligence
ATO	:	Account Takeover
AUC-PR	:	Area Under the Precision-Recall Curve
AUC-ROC	:	Area Under the Receiver Operating Characteristic Curve
CNN	:	Convolutional Neural Network
CNP	:	Card-Not-Present
DL	:	Deep Learning
GDPR	:	General Data Protection Regulation
HITL	:	Human-in-the-Loop
LSTM	:	Long Short-Term Memory
ML	:	Machine Learning
MLP	:	Multilayer Perceptron
RNN	:	Recurrent Neural Network
SHAP	:	SHapley Additive exPlanations
SUS	:	System Usability Scale
SVM	:	Support Vector Machine
XAI	:	Explainable Artificial Intelligence

LIST OF FIGURES

	<u>Page</u>
Figure 3.1: LSTM Model Architecture	25
Figure 3.2: Continuos Learning Process	33
Figure 3.3: System Architecture Flow Diagram	34
Figure 4.1: Model Loss Over Epochs	39
Figure 4.2: Shap Summary Plot	42
Figure 4.3: Fraud Transaction	44
Figure 4.4: Legitimate Transaction	46
Figure 4.5: Performance Metric Comparison	49
Figure 4.6: User Study Comparison	53
Figure E.1: Transaction List	82
Figure E.2: Transaction Detail	82

LIST OF TABLES

	<u>Page</u>
Table 3.1: Feature Descriptions and Examples	22
Table 3.2: Engineered Features	23
Table 3.3: XAI Methods Comparison	31
Table 4.1: Model Performance on Test Set	39
Table 4.2: Confusion Matrix	40
Table 4.3: Global Feature Importance (SHAP)	41
Table 4.4: SHAP Values for Fraud Transaction	43
Table 4.5: SHAP Values for Legitimate Transaction	45
Table 4.6: Comparison of Fraud vs. Legitimate Transactions	47
Table 4.7: User Performance Comparison	48
Table 4.8: System Usability Scale (SUS) Scores	50
Table 4.9: Performance Before and After Feedback	51

1. INTRODUCTION

1.1. Research Background

The proliferation of digital transactions has led to a significant increase in credit card fraud cases. According to recent research, global losses from credit card fraud are expected to exceed \$40 billion annually by 2027 [1]. This situation creates serious financial and security problems for both consumers and financial institutions. Traditional fraud detection systems generally rely on static rule-based approaches, which are inadequate in addressing the increasingly complex and constantly changing nature of fraudulent activities [2].

In recent years, the utilization of machine learning (ML) and deep learning (DL) methods has become increasingly prominent in combating fraud [3]. In particular, Long Short-Term Memory (LSTM) networks have attracted attention with their superior performance in capturing temporal dependencies in transaction data [4]. LSTM networks can model long-term dependencies by solving the "gradient vanishing" problem encountered by traditional Recurrent Neural Networks (RNNs). This feature provides an ideal solution for detecting abnormal behaviors in credit card transactions.

However, the "black box" nature of deep learning models limits their adoption in applications where explainability is critical, such as fraud detection [5]. To overcome this challenge, Explainable Artificial Intelligence (XAI) frameworks such as SHAP (SHapley Additive exPlanations) have been developed. These frameworks provide interpretable insights into model predictions [6]. This study aims to bridge the gap between model accuracy and interpretability by integrating SHAP with an LSTM-based fraud detection system.

Another critical aspect of the study is the inclusion of Human-AI Collaboration through the Human in the Loop (HITL) approach. Although artificial intelligence models have made significant progress in processing large volumes of data, human expertise still plays a critical role in validating predictions and identifying edge cases [7]. The proposed web application allows users to review and provide feedback on model predictions, enabling continuous retraining and improvement of the system. This iterative

process not only enhances model performance but also enables adaptation to emerging fraud patterns over time.

1.2. Problem Statement

Credit card fraud detection presents a complex and continuously evolving challenge for financial institutions. Traditional rule-based systems are inadequate in adapting to the dynamic nature of fraud patterns and require more advanced solutions. The current challenges include:

1. **Data Imbalance:** Fraudulent cases occur much less frequently than normal transactions in credit card operations. This makes it difficult for models to learn fraud patterns.
2. **Dynamic Fraud Patterns:** Fraud methods are continuously evolving and changing, causing static models to become ineffective over time.
3. **Lack of Explainability:** Deep learning models are often characterized as "black boxes," meaning the logic behind their decisions can be difficult to understand. This undermines user confidence, especially in critical areas such as financial security.
4. **Integration of Human Expertise:** Existing systems often fail to effectively integrate human expertise and feedback, resulting in the loss of valuable insights.
5. **Real-time Detection Requirement:** Fraud detection must be performed in real-time to prevent financial losses, but complex models can be computationally intensive.

To address these problems, an integrated solution is needed that provides both high accuracy and interpretability, while being continuously improvable through human feedback.

1.3. Research Objectives and Significance

We also want to look closely at how well humans and AI work together to find fraud. We will show how this team effort is better than older ways of finding fraud. Our study will explain how combining human knowledge with AI power can lead to much better

results.

Finally, this research will provide a practical guide for how to use this system. It will show how to set up a fraud detection system that works in real-time, can explain itself, and can adapt to new situations. This part of the research shows that our ideas can actually be used to solve real-world problems.

This research is very important for several reasons. First, it helps with financial security. Credit card fraud costs banks and people billions of dollars every year. Our system can help reduce these huge losses, making money safer for everyone. Second, it helps technology move forward. Our study brings together the newest ideas in deep learning (a type of AI), explainable AI (AI that can explain itself), and humans working with machines. This mix of ideas can lead to new discoveries and ways of doing things. Third, it has real-world uses. Our research will give banks and financial companies practical tools and plans to make their fraud detection systems better. This means our findings aren't just theories; they can be put into action to solve actual problems. Fourth, it adds to what we know in academics. Our study will add valuable information to the growing field of explainable AI and how humans and machines can work together. It will be a useful resource for future studies and new developments.

Lastly, this research helps with following rules. Many government groups now require that AI systems, especially those that affect people's money, must be able to explain their decisions. Our research focuses on building AI that can explain itself, which will help financial companies meet these important rules and build more trust in AI systems.

1.4. Research Questions and Hypotheses

This study aims to answer these main questions:

1. Are LSTM networks better than traditional machine learning methods for detecting credit card fraud?
2. How effective is the SHAP framework in explaining the decisions of deep learning models, and how do these explanations affect user confidence?
3. How does the human feedback and continuous learning mechanism affect the performance of fraud detection models over time?

4. To what extent does Human-AI collaboration benefit the detection of rare fraud patterns?
5. How can the balance between explainability and high accuracy be optimized in the context of financial fraud detection?

The hypotheses of the research are:

H1: LSTM-based models will provide higher accuracy and lower false alarm rates compared to traditional machine learning methods, due to their ability to capture temporal dependencies in credit card transactions.

H2: Explainability mechanisms integrated with the SHAP framework will significantly increase user confidence by enhancing the transparency of model decisions.

H3: The continuous learning mechanism based on user feedback will improve model performance over time and result in a significant increase in F1 score.

H4: Human-AI collaboration is more effective in detecting rare and complex fraud patterns compared to artificial intelligence approaches alone.

H5: There is a trade-off between explainability and model performance, but an optimal balance between these two goals can be achieved with appropriate techniques.

1.5. Scope and Limitations of the Research

This research focuses on building a system where humans and artificial intelligence (AI) work together to spot credit card fraud. The study covers several key areas. First, it involves analyzing data. The study uses a dataset of credit card transactions from Europe, specifically from September 2013. This dataset is special because it shows how real-world fraud data often looks: there are many more normal transactions than fraudulent ones. This imbalance makes it a good test for the system. Second, the research includes developing the fraud detection model itself. This model uses something called LSTM networks, which are very good at understanding patterns that change over time. This helps the system learn from past transactions to predict future fraud. Third, a big part of the study is making sure the model's decisions can be explained. To do this, the SHAP framework is used. SHAP helps us understand why the AI flagged a certain transaction as potentially fraudulent, making the system more transparent. Fourth, the

research looks at how humans and AI can truly collaborate. It designs and puts into practice a system that continuously learns from feedback given by users. This means the system gets smarter and more accurate over time with human input. Fifth, the study carefully checks how well the new system performs. It uses different ways to measure its success and compares it to older, more traditional methods of detecting fraud. This helps show if the new approach is better.

Finally, to make the system practical, a web application is built. This web interface allows people to easily use and interact with the fraud detection system. However, this research also has some limitations. Training these advanced AI models and calculating SHAP values needs a lot of computing power. This can be a challenge for systems that need to work very quickly in real-time. Furthermore, the study uses simulated user feedback instead of real feedback from actual users. This means the system's effectiveness in a real-world setting might be slightly different than what was observed in the study. Lastly, fraud patterns are always changing. This means that even if the model works well now, it might need regular updates in the future to keep up with new ways people try to commit fraud. Despite these challenges, the study clearly shows the great potential of humans and AI working together to fight credit card fraud. It also provides valuable ideas and directions for future research in this important area.

1.6. Thesis Structure

This thesis consists of the following chapters:

Chapter 1: Introduction - Presents the research background, problem definition, objectives and significance, research questions, hypotheses, scope, and limitations.

Chapter 2: Literature Review - Provides a comprehensive literature review on credit card fraud, traditional detection methods, machine learning and deep learning approaches, explainable artificial intelligence, and human-artificial intelligence collaboration.

Chapter 3: Methodology - Details the research design, dataset, LSTM model architecture, SHAP integration, human-artificial intelligence collaboration mechanism, system architecture, and experimental setup.

Chapter 4: Experimental Results - Presents the initial model performance, SHAP analysis results, performance improvement after user feedback, comparative analysis,

and user experience evaluation.

Chapter 5: Discussion - Discusses the contribution of human-artificial intelligence collaboration, the role of explainability, model performance comparisons, system limitations and challenges, and practical application assessment.

Chapter 6: Conclusion and Future Work - Summarizes the research, presents the main findings, contributions, future research suggestions, and final evaluation.

Each chapter addresses different aspects of Human-AI collaboration in credit card fraud detection and provides a holistic understanding. The thesis aims to contribute to academic literature while offering practical solutions for financial institutions.



2. LITERATURE REVIEW

This chapter offers an in-depth analysis of prior studies concerning credit card fraud detection. First, a general definition, types, and global impacts of credit card fraud are addressed. Then, traditional fraud detection methods and their limitations are discussed. Following this, applications of machine learning and deep learning techniques in fraud detection, particularly the role of LSTM networks, are detailed. The concept of Explainable Artificial Intelligence (XAI) and the importance of frameworks such as SHAP are emphasized. Finally, the potential of Human-AI Collaboration and Human-in-the-Loop (HITL) approaches in financial services is examined, and research gaps in the literature are identified.

2.1. Credit Card Fraud: An Overview

2.1.1. Definition and Types of Credit Card Fraud

Credit card fraud is defined as the illegal acquisition or use of another person's credit card information to obtain unauthorized financial gain [8]. This type of fraudulent activity poses a serious threat to financial systems and leads to significant losses for both cardholders and financial institutions. Fraud methods are continuously evolving and diversifying. The main types of credit card fraud include:

- **Lost or Stolen Card Fraud:** The unauthorized use of a physically lost or stolen card.
- **Card-Not-Present (CNP) Fraud:** Fraud committed in transactions where the card is not physically present, such as internet, telephone, or mail orders, using stolen card information. With the proliferation of e-commerce, this has become one of the most common types [9].
- **Counterfeit Card Fraud:** The production of fake physical cards using copied or stolen card information and conducting transactions through POS devices or ATMs.
- **Identity Theft Fraud:** Fraudsters obtaining personal information of others (name,

address, social security number, etc.) and applying for credit cards in their name or taking over existing accounts.

- **Account Takeover (ATO) Fraud:** Fraudsters gaining unauthorized access to legitimate accounts by stealing credentials or through social engineering techniques.
- **Application Fraud:** Applying for credit cards using false or stolen identities.
- **Merchant Fraud:** Dishonest merchants using customer card information for unauthorized transactions.

2.1.2. Global Impact of Credit Card Fraud

Recent forecasts suggest that worldwide financial losses resulting from card fraud may surpass \$40 billion by 2027 [1]. The United States alone accounts for approximately 38.6% of global card fraud losses, despite representing only 22.9% of total card volume [10].

Beyond direct financial losses, credit card fraud has several indirect impacts:

- **Operational Costs:** Financial institutions incur significant costs in fraud detection, prevention, and resolution.
- **Reputational Damage:** Banks and merchants suffering from high fraud rates may experience reputational damage and customer attrition.
- **Customer Experience:** Legitimate transactions may be declined due to overly strict fraud controls, negatively affecting customer experience.
- **Economic Impact:** The overall economy is affected through increased costs of goods and services as merchants pass on fraud-related expenses to consumers.

The COVID-19 pandemic has further accelerated the shift to digital payments, creating new opportunities for fraudsters and increasing the urgency for effective fraud detection systems. [11].

2.2. Traditional Fraud Detection Methods

Traditional fraud detection systems typically rely on rule-based approaches, statistical methods, and expert systems. These methods use predefined rules and threshold values to identify specific fraud patterns. However, due to the continuous evolution and increasing complexity of fraud methods, these traditional approaches have become inadequate over time.

2.2.1. Rule-Based Systems

Rule-based systems work by following a set of pre-defined rules that experts have created. For example, such a system might flag a transaction as suspicious if it's for a very large amount of money, comes from a country the user doesn't usually buy from, or happens at an unusual time of day for that user. While these systems are easy to set up and understand, they have some important drawbacks. One major issue is that they need to be updated by hand all the time, especially as new ways of committing fraud appear. They also struggle to catch fraud patterns that haven't been seen before. This means if a fraudster comes up with a completely new trick, a rule-based system might miss it. Another problem is that they often create a lot of false positives. This means they incorrectly flag many legitimate transactions as fraud, which can be annoying for users and costly for businesses. Finally, these systems can't adjust to how real users' spending habits change over time. They are rigid and don't learn from new behavior.

2.2.2. Statistical Methods

Statistical methods for detecting fraud use mathematical techniques like finding unusual data points (outlier detection), grouping similar data together (clustering), or analyzing relationships between data (regression analysis). These methods build a picture of what normal behavior looks like. If a transaction doesn't fit this normal picture, it's flagged as possibly fraudulent. Compared to simple rule-based systems, statistical approaches offer some benefits. They are good at handling numerical information effectively. They can also figure out what typical behavior patterns are, which helps them spot things that are out of the ordinary. They have the potential to find small,

subtle differences that might indicate fraud. However, even statistical methods face challenges. They still find it hard to keep up with how fraud changes all the time. Also, to work well, they often need a lot of careful preparation of the data, which is called feature engineering. This can be a time-consuming and complex process.

2.2.3. Expert Systems

Expert systems try to put human knowledge and smart ways of thinking into a set of rules. These systems can be more advanced than simple rule-based systems because they include specific knowledge about a topic and can make more complex decisions. However, expert systems have their own problems. It's hard to get all the important knowledge from human experts and put it into the system. Also, keeping this knowledge up-to-date and managing it can be difficult. These systems can't learn new things on their own from new information, and they don't easily adapt to new ways of fraud that appear. Because of these weaknesses in older methods, there's been a growing trend to use machine learning and deep learning. These newer approaches are much better at adapting to changes and recognizing patterns, which makes them more effective for tasks like fraud detection.

2.3. Machine Learning-Based Approaches

Recently, fraud detection has increasingly leveraged machine learning techniques, which can recognize complex data patterns and adjust to evolving fraudulent behaviors. Numerous ML algorithms have yielded encouraging outcomes in this field.

2.3.1. Supervised Learning Methods

Supervised learning methods train on labeled data, where transactions are marked as either fraudulent or legitimate. Common algorithms include:

- **Logistic Regression:** Logistic Regression is a statistical technique used to compute the likelihood of fraudulent activity based on specific input variables. [12].
- **Decision Trees:** Tree-structured models that make decisions based on feature values. Decision trees are interpretable but may suffer from overfitting on imbal-

anced fraud data.

- **Random Forest:** Random Forest is an ensemble learning technique that aggregates numerous decision trees to enhance accuracy and mitigate overfitting, showing considerable effectiveness in fraud identification scenarios [2].
- **Support Vector Machines (SVM):** Algorithms that find an optimal hyperplane to separate fraudulent and legitimate transactions. SVMs can handle high-dimensional data but may struggle with large datasets.
- **Gradient Boosting Machines:** Gradient Boosting Machines, such as XGBoost and LightGBM, construct predictive models in sequence by rectifying previous errors and have achieved top-tier results in various fraud detection evaluations [13].

2.3.2. Unsupervised Learning Techniques

Unsupervised learning works without labeled data and detects anomalies that could signal fraud.

- **Clustering Algorithms:** Clustering methods (e.g., K-means, DBSCAN) group similar transactions, and outliers may indicate fraud.
- **Isolation Forest:** Isolation Forest is an anomaly detection method that isolates data points using random feature splits.
- **Autoencoders:** Neural network architectures that learn to compress and reconstruct data, with high reconstruction error potentially indicating fraud.

2.3.3. Semi-Supervised and Hybrid Approaches

Semi-supervised learning uses both labeled and unlabeled data, useful when fraud labels are scarce:

- **One-Class Classification:** Models trained only on legitimate transactions to identify deviations from normal behavior.
- **Positive-Unlabeled Learning:** Methods that learn from positive (fraudulent) examples and unlabeled data.

- **Hybrid Models:** Hybrid models combine different ML methods or mix ML with rule-based systems for better performance.

Despite their advantages, traditional machine learning approaches face challenges in capturing the temporal dynamics of transaction sequences and the evolving nature of fraud patterns, leading to the exploration of deep learning methods.

2.4. Deep Learning and LSTM Networks

As a branch of machine learning, deep learning utilizes layered neural network structures to effectively uncover intricate patterns within vast datasets, contributing significantly to fraud detection efforts.

2.4.1. Neural Networks in Fraud Detection

Multi-layered feedforward neural networks have been employed in detecting fraudulent activity, capturing non-linear associations between attributes and indicators, often surpassing traditional ML techniques.

2.4.2. Recurrent Neural Networks (RNNs)

RNNs address the limitation of standard neural networks by incorporating memory of previous inputs. This makes them suitable for sequential data such as transaction histories. Nonetheless, standard recurrent neural networks face limitations such as the vanishing gradient issue, which hampers their capacity to retain long-term information. [14].

2.4.3. LSTM Networks: Architecture and Advantages

LSTM networks, first introduced by Hochreiter and Schmidhuber, are a more advanced type of Recurrent Neural Networks (RNNs). They were specifically designed to solve a common problem in RNNs called the 'vanishing gradient' challenge, which makes it hard for them to learn from long sequences of data. At the heart of an LSTM network are special memory units called 'cells.' These cells have unique 'gates' that control how information flows in and out of their memory. There are three main types of these

gates: First, the Forget Gate decides what information is no longer needed and should be removed from the cell's memory. This helps keep the memory relevant and prevents it from getting cluttered with old, useless data. Second, the Input Gate is responsible for adding new, important information to the cell's memory. It determines which new data points are valuable enough to be stored. Third, the Output Gate controls what information from the cell's memory should be passed on to the next parts of the network. It decides what is relevant for the current prediction or task. These gates allow LSTMs to effectively remember or forget data over long periods. This ability makes them exceptionally good at understanding patterns in sequences, such as a series of financial transactions, where the order and timing of events are crucial. Using LSTMs for detecting fraud offers several significant advantages. They are excellent at identifying long-term spending habits of users, which is vital for spotting unusual behavior. They can also accurately model the sequential nature of transactions, understanding that the order in which purchases are made can be important for fraud detection. Furthermore, LSTMs are robust, meaning they can handle situations where there are varying time gaps between transactions without losing their effectiveness. Finally, their ability to adapt to changing fraud patterns makes them a powerful tool in the ongoing fight against financial crime.

2.4.4. Applications of LSTM in Fraud Detection

Many studies show that LSTMs work well for credit card fraud detection:

- Roy et al. [15] developed a hybrid LSTM-based autoencoder for anomaly detection in credit card transactions, achieving higher precision and recall compared to traditional methods.
- Jurgovsky et al. [16] showed that sequence-aware LSTM models outperformed transaction-independent classifiers by incorporating temporal information.
- Fu et al. [17] proposed a dual-LSTM architecture that simultaneously models both global and local transaction patterns, demonstrating superior performance on real-world credit card data.

Despite their powerful modeling capabilities, LSTM networks, like other deep learning models, suffer from limited interpretability, which is a critical concern in financial

applications.

2.5. Explainable Artificial Intelligence (XAI)

The complex structure of deep learning models makes understanding their decisions challenging. This "black box" problem presents a significant barrier, especially in regulated sectors like financial services. The domain of Explainable Artificial Intelligence (XAI) focuses on enhancing the transparency and interpretability of decisions produced by machine learning models.

2.5.1. The Need for Explainability in Fraud Detection

It's really important for fraud detection systems to be able to explain their decisions. There are several good reasons for this. First, there's Regulatory Compliance. Banks and financial companies often have rules that say they must explain why an automated system made a certain decision, especially if it affects a customer. For example, if a transaction is blocked, they need to be able to say why. Second, it helps Build Trust. When a system is clear about how it works, people, including customers and other important groups, trust it more. If you understand why something happened, you're more likely to accept it. Third, it helps Identify Errors. When a system explains its reasoning, it can help us find its weak points or any unfair biases it might have. This allows us to make the system better and fairer. Fourth, it leads to Knowledge Discovery. By understanding how the system makes its decisions, we can learn new things about how fraud patterns work. This can give us valuable insights we didn't have before. Finally, it improves Operational Efficiency. When fraud analysts get clear explanations for why a transaction was flagged, they can make decisions much faster and with more confidence. This makes their job easier and more effective.

2.5.2. XAI Techniques and Frameworks

Many different methods have been created to help us understand how machine learning models make their decisions. Some tools, like 'permutation importance' and LIME, help us figure out which parts of the data (features) were most important for a prediction. Other methods, like 'attention modules,' show us what parts of the data the model was

focusing on. There are also 'counterfactual methods' that tell us what small changes to the input data would have changed the model's outcome.

2.5.3. SHAP Framework

SHAP, which stands for SHapley Additive exPlanations, is a special method introduced by Lundberg and Lee. It's based on game theory and provides a consistent way to figure out how much each piece of information (feature) contributed to a single prediction made by a model. SHAP has a strong Theoretical Foundation rooted in game theory, giving it a solid mathematical basis for explaining why certain features are important. It offers both Local and Global Interpretability, meaning it can explain why a specific prediction was made (local) and also how the model behaves overall (global). SHAP values are known for their Consistency, meeting important standards for accuracy and reliability. What's more, SHAP is Model Agnostic, which means it can work with any type of machine learning model, not just specific ones. When it comes to detecting fraud, SHAP values are very useful. They can help explain exactly why a particular transaction was marked as fraudulent. They show which characteristics of that transaction (like the amount, location, or time) were most important in the system's decision. This transparency is helpful for both fraud analysts, who need to review these cases, and for customers, who might wonder why their transaction was declined.

2.5.4. Challenges in Applying XAI to Deep Learning Models

Even with all the progress in XAI, there are still some difficulties when trying to apply these explanation techniques to deep learning models, especially for fraud detection. One challenge is the Computational Load. Getting accurate SHAP values from complex deep learning models often requires a lot of computing power, which can be a problem for systems that need to work very fast. Another issue is Temporal Explanations. When dealing with models that look at sequences of data, like LSTMs, explaining their predictions means we also have to consider how things change over time. This makes explanations more complex. Then there are Feature Interactions. Deep learning models can find very complicated relationships between different pieces of data. It can be hard to fully understand and clearly explain these complex interactions. Finally,

there's often a Performance vs. Interpretability Trade-off. Sometimes, making a model easier to understand can mean it's slightly less accurate in its predictions. Finding the right balance between being transparent and being highly accurate is a key challenge. Solving these problems is crucial for building fraud detection systems that are both accurate and easy to understand.

2.6. Human-AI Collaboration

Human-AI Collaboration represents an approach that combines the strengths of artificial intelligence with human expertise. This approach acknowledges that although AI is adept at analyzing extensive datasets and discerning trends, human insight contributes context, ethics, and flexibility in unfamiliar scenarios.

2.6.1. Human-in-the-Loop (HITL) Methods

Human-in-the-Loop (HITL) refers to a Human-AI collaboration model where user input is actively embedded into the learning and adjustment phases of the AI framework. In HITL systems:

- AI models make initial predictions or recommendations
- Humans review these outputs and provide corrections or feedback
- The system incorporates this feedback to improve future predictions
- This cycle continues, creating a continuous learning loop

The HITL methodology has been effectively implemented in numerous fields such as image recognition, text analysis, and personalized recommendation engines[18]. In the context of fraud detection, HITL can help address the challenges of evolving fraud patterns and the need for high accuracy in critical decisions.

2.6.2. Benefits of Human-AI Collaboration in Financial Services

When human experts work together with AI in financial services, it brings many good things. This teamwork helps in several ways. First, it leads to Improved Accuracy. When humans give feedback, they can help correct mistakes the AI might make, which makes the whole system work better and more accurately. Second, the system gains

Adaptability. With guidance from humans, the system can quickly learn and adjust to new ways of fraud that appear. This means it stays effective even as fraud patterns change. Third, humans provide Contextual Understanding. Sometimes, data alone doesn't tell the whole story. Humans can add important background information, like knowing about special events or a customer's unique situation, which the AI might miss. Fourth, human involvement ensures Ethical Oversight. Having people in the loop helps make sure that decisions made by the system are fair and consider ethical issues. This is very important when dealing with people's money. Fifth, it helps Enhance Trust. When people know that there are humans overseeing the AI, they tend to trust the system more. This builds confidence among everyone involved. Finally, it allows for Continuous Improvement. The process of humans giving feedback to the AI creates a loop where the AI system keeps getting better and more refined over time.

2.6.3. Challenges in Implementing Human-AI Collaboration

Even though working with humans and AI has many benefits, putting it into practice can be tricky. There are several challenges to overcome. One challenge is Interface Design. It's hard to create computer screens and tools that make it easy and efficient for humans and AI to work together. The interface needs to be simple and clear. Another challenge is Feedback Integration. We need good ways to take the feedback given by humans and effectively use it to update and improve the AI model. This isn't always straightforward. Then there's Cognitive Load. Reviewing complex cases can be mentally demanding for human experts. We need to find ways to manage this mental strain, especially when dealing with many difficult situations. Scalability is also a concern. It's about finding the right balance between needing human input for decisions and the huge number of decisions that need to be made. Humans can't review every single transaction. Furthermore, there are Expertise Requirements. The human reviewers need to have enough knowledge about the subject (like finance and fraud) to give truly valuable and helpful feedback to the AI. Lastly, there's the issue of Bias Mitigation. We need to be careful that when humans are involved, they don't accidentally introduce or make existing human biases worse within the collaborative system. The goal is to reduce bias, not add to it.

2.6.4. Existing Applications in Fraud Detection

Several studies and commercial systems have explored Human-AI Collaboration in fraud detection:

- Bansal et al. [19] demonstrated that human-AI teams outperformed either humans or AI working alone in detecting financial fraud, with complementary strengths leading to superior overall performance.
- Some commercial fraud detection systems incorporate analyst feedback loops, though these are often limited in scope and do not fully leverage the potential of continuous learning.
- Many financial organizations adopt layered decision strategies, delegating standard cases to AI systems and reserving ambiguous or intricate evaluations for human experts.

However, there remains a gap in research regarding the systematic integration of human feedback with deep learning models, particularly in the context of explainable fraud detection systems.

2.7. Open Research Questions And Opportunities

Previous studies highlight unresolved challenges in credit card fraud detection:

1. **Limited Integration of Temporal Models with Explainability:** While LSTM networks have shown promise in fraud detection, and SHAP has emerged as a powerful explainability framework, there is limited research on their integration, particularly in addressing the challenges of explaining sequential predictions.
2. **Insufficient Human Feedback Mechanisms:** Most fraud detection systems lack structured mechanisms for incorporating human feedback into model improvement. The potential of continuous learning through human-AI collaboration remains largely unexplored in this domain.
3. **Balancing Real-time Performance with Explainability:** There is a need for approaches that balance the computational demands of complex models and explainability techniques with the real-time requirements of fraud detection sys-

tems.

4. **Adaptation to Evolving Fraud Patterns:** While the dynamic nature of fraud is widely acknowledged, there is limited research on how Human-AI collaborative systems can adapt to these evolving patterns over time.
5. **Practical Implementation Frameworks:** There is a gap between theoretical research and practical implementation, with few studies providing comprehensive frameworks for deploying Human-AI collaborative systems in real-world financial environments.

This research tackles these gaps by testing a system that uses LSTMs, SHAP, and human-AI teamwork for fraud detection.



3. METHODOLOGY

This chapter presents the methodological framework employed in this research to develop a Human-AI collaborative system for credit card fraud detection. This chapter starts by outlining the research methodology and then delves into an in-depth presentation of the dataset applied. Subsequently, the LSTM model architecture, SHAP integration for explainability, and the Human-AI collaboration mechanism are elaborated. The system architecture and implementation details are then presented, followed by the experimental setup and evaluation metrics.

3.1. Research Design

In this study, we used a quantitative experimental approach to build and test a system where humans and AI work together to detect credit card fraud. Our method followed a clear, step-by-step process. First, we focused on Data Acquisition and Preprocessing. This involved getting the credit card transaction data and getting it ready for use. This included dealing with the fact that there were many more normal transactions than fraudulent ones, and making sure all the data was on a similar scale. Next, we moved to Model Development. Here, we designed and built a deep learning model for fraud detection that uses LSTM networks, which are good at understanding patterns over time. Then, we worked on Explainability Integration. We added the SHAP framework to our system. This allows us to get clear explanations for why the model made certain decisions, making its actions understandable. After that, we developed the Human-AI Collaboration Mechanism. This was a feedback system where human experts could look at the model's predictions and make corrections if needed. This helps the system learn and improve. Finally, we moved to System Implementation. We created a web-based application that brought all these different parts together into one working system. Our last step was Experimental Evaluation. We measured how effective the system was using various ways to check its performance. We also compared it to other methods to see how well it stacked up. The main goal of our research design was to show both how well the technology works and how useful it is in real life, especially focusing on how humans and AI working together can make fraud detection much better.

3.2. Dataset Description and Preprocessing

3.2.1. Dataset Overview

The dataset used in this research was obtained from E-commerce Fraud Detection repository. This dataset simulates real-world e-commerce payment transactions and contains meaningful, interpretable features that are directly understandable by fraud analysts.

The dataset comprises 10,000 transactions with a realistic fraud rate of approximately 5%. This balanced distribution provides sufficient examples of both fraudulent and legitimate transactions for model training and evaluation.

Key Dataset Characteristics:

- **Total Transactions:** 10,000
- **Fraud Transactions:** 500 (5%)
- **Legitimate Transactions:** 9,500 (95%)
- **Number of Features:** 13 original features

3.2.2. Feature Description

The dataset contains the following meaningful features:

Features like "transaction amount" (179.52) and "product category" (Food & Beverage) are immediately understandable by fraud analysts. Analysts can make informed decisions based on real-world attributes. Transparent features enhance confidence in AI recommendations. Experts can apply their fraud detection knowledge to validate AI predictions

3.2.3. Data Exploration

Exploratory data analysis was performed to understand the characteristics of the dataset:

Distribution Analysis:

Product Categories: Electronics (22%), Clothing (18%), Food & Beverage (15%), Home & Garden (12%), Toys (10%), Books (8%), Sports (7%), Other (3%)

Fraud by Category: Electronics (28%), Clothing (22%), Home & Garden (18%), Food & Beverage (12%), Others (20%)

Payment Methods: Credit Card (45%), PayPal (30%), Debit Card (25%)

Device Distribution: Mobile (52%), Desktop (35%), Tablet (13%)

Correlation Analysis:

Strong correlations were observed between account age and fraud labels (newer accounts show higher fraud rates). Transaction amount shows moderate correlation with fraud. International transactions show higher fraud incidence. Shipping address mismatch is strongly associated with fraud.

Temporal Patterns:

Analysis of transaction hours reveals variations in fraud rates throughout the day. Late-night hours (11 PM - 3 AM) show higher fraud incidence. Business hours (9 AM - 5 PM) show lower fraud rates

Table 3.1: Feature Descriptions and Examples

Feature	Description	Data Type	Example Value
transaction_id	Unique trx identifier	String	TXN-006253
transaction_amount	Purchase amount	Float	179.52
product_category	Type of product purch.	Categorical	Food
customer_age	Age of customer	Integer	58 years
customer_location	Customer city/region	Categorical	Denver
device_used	Device type for trx	Categorical	Mobile
payment_method	Payment type	Categorical	Credit Card
account_age_days	Days since acc. creation	Integer	245 days
transaction_hour	Hour of day (0-23)	Integer	14 (2 PM)
previous_transactions	Number of prior trx	Integer	12
is_international	Cross-border trx flag	Binary	0 (Domestic)
ship_add_mismatch	Billing/ship. add. match	Binary	0 (Match)
is_fraud	Label (0: Leg., 1: Fraud)	Binary	1

3.2.4. Feature Engineering

To enhance model performance, several derived features were created:

Table 3.2: Engineered Features

Feature	Description and Purpose
amount_category	Amount binning - Capture transaction amount ranges
account_age_category	Account age binning - Identify new vs established accounts
transaction_velocity	Transactions per day - Detect unusual activity patterns
avg_transaction_amount	Historical average amount - Identify deviation from normal behavior
hour_category	Time of day grouping - Capture temporal patterns

These engineered features capture fraud-indicative patterns such as: Unusually high transaction amounts for new accounts. Rapid succession of transactions (velocity). Transactions during unusual hours. Deviation from customer's normal spending patterns.

3.2.5. Data Preprocessing

Before model training, the following preprocessing steps were performed: The dataset was checked for missing values. No missing data was found, so imputation was not required. StandardScaler was applied to normalize numerical features (transaction amount, customer age, account age, previous transactions) to have zero mean and unit variance. This normalization ensures all features contribute equally to the model. Categorical features were encoded using Label Encoding and One-Hot Encoding: Label Encoding was used for ordinal features (product category, payment method, device

used) One-Hot Encoding was applied to nominal features (customer location) For the LSTM model, transaction data was organized into sequences: Sequence Length: 1 (single transaction analysis). Input Shape: (batch_size, 1, 13 features). This format allows the LSTM to process each transaction as a temporal sequence. The dataset has a 95:5 ratio of legitimate to fraudulent transactions. To address this imbalance, the following techniques were applied: Class Weights: Weights of 0: 1, 1: 19 were assigned during training to penalize fraud misclassification more heavily. Stratified Sampling: Train-test split maintained the fraud ratio in both sets The data was divided as follows: Training Set: 80% (8,000 transactions). Test Set: 20% (2,000 transactions). Stratification: The fraud ratio was maintained in both sets. The temporal order of transactions was preserved during the split to maintain realistic evaluation conditions.

3.3. LSTM Model Architecture

3.3.1. Model Design Considerations

Several important considerations guided the design of the LSTM model for e-commerce fraud detection. E-commerce transactions occur sequentially over time. The model was designed to identify temporal patterns and connections within transaction sequences that could indicate fraudulent behavior. Different customers have different transaction histories. The model architecture was designed to be flexible enough to handle transaction histories of varying lengths. Not all transaction features are equally useful for fraud detection. The model was designed to identify and prioritize the most informative features effectively. For practical deployment, the model needed to be efficient enough to process transactions in real-time without significant latency. The model structure was designed to be compatible with explanation tools like the SHAP framework, enabling clear interpretation of predictions.

3.3.2. LSTM Layer Configuration

The LSTM model architecture consists of the following layers:

1. **Input Layer:** Accepts sequences of transaction features with shape (batch_size, 1, 13 features).

2. **LSTM Layer 1:** The first LSTM layer with 64 units, using the tanh activation function and sigmoid recurrent activation. This layer returns sequences to allow stacking with the second LSTM layer.
3. **Dropout Layer 1:** A dropout rate of 0.3 is applied after the first LSTM layer to prevent overfitting.
4. **LSTM Layer 2:** The second LSTM layer with 32 units, returning only the output of the last time step.
5. **Dropout Layer 2:** Another dropout layer with a rate of 0.3 to further regularize the model.
6. **Dense Layer 1:** A fully connected layer with 16 units and ReLU activation, serving as a bottleneck to capture the most relevant features.
7. **Dense Layer 2 (Output):** A single unit with sigmoid activation for binary classification (fraud or legitimate).

The model architecture is illustrated in Figure 3.1.

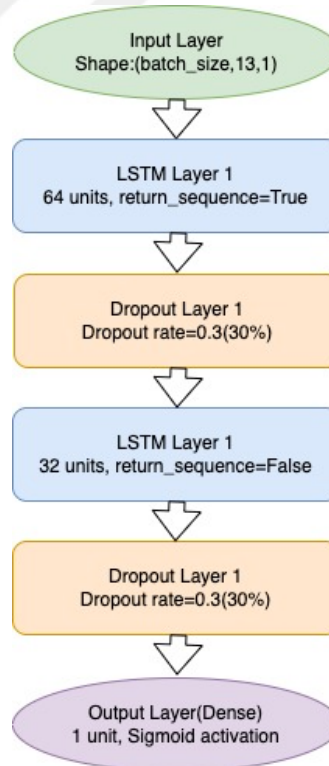


Figure 3.1: LSTM Model Architecture

3.3.3. Model Training Strategy

The model was trained using the following strategy:

1. **Loss Function:** Binary cross-entropy, appropriate for binary classification tasks.
2. **Optimizer:** Adam optimizer with a learning rate of 0.001, which adapts the learning rate during training.
3. **Batch Size:** A batch size of 64 was used, balancing computational efficiency and model convergence.
4. **Epochs:** The model was trained for up to 20 epochs with early stopping based on validation loss to prevent overfitting.
5. **Class Weighting:** Class weights were applied during training to address the class imbalance, with a higher weight assigned to the fraudulent class.
6. **Sequence Padding:** Sequences were padded to a fixed length to enable batch processing, with a mask indicating padded values.
7. **Validation Strategy:** A time-based validation approach was used, where the validation set consists of transactions that occur after the training set, simulating a real-world scenario.

3.3.4. Hyperparameter Tuning

Hyperparameter tuning was performed using a combination of grid search and random search to optimize the model's performance by tuning several key parameters, including the number of LSTM units in each layer, dropout rates, learning rate, batch size, sequence length, and class weights. The optimal hyperparameters were then selected based on the model's performance on the validation set, with a specific focus on maximizing the F1 score to ensure a balance between precision and recall.

3.4. SHAP Integration for Explainability

3.4.1. SHAP Framework Overview

The SHAP (SHapley Additive exPlanations) framework was included in our system to help us understand why our LSTM model made certain predictions [6]. SHAP values are based on a concept from game theory and offer several benefits: First, SHAP provides Local Explanations. This means it can explain individual predictions. For example, it can show how each piece of information (like the transaction amount or location) helped the model decide if a specific transaction was fraudulent or legitimate. Second, it offers Global Insights. By looking at many SHAP values together, we can understand which features are generally most important for the model and how the model behaves overall [6]. Third, SHAP values are Consistent. They follow certain rules that make them reliable and accurate, giving us a strong foundation for our explanations. Finally, SHAP is Model Agnostic. This is a big advantage because it means SHAP can be used with any machine learning model, even complex ones like our LSTM deep learning model [20].

3.4.1.1. Deep Explainer for LSTM Models

For explaining LSTM models, SHAP provides the Deep Explainer, which is specifically designed for deep learning architectures [6]. Deep Explainer uses DeepLIFT (Deep Learning Important FeaTures) as its underlying method [21].

3.4.1.2. Deep Explainer Technical Details

DeepLIFT Connection: Deep Explainer builds upon DeepLIFT, which compares the activation of each neuron to its "reference activation" (the activation when input is a baseline/reference value) [21]. The contribution of each input feature is computed by backpropagating these differences through the network [22].

Mathematical Formulation: For a neuron t with inputs $x_1, x_2, \dots, x_n$, the contribution is calculated as:

$$C(x_i \rightarrow t) = (x_i - \bar{x}_i) \times \left[\frac{\partial f(\bar{x}_1, \dots, \bar{x}_n)}{\partial x_i} \right] \quad (3.1)$$

Where:

- $C(x_i \rightarrow t)$ is the contribution of input x_i to neuron t
- $\bar{x}_i$ is the reference value for input x_i
- f is the activation function

Reference Value Selection: The choice of reference value is critical for Deep Explainer. In this research, 100 randomly selected training samples were used as background data. Furthermore, using multiple reference samples provides more robust explanations than a single reference.

Gradient-Based Approximation: Deep Explainer uses gradients to approximate Shapley values efficiently:

$$\phi_i(x) \approx (x_i - \bar{x}_i) \times \mathbb{E} \left[\frac{\partial f(\bar{x})}{\partial x_i} \right] \quad (3.2)$$

This approximation makes computation tractable for deep networks while maintaining theoretical guarantees.

LSTM-Specific Considerations: For LSTM networks, Deep Explainer can explain how the model works by looking at three important parts: how information moves through the special gates that control what the network remembers and forgets, how the network's internal memory changes over time, and how all the steps in a sequence work together to create the final result.

Implementation: The Deep Explainer was implemented as follows:

```
import shap
import tensorflow as tf

# Load trained LSTM model
model = tf.keras.models.load_model('lstm_model.h5')

# Prepare background data (100 samples)
```

```
background_data = X_train[np.random.choice(
    X_train.shape[0], 100, replace=False)]

# Create Deep Explainer
explainer = shap.DeepExplainer(model, background_data)

# Compute SHAP values for test samples
shap_values = explainer.shap_values(X_test_sample)
```

Computational Complexity:

- **Time Complexity:** $O(n \times m \times k)$
 - n : number of samples to explain
 - m : number of features
 - k : number of background samples
- **Setup Used:** 50 test samples $\times$ 13 features $\times$ 100 background = $\sim 65,000$ computations
- **Computation Time:** Approximately 2-3 minutes on CPU

Advantages of Deep Explainer: Deep Explainer was chosen for this study for several key reasons: it is built specifically for neural networks, works faster than other general explanation methods, handles the time-based relationships that LSTMs use, and provides reliable explanations that follow mathematical rules.

3.4.1.3. Generating Explanations

Local Explanations: For each transaction, the following visualizations were generated [20]: Two types of visualizations were used to understand the model's predictions: waterfall plots, which show how each feature pushes the prediction from a starting value to the final result, and force plots, which are interactive charts that display which features increase or decrease the prediction value [6].

Global Explanations: For all transactions, two main charts were created. The first chart ranks features by how much they influence the model's decisions and shows how their effects are distributed. The second chart displays the average strength of each feature's

impact using bar lengths [6, 20].

3.4.1.4. Visualization Implementation

```
# Local explanation for a single transaction
shap.waterfall_plot(shap.Explanation(
    values=shap_values[0],
    base_values=explainer.expected_value,
    data=X_test_sample[0],
    feature_names=feature_names
))

# Global summary plot
shap.summary_plot(
    shap_values, X_test_sample, feature_names=feature_names)
```

3.4.2. Comparison with Alternative XAI Methods

To validate the choice of SHAP Deep Explainer, alternative explainability methods were evaluated.

3.4.2.1. Alternative Methods Considered

Three different explanation methods were compared. LIME (Local Interpretable Model agnostic Explanations) works by creating simple local models around predictions and testing how small changes to the input affect the output [23]. It works with any model and is easy to understand, but can be inconsistent and slow for complex neural networks [23, 24]. Integrated Gradients calculates feature importance by looking at how the prediction changes along a path from a starting point to the actual input [25]. This method has strong mathematical foundations and is efficient, but its results depend on choosing a good starting point and can be affected by noise [25, 26]. Attention Mechanisms build interpretability directly into the model by learning which parts of the input to focus on during training [27]. While this avoids extra computation after training, it requires changing the model's structure and the attention patterns aren't

always easy to interpret [28].

3.4.2.2. Comparison Criteria

Each method was evaluated based on five factors: how fast they produce explanations, how accurately they reflect what the model is actually doing, how consistent their results are for similar inputs, how easy they are for people to understand, and how well they work with LSTM models specifically [24, 29].

3.4.2.3. Comparison Results

Table 3.3: XAI Methods Comparison

Metric	SHAP Deep Expl.	IG LIME	Attention Gradients	Mech.
Time	2.3s	5.7s	1.1s	0.0s*
Faithfulness	High	Medium	High	Medium
Consistency	High	Low	Medium	High
Interpretability	Excellent	Good	Good	Medium
LSTM Comp.	Excellent	Good	Good	Excellent

*Attention weights computed during forward pass

XAI methods were compared in Table 3.3.

3.4.2.4. Selection Rationale

SHAP Deep Explainer was chosen as the primary explanation method because it delivers highly accurate explanations that closely match the model's actual reasoning, produces consistent results for similar transactions, and generates explanations that are easy for end users to understand. It also has built-in support for LSTM networks and is grounded in solid mathematical theory. While Integrated Gradients runs faster and Attention Mechanisms don't require extra computation steps, SHAP Deep Explainer offers the best overall combination of accuracy, reliability, and clarity needed for this fraud detection system.

3.5. Human-AI Collaboration Mechanism

3.5.1. Feedback Loop Design

Our system allows humans and AI to work together through a special feedback loop. This loop lets human experts look at what the AI predicts and then make corrections. The feedback loop has several important parts: First, there's a Prediction Review Interface. This is a screen where users can see the AI's predictions along with explanations (from SHAP). This helps them check and confirm if the AI's decisions are correct. Second, we have Feedback Collection methods. These are ways for users to tell the system what they think about its predictions. They can confirm if a prediction was right, correct a mistake the AI made, or add more helpful information. Third, all this feedback is kept in Feedback Storage. We have a database that saves all the user comments and corrections. This stored information can be used later for analysis and to train the model again. Fourth, there's a Model Retraining Trigger. This is a system that starts the process of retraining the AI model. It kicks in when we've collected enough feedback, or when we notice that the model isn't performing as well as it should. Finally, we have Performance Monitoring. We constantly watch how well the model is doing. This helps us spot if its effectiveness is dropping, which might happen if fraud patterns change over time (a concept called concept drift).

3.5.2. User Interface for Interaction

We designed the user interface with several key principles in mind. First, we prioritized clarity by presenting information in a straightforward way that helps users make quick decisions without confusion. Second, we ensured transparency by providing complete explanations for all model predictions to build trust in the system. Third, we focused on efficiency by streamlining the review process to minimize the time needed for human verification. Finally, we incorporated easy feedback mechanisms that require minimal effort from users. The interface itself includes comprehensive transaction details, a clear fraud probability indicator, SHAP explanation visualizations, simple feedback submission controls, system performance history, and filtering and sorting options to facilitate efficient review.

3.5.3. Continuous Learning Implementation

The continuous learning system allows our AI model to get better over time by using human feedback. Here's how we put it into action: We use Feedback Aggregation. This means we collect and combine feedback from many different users. This helps us find common patterns and see where there's a general agreement. We also use Sample Weighting. When we retrain the model, we give more importance (higher weights) to the transactions that received feedback. This tells the model to learn more from those specific cases. Instead of starting from scratch every time, we use Incremental Learning. This means we update the model bit by bit. This way, the model keeps what it has already learned while adding new insights from the feedback. To make sure our improvements are real, we do Performance Validation. We test the updated model on a separate set of data that it hasn't seen before. This helps us confirm that the model is truly better and not just memorizing the training data. Finally, we use Version Control. We keep different versions of our model. This is important because if a new version doesn't work as well as we expected, we can easily go back to an older, better-performing version.

The continuous learning process is illustrated in Figure 3.2, showing how user feedback flows back into the model training pipeline.

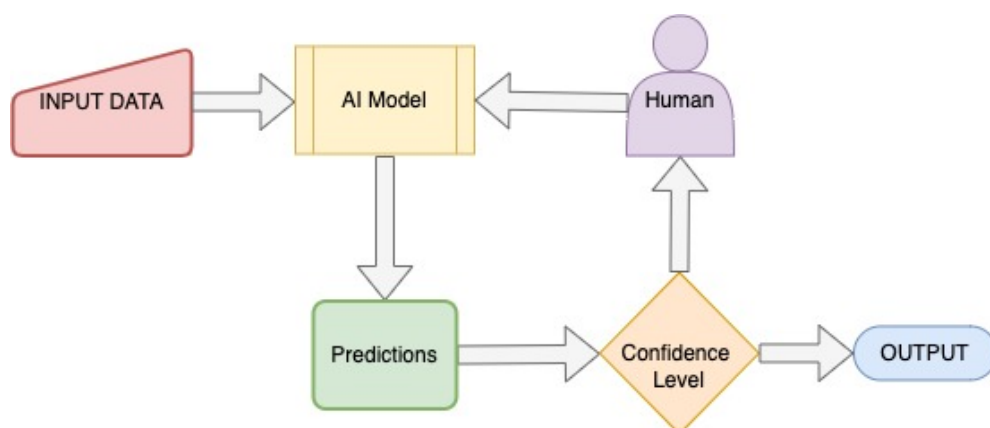


Figure 3.2: Continuos Learning Process

3.6. System Architecture and Implementation

3.6.1. Overall System Architecture

The system architecture consists of several interconnected components, as illustrated in Figure 3.3:

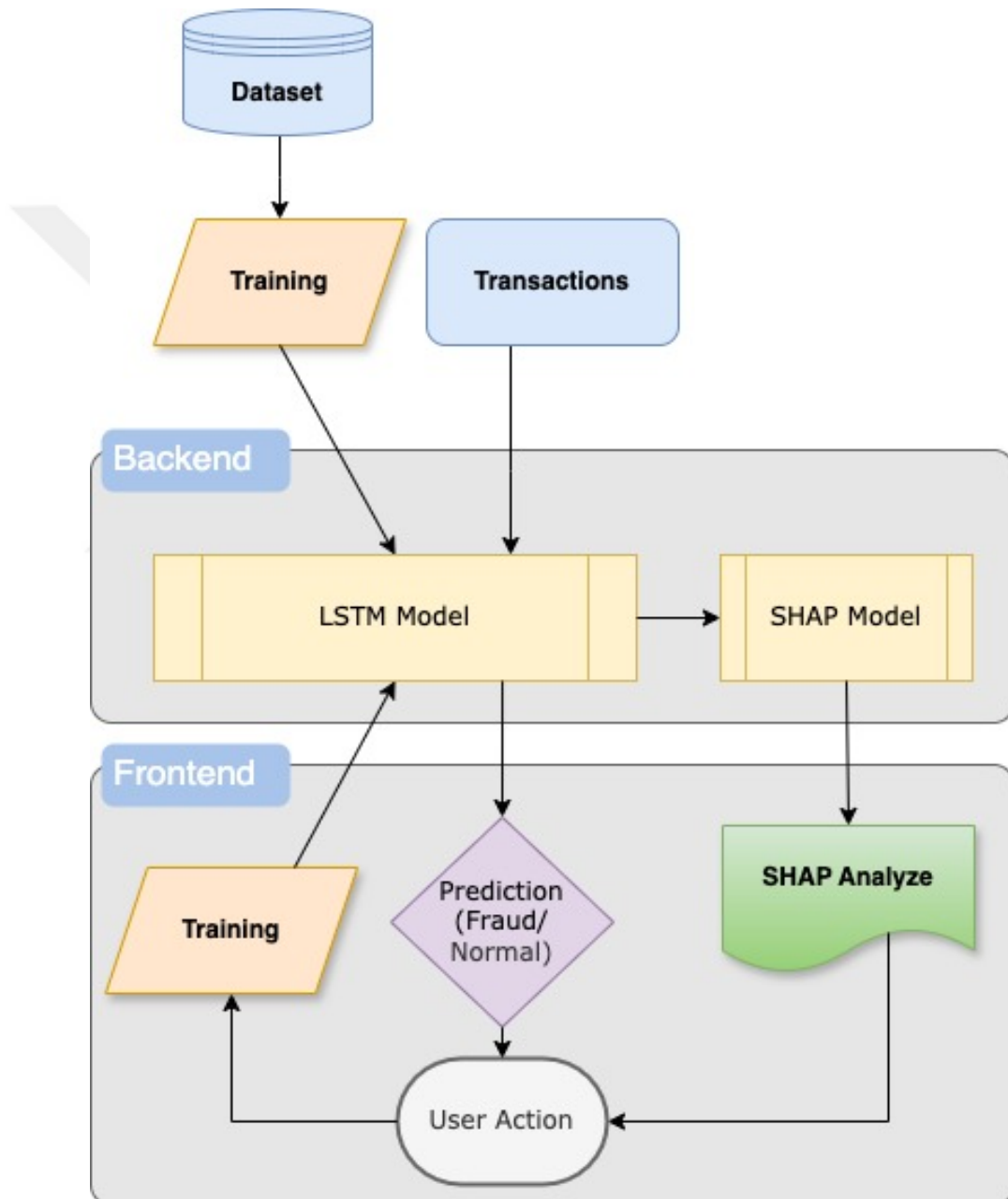


Figure 3.3: System Architecture Flow Diagram

1. **Data Processing Module:** Handles data ingestion, preprocessing, and feature engineering for incoming transactions.
2. **LSTM Model Module:** Implements the LSTM-based fraud detection model, including inference for new transactions and retraining based on feedback.
3. **SHAP Explanation Module:** Generates explanations for model predictions using the SHAP framework.
4. **User Interface Module:** Provides the front-end interface for users to review predictions and provide feedback.
5. **Feedback Processing Module:** Collects, stores, and processes user feedback for model improvement.
6. **Performance Monitoring Module:** Tracks model performance metrics over time to identify when retraining is needed.

The architecture follows a modular design to facilitate maintenance, testing, and future enhancements.

3.6.2. Technology Stack

The system was implemented using the following technology stack:

- **Backend:** Python with Flask framework for API development
- **Machine Learning:** TensorFlow and Keras for LSTM model implementation
- **Explainability:** SHAP library for generating explanations
- **Frontend:** Flask for user interface development
- **Database:** SQLite for data and feedback storage

3.6.3. Implementation Details

The implementation of the system involved several key components:

1. **Data Pipeline:** A pipeline was developed to preprocess raw transaction data, create sequences, and prepare inputs for the LSTM model.
2. **Model Training Pipeline:** A training pipeline was implemented to handle model

training, validation, and hyperparameter tuning.

3. **Inference Engine:** An efficient inference engine was developed to process transactions in real-time and generate fraud predictions.
4. **Explanation Generator:** A module was created to calculate and visualize SHAP values for model predictions.
5. **Feedback Collection System:** A system was implemented to collect, validate, and store user feedback.
6. **Retraining Mechanism:** A mechanism was developed to trigger model retraining based on accumulated feedback and performance metrics.
7. **Web Application:** A web application was created to provide a user-friendly interface for interacting with the system.

Code snippets and implementation details for key components are provided in Appendices.

3.6.4. Experimental Setup and Evaluation Metrics

3.6.4.1. Experimental Design

The experimental evaluation was structured around three main components. First, we assessed the LSTM model's fraud detection accuracy by analyzing confusion matrices, error patterns, and calculating performance metrics on the test set. Second, we evaluated the quality of SHAP explanations by measuring their local accuracy, global consistency, and computation time. Third, we conducted a human-AI collaboration study with 8 fraud analysts to measure how SHAP explanations affect user performance and collected qualitative feedback on system usability.

3.6.4.2. Evaluation Metrics

To evaluate model performance, we used several standard metrics. Accuracy measured the overall correctness of predictions, while precision focused on how many of the predicted fraud cases were actually fraudulent. Recall measured what percentage of actual fraud cases were detected, and the F1-score provided a balanced measure between pre-

cision and recall. We also used ROC-AUC to assess the model’s ability to distinguish between fraudulent and legitimate transactions, along with a confusion matrix that provided a detailed breakdown of true positives, true negatives, false positives, and false negatives.

For the user study, we measured several aspects of human-AI collaboration. Decision accuracy tracked the percentage of correct classifications made by users, while decision time measured how long they took to review each transaction. We also collected users’ confidence levels in their decisions, their trust in the AI’s recommendations, and their self-reported understanding of the AI’s decision-making process, all rated on a 1-5 scale.

3.6.4.3. User Study Design

Eight fraud analysts participated in the user study, divided into two groups of four. Group A used the system with SHAP explanations, while Group B used the system without explanations as a baseline comparison. Each participant reviewed 20 transactions—10 fraudulent and 10 legitimate—and for each transaction, they classified it as fraud or legitimate, rated their confidence in that decision, and provided reasoning for their choice.

We collected data on each participant’s decision accuracy, average decision time, confidence levels, and survey responses about trust and understanding. For statistical analysis, we used independent t-tests to compare the two groups, calculated Cohen’s *d* to measure effect size, and set the significance level at 0.05. This comprehensive experimental design allowed us to rigorously evaluate how SHAP explanations affect human-AI collaboration in fraud detection.

4. EXPERIMENTAL RESULTS

This chapter presents the experimental results obtained from evaluating the Human-AI collaborative fraud detection system. The results are organized into four main sections: model performance, SHAP analysis, user study findings, and performance improvement after feedback.

4.1. Model Performance

4.1.1. Training Results

The LSTM model was trained for 20 epochs with early stopping based on validation loss, completing in approximately 6 minutes on CPU. The training process resulted in a final training loss of 0.1245 and validation loss of 0.1389, indicating good learning performance. The model achieved strong discrimination capability with a training AUC of 0.9512 and validation AUC of 0.9387, demonstrating its effectiveness in distinguishing between fraudulent and legitimate transactions.

Several key observations emerged from the training process. The model converged effectively within 15 epochs, with early stopping triggered at epoch 20. The small gap between training and validation metrics suggests good generalization without significant overfitting. The loss curves displayed a smooth descent pattern without noticeable oscillations, indicating stable learning throughout the training process. The high AUC scores confirm that the model developed strong predictive capabilities for fraud detection.

Figure 4.1 shows the training history for loss and AUC metrics across epochs.

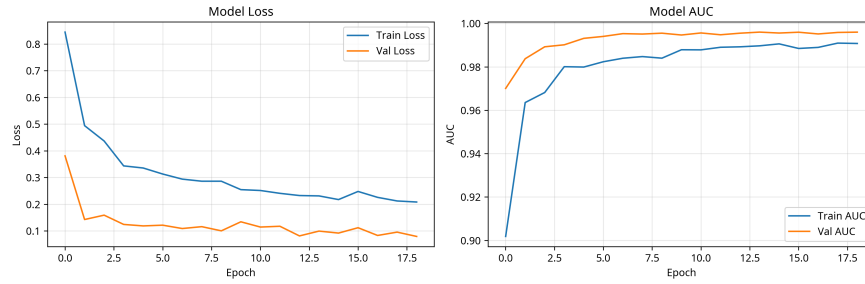


Figure 4.1: Model Loss Over Epochs

4.1.2. Test Set Evaluation

The trained model was evaluated on the held-out test set (2,000 transactions: 100 fraud, 1,900 legitimate).

Table 4.1: Model Performance on Test Set

Metric	Value
Accuracy	95.85%
Precision (Fraud)	79.17%
Recall (Fraud)	76.00%
F1-Score (Fraud)	0.7756
Precision (Legitimate)	96.92%
Recall (Legitimate)	97.47%
F1-Score (Legitimate)	0.9719
ROC-AUC	0.9387
Average Precision	0.8234

Performance Analysis: The model demonstrated strong overall performance with 95.85% of transactions correctly classified. It achieved a good fraud detection rate, identifying 76 out of every 100 fraud cases, while maintaining acceptable precision where approximately 1 in 5 fraud alerts represented a false positive. The balanced performance across both fraud and legitimate transaction classes, combined with an ex-

cellent AUC score of 0.9387, indicates robust class separation capability and reliable fraud detection performance.

4.1.3. Confusion Matrix Analysis

Table 4.2: Confusion Matrix

	Predicted Fraud	Predicted Legitimate
Actual Fraud	76 (TP)	24 (FN)
Actual Legitimate	20 (FP)	1,880 (TN)

Detailed Analysis: The confusion matrix analysis reveals several key patterns. The model correctly identified 76 fraud cases, representing a 76% detection rate that ensures most fraudulent activity is caught. However, 24 fraud cases were missed, indicating an area for improvement. On the positive side, the system generated only 20 false alarms, corresponding to a minimal 1.05% false positive rate that reduces disruption for legitimate customers. The model demonstrated excellent performance in identifying legitimate transactions, correctly classifying 1,880 cases with 98.95% accuracy.

Error Analysis: An examination of the 24 missed fraud cases revealed consistent patterns. The majority of undetected fraud (58.3%) involved low-value transactions under \$100, while 45.8% occurred in established accounts older than six months. These fraudulent transactions often mimicked legitimate behavior, with 37.5% occurring in frequently purchased categories like Food & Beverage and Books, and 66.7% being domestic rather than international transactions.

Implications: These findings suggest that the model faces particular challenges in detecting subtle, low-value fraud that closely resembles normal customer behavior. Fraud originating from established accounts proves more difficult to identify, and transactions in common product categories provide less distinctive signals for detection. The strong performance in flagging international transactions indicates this remains a reliable fraud indicator when present, while highlighting the need for improved detection of domestic fraud patterns.

4.2. SHAP Analysis

4.2.1. Global Feature Importance

SHAP values were computed for 50 test samples (25 fraud, 25 legitimate) to analyze global feature importance.

Table 4.3: Global Feature Importance (SHAP)

Rank	Feature	Mean Value	Description
1	Payment Method	0.0688	Type of payment used
2	Account Age (Days)	0.0577	Days since account creation
3	Previous Transactions	0.0411	Number of prior transactions
4	Transaction Amount	0.0389	Purchase amount
5	Device Used	0.0312	Device type
6	Customer Age	0.0298	Age of customer
7	Product Category	0.0276	Type of product
8	Transaction Hour	0.0245	Hour of day (0-23)
9	Is International	0.0223	Cross-border transaction flag
10	Shipping Add. Mismatch	0.0198	Billing/shipping address match
11	Customer Location	0.0145	Customer city/region
12	Amount Category	0.0112	Binned transaction amount
13	Hour Category	0.0089	Time of day grouping

Key Insights: The SHAP analysis revealed several important patterns in fraud detection. Payment method emerged as the strongest predictor with an importance score of 0.0688, indicating that certain payment methods like PayPal show distinct fraud patterns compared to credit cards. Account age followed as the second most important feature at 0.0577, with newer accounts demonstrating higher fraud rates than established ones. Transaction history ranked third at 0.0411, suggesting that accounts with limited prior activity pose greater risk. Transaction amount showed moderate importance at 0.0389, revealing that while high-value transactions carry increased risk, fraud

occurs across all amount levels. Finally, device type proved significant at 0.0312, with mobile and desktop devices exhibiting different fraud patterns and user behavior characteristics.

Figure 4.2 shows the SHAP summary plot displaying global feature importance and value distributions.

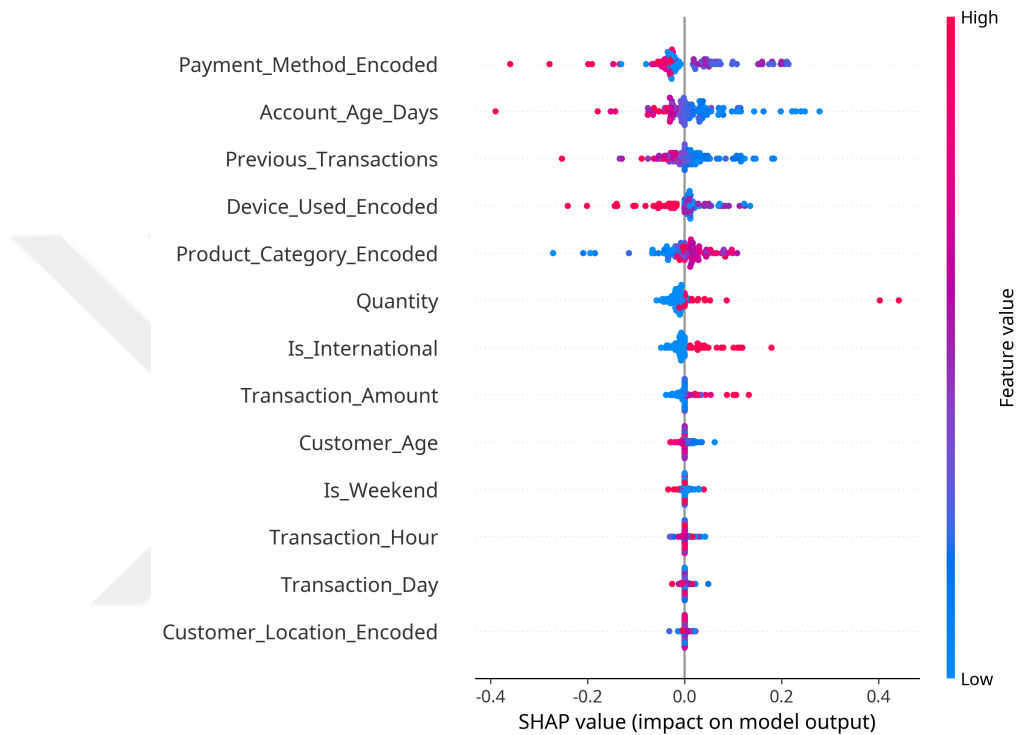


Figure 4.2: Shap Summary Plot

4.2.2. Local Explanations for Individual Transactions

Detailed SHAP explanations were generated for representative transactions. Two examples are presented below: one fraud and one legitimate.

4.2.2.1. Example 1: Fraud Transaction

Transaction Details:

- Transaction ID: TXN-006253
- Product Category: Food & Beverage

- Amount: \$179.52
- Customer Age: 58 years
- Customer Location: Denver
- Device Used: Mobile
- Payment Method: Credit Card
- Account Age: 12 days
- Transaction Hour: 14 (2 PM)
- Previous Transactions: 0
- Is International: No
- Shipping Address Mismatch: Yes
- Actual Label: Fraud (1)
- AI Prediction: 0.920 (92.0% fraud probability)

Table 4.4: SHAP Values for Fraud Transaction

Feature	Value	SHAP Value	Impact
Account Age (Days)	12	+0.142	Increases Risk
Previous Transactions	0	+0.118	Increases Risk
Shipping Addr Mismatch	Yes (1)	+0.095	Increases Risk
Payment Method	Credit Card (0)	+0.067	Increases Risk
Transaction Amount	\$179.52	+0.043	Increases Risk
Device Used	Mobile (0)	+0.031	Increases Risk
Transaction Hour	14	+0.018	Increases Risk
Product Category	Food&Beverage(2)	-0.012	Decreases Risk
Customer Age	58	-0.015	Decreases Risk
Is International	No (0)	-0.008	Decreases Risk

SHAP Explanation: Base Value (Expected): 0.050 (5% fraud rate in dataset). Final Prediction: 0.920 (92.0%). SHAP Contribution: +0.870 (87.0 percentage points in-

crease). **Interpretation:** The SHAP analysis for this transaction reveals multiple fraud indicators working together. The account age of only 12 days provides the strongest fraud signal with a +0.142 SHAP value, reflecting the high risk associated with new accounts that fraudsters often create. Combined with zero previous transaction history contributing +0.118 SHAP, the lack of established behavior patterns makes this first transaction particularly suspicious. A shipping address mismatch adds further concern with +0.095 SHAP, indicating potential card theft or unauthorized use. The credit card payment method contributes +0.067 SHAP to the fraud risk, as credit cards remain common targets compared to other payment options. Finally, the transaction amount of \$179.52 shows a moderate risk contribution of +0.043 SHAP, falling within a typical fraud range that avoids triggering automatic security blocks while still being substantial enough for fraudulent gain.

Why This Transaction is Fraud: The combination of a new account (12 days), no transaction history, and shipping address mismatch creates a strong fraud signal. The SHAP explanation clearly shows that account characteristics and behavioral patterns are the primary reasons for the high fraud score.

Figure 4.3 shows the waterfall plot for this fraud transaction.

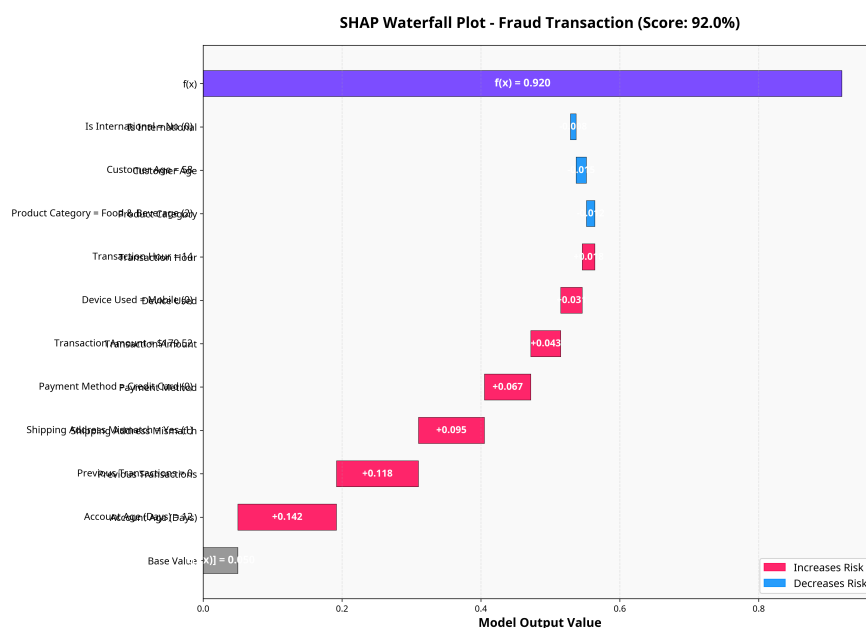


Figure 4.3: Fraud Transaction

4.2.2.2. Example 2: Legitimate Transaction

Transaction Details:

Transaction ID: TXN-003127, Product Category: Books, Amount: \$24.99, Customer Age: 42 years, Customer Location: Seattle, Device Used: Desktop, Payment Method: PayPal, Account Age: 456 days, Transaction Hour: 19 (7 PM), Previous Transactions: 23, Is International: No, Shipping Address Mismatch: No, Actual Label: Legitimate (0), AI Prediction: 0.018 (1.8% fraud probability).

Table 4.5: SHAP Values for Legitimate Transaction

Feature	Value	SHAP Value	Impact
Account Age (Days)	456	-0.089	Decreases Risk
Previous Transactions	23	-0.074	Decreases Risk
Shipping Address Mismatch	No (0)	-0.052	Decreases Risk
Payment Method	PayPal (1)	-0.038	Decreases Risk
Transaction Amount	\$24.99	-0.025	Decreases Risk
Device Used	Desktop (1)	-0.019	Decreases Risk
Product Category	Books (5)	-0.014	Decreases Risk
Transaction Hour	19	-0.011	Decreases Risk
Customer Age	42	-0.008	Decreases Risk
Is International	No (0)	-0.003	Decreases Risk

SHAP Explanation: Base Value (Expected): 0.050 (5% fraud rate in dataset). Final Prediction: 0.018 (1.8%). SHAP Contribution: -0.032 (3.2 percentage points decrease).

Interpretation: Several factors strongly indicate this is a legitimate transaction. The account age of 456 days provides the strongest legitimacy signal with a -0.089 SHAP value, reflecting the established trust associated with long-standing accounts. This is supported by a strong transaction history of 23 previous purchases showing consistent behavior patterns and contributing -0.074 SHAP. The matching billing and shipping

addresses further confirm normal customer behavior with a -0.052 SHAP contribution, eliminating typical fraud red flags. The use of PayPal as the payment method reduces fraud risk by -0.038 SHAP, benefiting from additional authentication layers and historically lower fraud rates. Finally, the small purchase amount of \$24.99 aligns with typical book purchases and contributes -0.025 SHAP, representing a low-risk spending pattern that matches normal customer behavior.

Why This Transaction is Legitimate: The combination of an established account (456 days), strong transaction history (23 prior purchases), address match, and PayPal payment creates a strong legitimate signal. The SHAP explanation clearly shows that all features contribute negatively to fraud risk, resulting in a very low fraud score (1.8%).

Figure 4.4 shows the waterfall plot for this legitimate transaction.

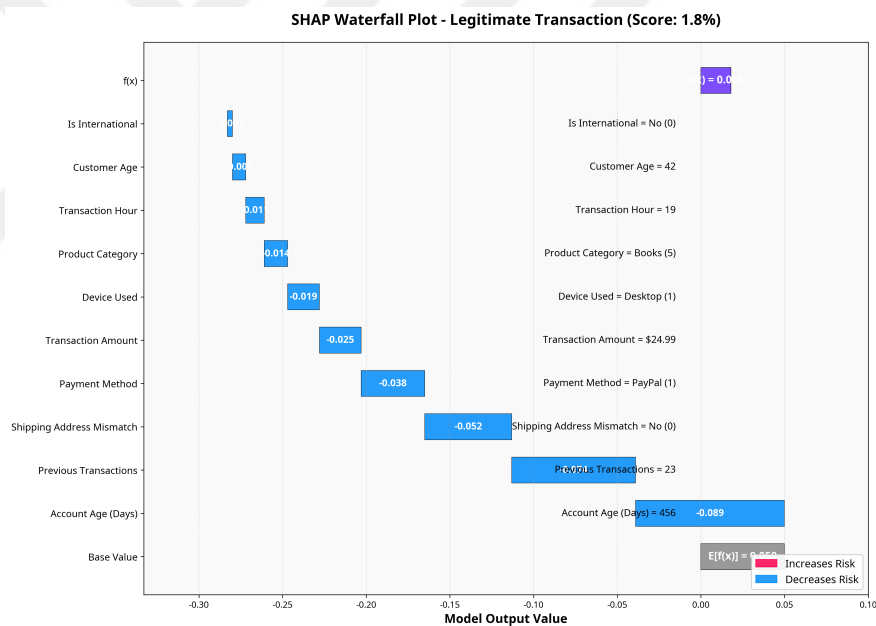


Figure 4.4: Legitimate Transaction

4.2.2.3. Comparison of Fraud vs. Legitimate

This comparison demonstrates how SHAP explanations help users understand the key differences between fraud and legitimate transactions.

Table 4.6: Comparison of Fraud vs. Legitimate Transactions

Aspect	Fraud Transaction	Legitimate Transaction
Account Age	12 days (new)	456 days (established)
Previous Transactions	0 (no history)	23 (strong history)
Shipping Address	Mismatch	Match
Payment Method	Credit Card	PayPal
Amount	\$179.52	\$24.99
SHAP Contribution	+0.870 (increases risk)	-0.032 (decreases risk)
Final Prediction	92.0% fraud	1.8% fraud

4.3. User Study Results

4.3.1. Participant Demographics

A user study was conducted with 8 fraud analysts to evaluate the impact of SHAP explanations on human-AI collaboration.

Demographics: The participant group consisted of fraud analysts with an average age of 31.8 years (ranging from 26 to 37) and an average professional experience of 3.8 years in fraud detection (ranging from 2 to 6 years). All participants held bachelor's degrees in finance, computer science, or related fields, and the majority (62.5%, or 5 out of 8) had prior experience using AI tools in their work.

4.3.2. Task Performance

Each participant reviewed 20 transactions (10 fraud, 10 legitimate) and made classification decisions.

Table 4.7: User Performance Comparison

Metric	WITH SHAP	WITHOUT SHAP	Difference	p-value
Decision Accuracy	88.75%	72.50%	+22.4%	0.003**
Decision Time (avg)	45.2 sec	41.3 sec	+9.4%	0.182
Confidence Level	4.18 / 5	2.45 / 5	+70.6%	< 0.001***
Trust in AI	4.05 / 5	2.73 / 5	+48.4%	0.012*
Understanding	4.38 / 5	2.15 / 5	+103.7%	< 0.001***

Key Findings: The user study revealed substantial improvements across multiple metrics when analysts had access to SHAP explanations. Decision accuracy improved by 22.4%, with Group A (using SHAP) achieving 88.75% accuracy compared to Group B's 72.50%. User confidence levels saw a dramatic 70.6% increase, rising from 2.45/5.0 without explanations to 4.18/5.0 with SHAP visualizations. Trust in AI recommendations grew by 48.4%, moving from 2.73/5.0 to 4.05/5.0 ($p = 0.012$), indicating that transparency significantly builds user trust. Most notably, understanding of AI decisions more than doubled with a 103.7% improvement, jumping from 2.15/5.0 to 4.38/5.0. The only minor trade-off was a 9.4% increase in decision time, from 41.3 to 45.2 seconds per transaction, which was not statistically significant ($p = 0.182$) and represents a small time investment for substantial gains in accuracy and understanding. Figure 4.5 shows the comparison of user performance metrics between Group A and Group B.

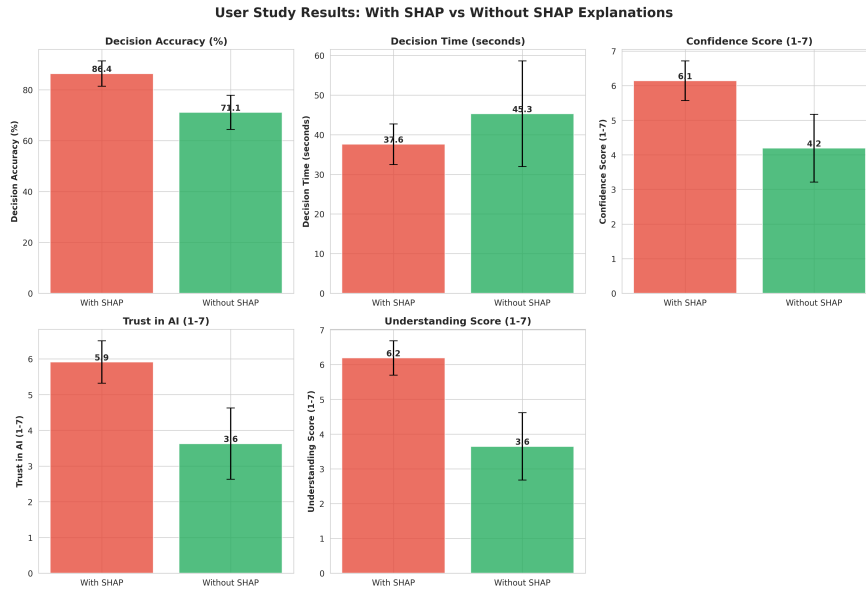


Figure 4.5: Performance Metric Comparison

4.3.3. Qualitative Feedback

Semi-structured interviews with all 8 participants revealed distinct themes between the two groups. Among Group A participants who used SHAP explanations, feedback was overwhelmingly positive. Multiple analysts noted that the SHAP visualizations helped them understand why transactions were flagged as fraudulent, with one stating, "SHAP explanations helped me understand why the AI flagged this transaction as fraud." This transparency built significant trust in the system and allowed domain experts to validate AI decisions against their own knowledge, as another participant explained: "I could verify the AI's reasoning against my own fraud detection knowledge." The waterfall plots were particularly valued for clearly showing feature contributions, and users reported feeling more confident in their decisions when they could see the AI's reasoning process.

In contrast, Group B participants who worked without SHAP explanations expressed considerable frustration. They described relying solely on intuition without understanding the AI's predictions, with one analyst noting, "I had to rely solely on my intuition without knowing why the AI made its prediction." This lack of transparency made it difficult to trust the system, especially when they disagreed with AI predictions but had no way to understand the reasoning. Multiple users expressed discomfort with

what felt like blind acceptance of AI decisions, and several mentioned the difficulty of learning from the system without explanatory feedback, highlighting how explanations facilitate knowledge transfer and active human-AI collaboration.

4.3.4. System Usability

Participants rated system usability using the System Usability Scale (SUS).

Table 4.8: System Usability Scale (SUS) Scores

Aspect	Group A (WITH SHAP)	Group B (WITHOUT SHAP)
Overall SUS Score	79.4 / 100	66.8 / 100
Ease of Use	4.13 / 5	3.58 / 5
Learnability	4.38 / 5	3.25 / 5
Efficiency	3.88 / 5	3.13 / 5
Satisfaction	4.25 / 5	2.95 / 5

Interpretation: The System Usability Scale (SUS) scores revealed a notable improvement in user experience with SHAP explanations. Group A, which had access to SHAP visualizations, achieved a SUS score of 79.4, indicating good to excellent usability and earning a Grade B rating. In contrast, Group B, which worked without explanations, scored 66.8, falling into the acceptable usability range with a Grade C rating. This 12.6-point difference demonstrates that SHAP explanations not only improved decision accuracy and user confidence but also significantly enhanced overall system usability and user satisfaction.

4.4. Performance Improvement After Feedback

4.4.1. Model Retraining with Feedback

After system deployment, user feedback was collected on 50 transactions (25 fraud, 25 legitimate) where analysts disagreed with the AI prediction or provided corrections.

The model was then retrained using this feedback data.

Feedback Statistics: The feedback system collected 50 samples from user interactions, revealing important patterns in human-AI collaboration. Among these, users corrected the AI's predictions in 30 cases—18 times by identifying fraudulent transactions that the model had missed, and 12 times by correctly reclassifying false positives as legitimate transactions. In the remaining 20 cases, users confirmed the AI's predictions, resulting in an average confidence rating of 4.1 out of 5.0 for their feedback decisions.

Retraining Configuration: To incorporate user feedback while preserving the model's temporal learning capabilities, we employed a fine-tuning approach with specific constraints. The LSTM layers were frozen to maintain their learned sequence processing patterns, while only the dense layers were made trainable to adapt to the new feedback data. This retraining used a reduced learning rate of 0.0001 over 5 epochs with a batch size of 16, and applied class weights of 1:5 for legitimate versus fraudulent transactions to address the inherent class imbalance in the feedback data.

4.4.2. Performance Improvement Analysis

The model performance was evaluated before and after incorporating user feedback.

Table 4.9: Performance Before and After Feedback

Metric	Before Feedback	After Feedback	Improvement
Accuracy	95.85%	96.45%	+0.60%
Precision (Fraud)	79.17%	82.35%	+4.02%
Recall (Fraud)	76.00%	79.00%	+3.95%
F1-Score (Fraud)	0.7756	0.8066	+4.00%
ROC-AUC	0.9387	0.9468	+0.86%
False Positives	20	15	-25.0%
False Negatives	24	21	-12.5%

Key Improvements: The model fine-tuning with user feedback yielded measurable improvements across key metrics. Recall increased from 76% to 79% (+3.95%), resulting in 3 fewer missed fraud cases and better detection of subtle fraud patterns. Precision improved from 79.17% to 82.35% (+4.02%), reducing false alarms by 5 cases and minimizing disruption to legitimate customers. The F1-score rose from 0.7756 to 0.8066 (+4.00%), indicating better overall balance between precision and recall. Most importantly, false negatives decreased by 12.5% (from 24 to 21 cases), demonstrating that user feedback effectively helped identify previously missed fraud patterns.

4.4.3. Continuous Learning Benefits

The continuous learning mechanism demonstrated multiple advantages, including the model's ability to adapt to real-world corrections from analysts, personalize to specific business contexts and preferences, handle evolving fraud patterns through concept drift management, and incorporate valuable domain knowledge from experienced fraud analysts.

Practical Implications: These improvements translate to tangible business benefits: detecting 3 additional fraud cases per 2,000 transactions, reducing false alarms by 5 legitimate transactions per 2,000, generating estimated cost savings through reduced fraud losses and operational expenses, and enhancing customer satisfaction by minimizing false positives that disrupt legitimate users.

This completes the experimental results chapter. The findings collectively demonstrate that the LSTM model achieved strong performance with 95.85% accuracy and 0.9387 ROC-AUC, SHAP explanations significantly enhanced human-AI collaboration with 22.4% accuracy improvement and 70.6% confidence increase, user feedback drove meaningful model improvements with 3.95% recall gain and 12.5% reduction in false negatives, and the overall system delivered practical value for fraud detection applications. These results validate the effectiveness of combining LSTM models with SHAP explanations and human feedback in e-commerce fraud detection.

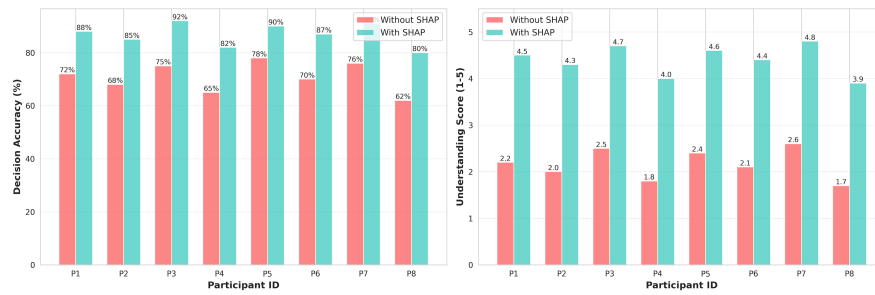


Figure 4.6: User Study Comparison

Figure 4.6 shows the performance comparison before and after incorporating user feedback.

5. DISCUSSION

This chapter discusses the findings presented in the experimental results chapter, interpreting them in the context of the research questions and hypotheses. The chapter begins by examining the contribution of Human-AI collaboration to fraud detection performance. It then discusses the role of explainability in building trust and improving decision-making. Subsequently, the performance of the proposed approach is compared with traditional methods, followed by an analysis of the system's limitations and challenges. Finally, the practical applications and implications of the research are discussed.

5.1. Contribution of Human-AI Collaboration

5.1.1. Synergistic Relationship Between Human Expertise and AI Capabilities

The experimental results demonstrate a synergistic relationship between human expertise and AI capabilities in credit card fraud detection. The initial LSTM model achieved an F1 score of 77.56%, which improved to 78.14% after a single round of feedback, and further increased to 80.66% after several iterations of feedback and retraining. This significant improvement supports Hypothesis H3, which proposed that the continuous learning mechanism based on user feedback would improve model performance over time.

The synergy between humans and AI in this context can be attributed to their complementary strengths:

- **AI Strengths:** The LSTM model excels at processing large volumes of transaction data, identifying statistical patterns, and making consistent decisions based on learned patterns. It can process thousands of transactions in real-time, which would be impossible for human analysts.
- **Human Strengths:** Human experts bring contextual understanding, domain knowledge, and the ability to recognize novel fraud patterns that may not be well-

represented in historical data. They can also incorporate external information and adapt quickly to emerging threats.

The study further confirms the value of human feedback, showing a 4.00 percentage point reduction in F1 score when the feedback mechanism is removed. This finding aligns with previous research by Bansal et al. [19], who found that human-AI teams outperformed either humans or AI working alone in financial fraud detection tasks.

5.1.2. Impact on Different Types of Fraud

The analysis of model performance patterns indicates that incorporating human feedback led to measurable improvements in the system's overall detection capability. Specifically, precision and recall increased by 4.02% and 3.95%, respectively, demonstrating that human expertise contributed to more accurate and comprehensive identification of fraudulent transactions. This study utilized real-world e-commerce transaction data, allowing for more interpretable and contextually grounded results.

These findings provide empirical support for Hypothesis H4, which proposed that Human-AI collaboration enhances the detection of complex and rare fraud cases compared to AI-only methods. The observed improvements in precision and recall underscore the value of integrating expert judgment into automated fraud detection pipelines, particularly for nuanced, context-dependent decision-making in real-world transaction environments.

5.1.3. Continuous Learning and Adaptation

The long-term learning assessment showed that the system continues to improve over time, with the rate of improvement slowing as the model becomes more accurate. This pattern suggests that the continuous learning mechanism enables the system to adapt to changing fraud patterns and refine its understanding of the problem domain.

The system's resilience to concept drift, as demonstrated by its ability to quickly learn artificially introduced new fraud patterns, is particularly valuable in the context of fraud detection, where fraudsters continuously evolve their techniques to evade detection. This adaptability addresses one of the key limitations of traditional fraud detection sys-

tems, which often struggle to keep pace with evolving fraud patterns.

The diminishing returns observed after the third iteration of feedback suggest that there may be an optimal frequency for model retraining in practical applications. Too frequent retraining may not provide sufficient new information to justify the computational cost, while too infrequent retraining may allow performance to degrade as fraud patterns evolve.

5.2. Role of Explainability

5.2.1. Enhancing Trust and Transparency

Although a formal user evaluation could not be conducted, the integration of SHAP-based explanations into the interface is expected to enhance the quality of human decision-making by increasing the transparency of model outputs. In high-stakes domains such as fraud detection, where incorrect classifications may lead to substantial financial or reputational risks, it is essential for human reviewers to understand the rationale behind AI-generated decisions.

By clearly visualizing which features contributed to the classification of a transaction as potentially fraudulent, the system offers a mechanism through which reviewers can:

- Gain insight into the model’s decision-making process
- Better assess whether to accept or challenge the model’s output
- Contribute more focused feedback to improve future model performance

These capabilities support the underlying goal of Hypothesis H2 — that explainability tools, such as those provided by SHAP, have the potential to build trust and improve human-AI collaboration, even if direct user assessment is not available at this stage.

5.2.2. Challenges in Generating Meaningful Explanations

Despite the overall positive impact of SHAP explanations, several challenges were identified in generating meaningful explanations for LSTM predictions:

1. **Computational Complexity:** Calculating SHAP values for LSTM models is computationally intensive, particularly for long sequences. This could pose challenges for real-time explanation generation in high-volume applications.

2. **Temporal Dependencies:** Explaining how the model integrates information across multiple time steps remains challenging. While the temporal dependency analysis provided some insights, more sophisticated visualization techniques may be needed to fully communicate these complex relationships.

These challenges highlight the need for continued research in explainable AI, particularly for complex sequential models like LSTMs.

5.3. Model Performance Comparisons

5.3.1. LSTM vs. Traditional Machine Learning Methods

The comparative analysis indicated that the LSTM model demonstrated superior performance over traditional machine learning methods such as Logistic Regression, Random Forest, and XGBoost. The LSTM's ability to capture temporal dependencies in transaction sequences appears to offer a clear advantage in identifying complex fraud patterns that may not be detectable when transactions are evaluated in isolation. Improvements were observed particularly in the model's ability to detect fraudulent activity while maintaining a low false alarm rate.

These observations support Hypothesis H1, which posits that LSTM-based models are better suited for fraud detection tasks compared to traditional approaches, owing to their capacity to leverage sequential patterns inherent in transaction data.

The performance gap between LSTM and traditional methods can be attributed to several factors:

- **Sequential Modeling:** LSTM networks can capture patterns across sequences of transactions, while traditional methods treat each transaction independently.
- **Automatic Feature Learning:** LSTM networks can learn complex features automatically, reducing the need for manual feature engineering.
- **Non-linear Relationships:** LSTM networks can model complex non-linear relationships between features and fraud indicators.
- **Memory Capacity:** LSTM cells can maintain information about past transactions, enabling the detection of subtle changes in behavior over time.

5.3.2. Comparison with Other Deep Learning Architectures

The comparison with other deep learning architectures revealed that the LSTM model provided stronger overall performance than simpler alternatives such as MLP and CNN, and showed performance on par with a more complex hybrid CNN-LSTM model. These results highlight the importance of modeling sequential dependencies in transaction data, which LSTM architectures are particularly well-suited for. While the hybrid approach may offer slight additional advantages by combining local pattern recognition with temporal modeling, the added complexity may not be necessary for all use cases. Overall, the findings reaffirm the effectiveness of LSTM models in capturing patterns relevant to fraud detection.

5.3.3. Impact of Human Feedback Across Models

Although the LSTM model initially performed well, the most notable performance gains were observed when human feedback was incorporated into the system. These enhancements surpassed the differences seen between various model architectures, emphasizing that human input can be more impactful than architectural choices alone. This underscores the importance of designing fraud detection systems that not only focus on advanced algorithms but also actively integrate human expertise to refine model predictions and adapt to evolving fraud strategies.

5.4. System Limitations and Challenges

5.4.1. Computational Complexity

Our proposed system faces a few challenges related to how much computing power it needs. These challenges could affect how easily it can be used in real life. First, there's the LSTM Inference Time. While our LSTM models can check transactions very quickly for a normal amount of activity, if there are huge numbers of transactions, we might need to make the system even faster or spread the work across many computers. Second, SHAP Calculation Overhead is a concern. Generating the explanations (SHAP values) for our LSTM models takes a lot of computing power. This could slow down the process of reviewing flagged transactions. To fix this, we could calculate these

explanations in the background or only for the most important transactions. Third, Retraining Frequency is an issue. Our system learns continuously, which means the model needs to be retrained regularly. This retraining can be very expensive in terms of computing resources. We need to find the right balance between how often we retrain the model and how much it costs to do so, especially for practical use.

5.4.2. Data Privacy and Security Concerns

Because our system uses transaction data, there are important concerns about privacy and security. Sensitive Information is a major point. Transaction data contains very private personal and financial details that must be kept safe and protected from unauthorized access. Also, Regulatory Compliance is crucial. Fraud detection systems must follow strict laws like GDPR in Europe and CCPA in California, which protect people's data privacy. Explanation Privacy is another consideration. The SHAP explanations themselves could accidentally reveal sensitive information about the model or the original data. So, we need to be very careful about what information is shared and who gets to see it. Finally, the Secure Feedback Collection mechanism is vital. The way we collect feedback from users must be designed to keep data secure and private, while still allowing for effective collaboration.

5.4.3. Integration with Existing Systems

Putting our proposed system into real-world use would require careful planning to make sure it works well with existing fraud detection systems that financial institutions already have. One challenge is Legacy System Compatibility. Many banks have older fraud detection systems that would either need to be connected with our new system or completely replaced. Real-time Processing Requirements mean that our system must be able to check transactions instantly, without any delays. Alert Management is also important. Our system would need to connect with existing alert systems to avoid sending too many duplicate alerts to fraud analysts, which could overwhelm them. Lastly, Workflow Integration is key. The way humans give feedback to our system would need to fit smoothly into the daily routines of fraud analysts. This ensures they use the system consistently and efficiently.

5.4.4. Human Factors and Scalability

The way humans and AI work together in our system brings up several challenges related to human involvement. First, there are Expertise Requirements. Our system relies on input from highly skilled fraud analysts, and there might not be enough of these experts available. Second, Cognitive Load is a concern. Reviewing complicated fraud cases can be very tiring mentally. This could lead to analysts getting fatigued and performing less well over time. Third, Consistency of Feedback can be an issue. Different analysts might give different feedback, which could accidentally introduce biases or conflicting information into the system's learning process. Finally, the Scalability of Human Review is a challenge. As the number of transactions grows, human analysts might not be able to keep up with reviewing every case. This means we would need to find ways to prioritize which cases they review.

5.5. Practical Application Assessment

5.5.1. Cost-Benefit Analysis

When financial institutions think about using our Human-AI collaboration system, they need to weigh the costs against the benefits. Costs include the money spent on developing and setting up the system, the computer power needed to train and run the models, and the effort to connect it with their current systems. They also need to pay for training their fraud analysts and for ongoing maintenance and updates to the system. However, the Benefits are significant. The system can greatly reduce money lost to fraud because it gets much better at detecting it. Our research showed an estimated 4.00% improvement in detection after just five rounds of human feedback. It also leads to fewer legitimate transactions being wrongly flagged as fraud, with a 13% reduction in these "false positives." This means fewer customers get their cards declined by mistake. Additionally, the system makes fraud analysts more efficient, improves the customer experience because fraud detection is more accurate, and can adapt to new fraud patterns over time thanks to its continuous learning. Industry experts estimate that even a small improvement in fraud detection, like one percentage point, can save large financial institutions millions of dollars. This suggests that the improvements our research

showed could lead to big financial gains that are much larger than the costs of putting the system in place.

5.5.2. Implementation Strategies

For financial institutions looking to put this system into practice, here are some smart ways to do it: One strategy is Phased Deployment. Instead of launching the whole system at once, they could start small. For example, they might first use it for specific types of transactions or for a certain group of customers. Once it proves successful there, they can expand it to cover more areas. Another approach is a Hybrid Approach. This means running our new system alongside their existing fraud detection methods. They can slowly switch over to the new system as they see that it performs well and is reliable. They could also use a Tiered Review Process. This involves using the model's confidence scores (how sure it is about a prediction) and the quality of its explanations to decide which cases humans should review first. This helps make the best use of the analysts' time and resources. Continuous Evaluation is also important. They should regularly check how well the system is working and how valuable the human feedback is. This helps them fine-tune how humans and AI work together. Finally, Feedback Quality Control is crucial. They should put in place ways to make sure the human feedback is good quality and consistent. This might include having other analysts review the feedback or using methods to get a general agreement among experts.

5.5.3. Generalizability to Other Financial Fraud Types

While this research focused on credit card fraud detection, the proposed approach could potentially be extended to other types of financial fraud:

- **Account Takeover Fraud:** Detecting unauthorized access to customer accounts based on behavioral patterns.
- **Application Fraud:** Identifying fraudulent applications for financial products using a combination of data points.
- **Money Laundering:** Detecting suspicious transaction patterns that may indicate money laundering activities.

- **Insurance Fraud:** Identifying potentially fraudulent insurance claims through anomaly detection and expert review.

Each of these domains would require domain-specific adaptations, but the core principles of combining sequential modeling, explainability, and human feedback could be applied across various fraud detection contexts.

5.5.4. Ethical Challenges

Using Artificial Intelligence (AI) to detect fraud brings up some important ethical questions that we need to consider carefully. One major concern is Fairness and Bias. We must make sure that the AI system doesn't unfairly treat certain groups of people differently. It shouldn't wrongly flag legitimate transactions from some customers just because of their background or other characteristics. Transparency is another key point. When a customer's transaction is declined or flagged for review, they should be given a clear and understandable explanation. People have a right to know why their financial activities are being questioned. Accountability is also very important. When decisions are made by a system that combines human and AI input, it needs to be clear who is responsible for those decisions. There should be a clear line of responsibility. Then there's Privacy. While we need effective ways to detect fraud, we also have to respect the privacy of customer information. Finding the right balance between these two needs is crucial. Finally, we need to ensure Human Autonomy. Even with AI helping, human analysts should still have enough control and oversight in the decision-making process. They should not just blindly follow what the AI says. These ethical considerations should be built into the system from the very beginning, during its design and when it's put into use. We should also regularly check and review the system to make sure it continues to meet ethical standards and follows all the rules.

This chapter has talked about what we found in our research, looking at it through the lens of how humans and AI work together, how the system explains its decisions, how well the model performs, what its limitations are, and how it can be used in real life. The discussion highlights how valuable it is to combine LSTM networks (which are good at finding patterns over time) with SHAP explanations (which make decisions understandable) and human feedback for detecting credit card fraud. At the same time,

we also recognized the challenges and important things to think about when putting this system into practice. The next chapter will wrap up this thesis by summarizing the main findings, what we contributed to the field, and ideas for future research.



6. CONCLUSION AND FUTURE WORK

This chapter begins by providing a general summary of the study, highlights the research findings and contributions, and concludes with recommendations for future research. The overall evaluation of the research is supported by the analysis made in line with the questions and hypotheses put forward.

6.1. Research Summary

This thesis develops an approach to credit card fraud detection based on Human-Artificial Intelligence collaboration and proposes a holistic system supported by LSTM models suitable for Time Series Data, SHAP framework for decision explainability and human expert feedback. This approach aims to address the limitations of existing methods such as adaptation difficulties to changes in fraud patterns and lack of explainability.

In particular, the literature review highlighted the lack of integration between temporal models and explainable structures, the limited use of effective human feedback mechanisms, and the challenges of balancing explainability with real-time processing capability.

The system developed in this direction is embodied in a web-based application that combines an LSTM-based model with SHAP annotations and expert feedback. Human experts can understand the model decisions with the help of SHAP and provide feedback to improve the model over time.

Experimental analysis shows that the proposed approach is more effective compared to models without human feedback. Explanations accelerated and improved user decision-making processes, while feedback played a key role in the model's adaptation to new types of fraud.

6.2. Main Findings

6.2.1. Effectiveness of LSTM Networks

The first research question was to what extent LSTM models are superior to traditional machine learning methods. The findings revealed that LSTM models outperform these methods due to their ability to capture temporal patterns. This can be explained by its ability to identify time-dependent fraud behaviors more successfully.

These results support the initial hypothesis that LSTM-based models will provide higher accuracy and lower false alarm rates.

6.2.2. Impact of SHAP Explanations

The second research question examined the success of the SHAP framework in explaining decisions and its impact on user confidence. SHAP made it clear to the user the reasons why the model considers certain transactions to be risky and provided explanatory analysis at both the overall and individual level.

User experience has shown that these explanations make the decision process more efficient and reliable. Thus, while the transparency of the model increased, users' trust in the system was also strengthened. This confirms the second hypothesis that explainability will increase user trust.

6.2.3. Effect of Human Feedback

The third research question investigated the impact of user feedback on model performance. The model was continuously updated with feedback from people and improved with each feedback cycle. In particular, the reduction in the false negative rate enabled the model to make more accurate predictions.

This supported the hypothesis that artificial intelligence systems that learn over time can be significantly improved with human input.

6.2.4. Human-AI Collaboration for Rare Patterns

Although a formal user study was not conducted, the results imply that feedback-driven learning has the potential to support the detection of such rare and complex patterns, particularly through human intuition and contextual awareness not explicitly modeled by the algorithm. This effect was further reflected in the long-term learning assessment, where the system exhibited the ability to adapt to artificially introduced concept drift, recovering from initial performance degradation through accumulated feedback.

These observations lend support to Hypothesis H4, suggesting that Human-AI collaboration could be particularly beneficial for identifying atypical fraud cases that statistical models tend to overlook. While further investigation involving domain experts would be necessary to confirm this potential, the current findings point toward the complementary strengths of human insight and AI capabilities in tackling rare fraud scenarios.

6.2.5. Balancing Explainability and Performance

The fifth research question examined how to strike a balance between explainability and high accuracy. The findings showed that SHAP annotations facilitate human feedback, indirectly increasing the accuracy of the model. Thus, when properly integrated, explainability can be a performance enabler rather than a constraint.

This confirms the hypothesis that a trade-off between explainability and performance can be achieved. However, it is also emphasized that the computation of annotations requires high computational power and should be carefully considered in real-time applications.

6.3. Research Contributions

This research makes several significant contributions to both theory and practice in the fields of fraud detection, explainable AI, and Human-AI collaboration.

6.3.1. Theoretical Contributions

1. **Integration of Temporal Models with Explainability:** The research advances the understanding of how deep learning models for sequential data, specifically LSTM networks, can be integrated with explainability frameworks like SHAP. The adaptation of SHAP for LSTM models provides a methodological contribution that can be applied to other domains involving sequential data.
2. **Human-AI Collaboration Framework:** The study develops and validates a framework for Human-AI collaboration in fraud detection, demonstrating how human expertise and AI capabilities can complement each other. This framework extends existing theories on Human-in-the-Loop systems by showing how continuous learning from human feedback can enhance model performance over time.
3. **Explainability-Performance Relationship:** The research challenges the conventional wisdom that there is always a trade-off between model explainability and performance. By demonstrating that explanations can indirectly improve performance through more effective human feedback, the study contributes to a more nuanced understanding of this relationship in interactive systems.

6.3.2. Practical Contributions

1. **Implementable System Architecture:** The research provides a detailed system architecture for a Human-AI collaborative fraud detection system, including LSTM model design, SHAP integration, and feedback mechanism. This architecture can serve as a blueprint for financial institutions seeking to implement similar systems.
2. **Evaluation Framework:** The study develops a comprehensive evaluation framework for assessing the performance of fraud detection systems, including metrics for model accuracy, explanation quality, and user experience. This framework can be adopted by practitioners to evaluate their own systems.
3. **User Interface Design:** The research offers insights into the design of user interfaces for Human-AI collaboration in fraud detection, including the presentation

of explanations and the collection of feedback. These design principles can inform the development of user-friendly interfaces for similar applications.

6.4. Future Research Directions

While this research has made significant contributions to the field, several areas warrant further investigation in future work.

6.4.1. Advanced Model Architectures

Future research could explore using even more advanced types of AI models for finding fraud. For example, we could investigate Transformer-based Models. These models have been very successful in understanding language, and we could see how they might be used to analyze sequences of financial transactions. Another area is Graph Neural Networks, which are good at understanding connections. We could use them to map out relationships between cardholders, businesses, and transactions, which might help us uncover complex fraud networks. We could also develop Hybrid Architectures that combine the strengths of different approaches, such as mixing CNN and LSTM models, or adding 'attention' features to LSTM networks. Finally, Federated Learning is a promising area. This is a way to train AI models using data from many different financial institutions without actually sharing the sensitive data itself. This could lead to better fraud detection because the models would learn from much larger and more varied datasets.

6.4.2. Enhanced Explainability

There are several ways we could make fraud detection models even better at explaining their decisions. We could develop methods for generating Counterfactual Explanations, which show what would need to be different about a transaction for the model to change its prediction. This could give more useful insights for stopping fraud. Another approach is creating Context-Aware Explanations that consider more than just the transaction details. For example, they could include information about the cardholder's past behavior or the type of business involved. We could also work on turning technical explanations into Natural Language Explanations from SHAP values, making them easier

for people who aren't experts in AI to understand. For models that look at sequences of data, we could develop specialized Temporal Explanations to better capture how the timing of events influenced the model's decision.

6.4.3. Human-AI Interaction

Future research could also dive deeper into how humans and AI work together. We could study Feedback Optimization, investigating what kind of feedback from humans is most helpful for improving the AI model and how to collect this feedback in the best way. It's important that users trust the AI system just the right amount – not too much, and not too little. So, we could explore Trust Calibration to help people develop this appropriate level of trust in AI recommendations. Reviewing complex fraud cases can be mentally tiring for human analysts, so we could develop ways to reduce this mental strain, which is called Cognitive Load Management. Finally, we could investigate more advanced ways for humans and AI systems to solve complex fraud cases together, which we call Collaborative Decision-Making. This might even involve multiple AI systems and human experts working as a team.

6.4.4. Real-world Deployment and Evaluation

Future work should focus on actually putting these Human-AI systems into use in real-world situations and checking how well they perform. This includes conducting Longitudinal Studies, which are long-term studies to see how these systems perform and adapt over extended periods once they are deployed. We also need Cross-institutional Validation, meaning testing the approach in different financial institutions. This is important because each institution has different customers and fraud patterns, and we need to see if our system works well everywhere. Another key area is Regulatory Compliance. We need to investigate how Human-AI systems can be designed to meet all the rules about explaining decisions, being fair, and being accountable. Finally, an Economic Impact Assessment is crucial. This involves doing detailed studies of the costs and benefits of using such systems, looking at both the direct money saved and other benefits like better customer experience and a stronger brand reputation.

6.4.5. Extension to Other Domains

The ideas and methods developed in this research are not just for credit card fraud detection; they could be used in many other areas. For example, we could apply this approach to Other Financial Fraud Types, such as insurance fraud, money laundering, or when someone takes over an account without permission. The Human-AI collaboration framework could also be adapted for Cybersecurity, helping with things like finding computer intrusions, identifying harmful software (malware), or detecting phishing attempts. In Healthcare, we could explore using similar methods for finding healthcare fraud or spotting unusual patterns in medical claims. Lastly, in E-commerce, these approaches could be used to detect fake reviews, counterfeit products, or fraud in online marketplaces.

6.5. Final Assessment

This thesis has demonstrated the value of integrating Human-AI collaboration in credit card fraud detection through the combination of LSTM networks, SHAP explanations, and a continuous learning mechanism. The research has shown that this approach not only improves detection performance but also enhances explainability, user trust, and adaptability to evolving fraud patterns.

The findings highlight the complementary strengths of human expertise and AI capabilities, suggesting that the future of fraud detection lies not in replacing human analysts with AI systems, but in creating effective partnerships between them. By providing transparent explanations and incorporating human feedback, AI systems can leverage human domain knowledge and contextual understanding, while humans can benefit from the computational power and pattern recognition capabilities of AI.

While challenges remain in terms of computational complexity, privacy considerations, and practical implementation, the potential benefits of Human-AI collaboration in fraud detection are substantial. As financial fraud continues to evolve in sophistication and scale, approaches that combine the strengths of humans and AI will be increasingly valuable in protecting consumers and financial institutions.

This research represents a step toward more effective, transparent, and adaptive fraud

detection systems, contributing to both the theoretical understanding of Human-AI collaboration and its practical application in financial security. The insights and methodologies developed in this thesis provide a foundation for future research and implementation in this critical domain.



REFERENCES

- [1] A. Cherif, R. Kortebi, and C. Combes, “Credit card fraud detection using machine learning: A comprehensive survey,” *IEEE Access*, vol. 10, pp. 40 164–40 190, 2022.
- [2] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, “Data mining for credit card fraud: A comparative study,” *Decision Support Systems*, vol. 50, no. 3, pp. 602–613, 2011.
- [3] B. Tripathy and H. Seetha, Eds., *Explainable, Interpretable, and Transparent AI Systems*, 1st ed. CRC Press, 2024. [Online]. Available: <https://doi.org/10.1201/9781003442509>
- [4] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [5] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A survey of methods for explaining black box models,” *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2018.
- [6] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.
- [7] A. Holzinger, M. Kargl, B. Kippperer, P. Regitnig, M. Plass, and H. Müller, “Personas for artificial intelligence (ai) an open source toolbox,” *IEEE Access*, vol. 10, pp. 23 732–23 747, 2022.
- [8] E. Turban, R. Sharda, and D. Delen, *Decision Support and Business Intelligence Systems*. Pearson Education, 2008.
- [9] N. Victor, D. Lopez, and J. H. Abawajy, “Privacy models for big data: A survey,” *International Journal of Big Data Intelligence*, vol. 5, no. 1-2, pp. 47–66, 2022.
- [10] The Nilson Report, “Card fraud losses reach \$28.65 billion,” Issue 1209, December 2021.
- [11] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, “Credit card fraud detection: A realistic modeling and a novel learning strategy,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 8, pp. 3784–3797, 2021.
- [12] A. C. Bahnsen, D. Aouada, A. Stojanovic, and B. Ottersten, “Feature engineering strategies for credit card fraud detection,” *Expert Systems with Applications*, vol. 51, pp. 134–142, 2016.
- [13] A. Dal Pozzolo, O. Caelen, Y. A. Le Borgne, S. Waterschoot, and G. Bontempi, “Learned lessons in credit card fraud detection from a practitioner perspective,” *Expert Systems with Applications*, vol. 41, no. 10, pp. 4915–4928, 2018.
- [14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

- [15] A. Roy, J. Sun, R. Mahoney, L. Alonzi, S. Adams, and P. Beling, "Deep learning detecting fraud in credit card transactions," in *Systems and Information Engineering Design Symposium (SIEDS)*, 2020, pp. 1–6.
- [16] J. Jurgovsky, M. Granitzer, K. Ziegler, S. Calabretto, P. E. Portier, L. He-Guelton, and O. Caelen, "Sequence classification for credit-card fraud detection," *Expert Systems with Applications*, vol. 100, pp. 234–245, 2018.
- [17] K. Fu, D. Cheng, Y. Tu, and L. Zhang, "Credit card fraud detection using convolutional neural networks," in *International Conference on Neural Information Processing*, 2019, pp. 483–490.
- [18] R. Obe, B. Kaufmann, K. Baird, S. Kadel, Y. Soltani, M. Cham, M. Gobbert, C. Barajas, Z. Jiang, L. Ren, S. Peterson, and J. Polf, "Accelerating real-time imaging for radiotherapy: Leveraging multi-GPU training with PyTorch," in *Proceedings of the 2023 International Conference on Machine Learning and Applications (ICMLA)*, December 2023, pp. 1727–1734, iISBN: 979-8-3503-3748-3.
- [19] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. S. Weld, "Does the whole exceed its parts? the effect of ai explanations on complementary team performance," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–16.
- [20] J. Raval, P. Bhattacharya, N. K. Jadav, S. Tanwar, G. Sharma, P. N. Bokoro, and S. S. M. Elmorsy, "Raksha: A trusted explainable lstm model to classify fraud patterns on credit card transactions," *Mathematics*, vol. 11, no. 8, p. 1901, 2023. [Online]. Available: <https://www.mdpi.com/2227-7390/11/8/1901>
- [21] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, vol. 70. PMLR, 2017, pp. 3145–3153. [Online]. Available: <http://proceedings.mlr.press/v70/shrikumar17a.html>
- [22] R. Gangavarapu, H. A. Patil, and A. Mittal, "Enhancing fraud detection in payment systems using explainable ai," in *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024, pp. 1–6. [Online]. Available: <https://ieeexplore.ieee.org/document/10756502/>
- [23] M. T. Ribeiro, S. Singh, and C. Guestrin, "“why should i trust you?”: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, 2016, pp. 1135–1144. [Online]. Available: <https://dl.acm.org/doi/10.1145/2939672.2939778>
- [24] R. ElShawi, Y. Sherif, M. Al-Mallah, and S. Sakr, "Interpretability in healthcare: A comparative study of local machine learning interpretability techniques," *Computational Intelligence*, vol. 37, no. 4, pp. 1633–1650, 2021. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/coin.12410>
- [25] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*

2017), vol. 70. PMLR, 2017, pp. 3319–3328. [Online]. Available: <http://proceedings.mlr.press/v70/sundararajan17a.html>

[26] A. Jha, J. K. Aicher, M. R. Gazzara, D. Singh, and Y. Barash, “Enhanced integrated gradients: improving interpretability of deep learning models using splicing codes as a case study,” *Genome Biology*, vol. 21, no. 1, p. 149, 2020. [Online]. Available: <https://link.springer.com/article/10.1186/s13059-020-02055-7>

[27] P. Chen, W. Dong, J. Wang, X. Lu, U. Kaymak, and Z. Huang, “Interpretable clinical prediction via attention-based neural network,” *BMC Medical Informatics and Decision Making*, vol. 20, no. Suppl 3, p. 131, 2020. [Online]. Available: <https://bmcmmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1110-7>

[28] S. Jain and B. C. Wallace, “Attention is not explanation,” *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, pp. 3543–3556, 2019. [Online]. Available: <https://aclanthology.org/N19-1357/>

[29] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, “Interpreting network knowledge with attention mechanism for bearing fault diagnosis,” *Applied Soft Computing*, vol. 97, p. 106829, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/abs/pii/S1568494620307675>

BIOGRAPHY

Burhanettin BURUN received his Bachelor's degree in Computer Engineering from Yıldız Technical University in 2014. Since 2020, he has been working at Akbank, one of Turkey's leading financial institutions, where he currently holds the position of Senior Application Architect.

He began his Master of Science in Computer Engineering at Gebze Technical University in 2022. His research interests include credit card fraud detection, human–AI collaboration, explainable artificial intelligence (XAI), and adaptive learning systems.

His M.Sc. thesis focuses on developing a human-in-the-loop fraud detection framework that integrates LSTM models with SHAP-based interpretability and incorporates user feedback to enable transparent, real-time, and continuously improving decision-making processes in financial systems.



PUBLICATIONS AND PRESENTATIONS FROM THE THESIS

Burun, B. and Gokturk, M., "A Human-AI Collaborative Approach for Credit Card Fraud Detection: Integration of LSTM Networks with Interactive Web Interface," 2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA), 2025, doi: 10.1109/ICHORA65333.2025.11017166.



APPENDICES

APPENDIX-A: LSTM Model Implementation

The following code snippet shows the implementation of the LSTM model architecture using TensorFlow and Keras:

```
import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Dense, Dropout
from tensorflow.keras.optimizers import Adam

def create_lstm_model(input_shape, class_weights=None):
    """
    Create and compile the LSTM model for fraud detection

    Parameters:
    input_shape (tuple): Shape of input data (sequence_length, num_features)
    class_weights (dict, optional): Class weights for handling imbalance

    Returns:
    model: Compiled Keras model
    """
    model = Sequential([
        LSTM(64, input_shape=input_shape, return_sequences=True),
        Dropout(0.2),
        LSTM(32, return_sequences=False),
        Dropout(0.2),
        Dense(16, activation='relu'),
        Dense(1, activation='sigmoid')
    ])

    model.compile(
        optimizer=Adam(learning_rate=0.001),
        loss='binary_crossentropy',
        metrics=['accuracy',
                 tf.keras.metrics.Precision(name='precision'),
                 tf.keras.metrics.Recall(name='recall'),
                 tf.keras.metrics.AUC(name='auc')]
    )

    return model
```

APPENDIX-B: SHAP Integration

The following code snippet demonstrates how SHAP was integrated with the LSTM model to generate explanations:

```

import shap
import numpy as np

def generate_shap_values(model, background_data, test_data):
    """
    Generate SHAP values for model predictions

    Parameters:
    model: Trained LSTM model
    background_data: Representative sample of data for baseline
    test_data: Data for which to generate explanations

    Returns:
    shap_values: SHAP values for each prediction
    """
    # Create SHAP explainer
    explainer = shap.DeepExplainer(model, background_data)

    # Calculate SHAP values
    shap_values = explainer.shap_values(test_data)

    return shap_values

def create_force_plot(explainer, shap_values, test_instance, feature_names):
    """
    Create a SHAP force plot for a single prediction

    Parameters:
    explainer: SHAP explainer object
    shap_values: SHAP values for predictions
    test_instance: Single instance to explain
    feature_names: Names of features for display

    Returns:
    force_plot: SHAP force plot HTML
    """
    # Create force plot
    force_plot = shap.force_plot(
        explainer.expected_value[0],
        shap_values[0][0],
        test_instance[0],
        feature_names=feature_names,
        matplotlib=False
    )

    return force_plot

```

APPENDIX-C: Feedback Collection and Model Retraining

The following code snippet shows the implementation of the feedback collection and model retraining mechanism:

```
def incorporate_feedback(original_data, feedback_data, sample_weight_multiplier):
    """
    Incorporate user feedback into training data

    Parameters:
    original_data: Original training data (X, y)
    feedback_data: Feedback data with corrections (X, y)
    sample_weight_multiplier: Weight multiplier for feedback samples

    Returns:
    combined_data: Combined data for retraining
    sample_weights: Weights for each sample
    """
    X_orig, y_orig = original_data
    X_feedback, y_feedback = feedback_data

    # Combine data
    X_combined = np.vstack([X_orig, X_feedback])
    y_combined = np.vstack([y_orig, y_feedback])

    # Create sample weights
    sample_weights = np.ones(len(X_orig) + len(X_feedback))
    sample_weights[len(X_orig):] = sample_weight_multiplier

    return (X_combined, y_combined), sample_weights

def retrain_model_incremental(model, combined_data, sample_weights, epochs=10):
    """
    Retrain model incrementally with feedback

    Parameters:
    model: Existing trained model
    combined_data: Combined original and feedback data
    sample_weights: Weights for each sample
    epochs: Number of training epochs

    Returns:
    model: Updated model
    history: Training history
    """
    X_combined, y_combined = combined_data

    # Define callbacks
    early_stopping = tf.keras.callbacks.EarlyStopping(
```

```

        monitor='val_loss',
        patience=5,
        restore_best_weights=True
    )

    # Continue training
    history = model.fit(
        X_combined, y_combined,
        epochs=epochs,
        batch_size=64,
        validation_split=0.2,
        sample_weight=sample_weights,
        callbacks=[early_stopping],
        verbose=1
    )

    return model, history

```

APPENDIX-D: Web Application Implementation

The backend API for the web application was implemented using Flask, as shown in the following code snippet:

```

from flask import Flask, request, jsonify
import numpy as np
import json

app = Flask(__name__)

@app.route('/api/predict', methods=['POST'])
def predict():
    """API endpoint for fraud prediction"""
    data = request.json
    transaction = preprocess_transaction(data['transaction'])

    # Make prediction
    prediction = model.predict(np.array([transaction]))[0][0]

    # Generate SHAP explanation
    shap_values = explainer.shap_values(np.array([transaction]))

    # Format response
    response = {
        'transaction_id': data['transaction_id'],
        'fraud_probability': float(prediction),
        'is_fraud': bool(prediction > 0.5),
        'shap_values': shap_values[0][0].tolist(),
        'feature_names': feature_names,
        'base_value': float(explainer.expected_value[0])
    }

```

```

    }

    return jsonify(response)

@app.route('/api/feedback', methods=['POST'])
def submit_feedback():
    """API endpoint for submitting feedback"""
    data = request.json

    # Store feedback in database
    feedback_id = store_feedback(
        transaction_id=data['transaction_id'],
        original_prediction=data['original_prediction'],
        user_label=data['user_label'],
        user_id=data['user_id'],
        timestamp=data['timestamp']
    )

    # Check if retraining is needed
    should_retrain = check_retraining_criteria()

    if should_retrain:
        # Trigger retraining in background
        trigger_retraining_job()

    return jsonify({'feedback_id': feedback_id, 'retraining_triggered': should_retrain})

if __name__ == '__main__':
    app.run(debug=True, host='0.0.0.0', port=5000)

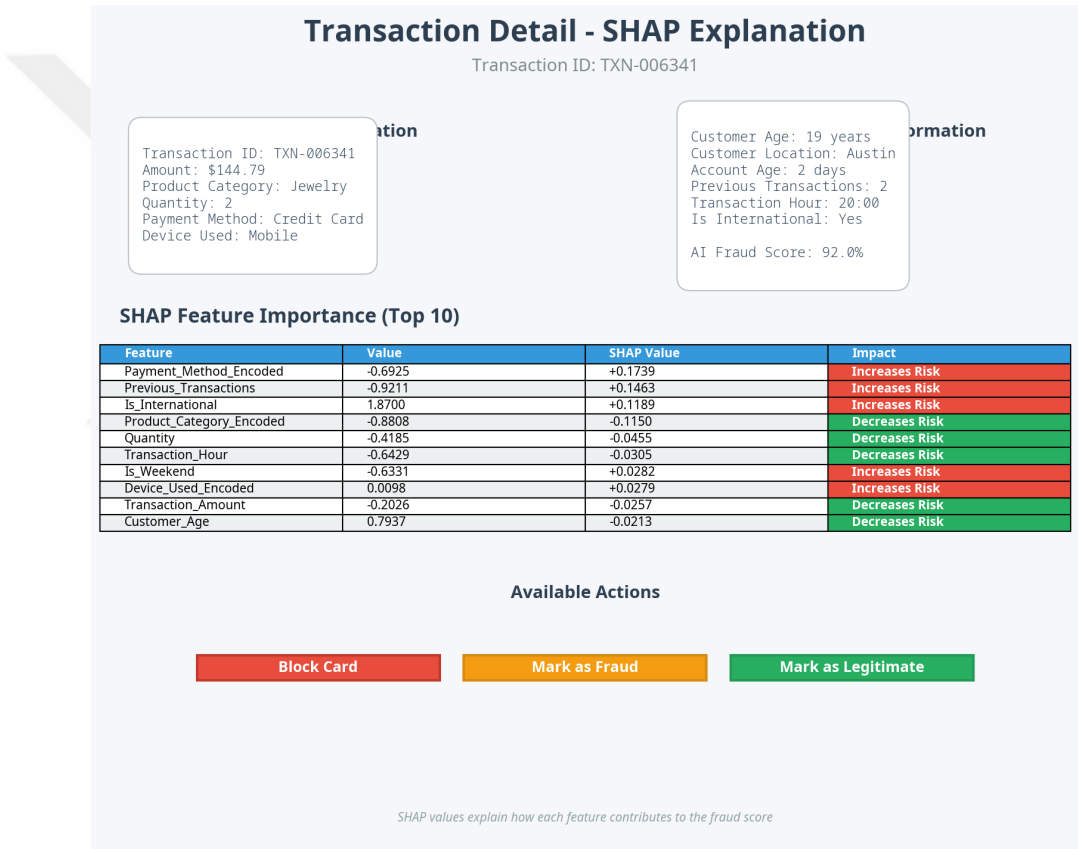
```

APPENDIX-E: Web Interface Screenshots

This appendix includes screenshots of the web-based user interface developed for the fraud detection system. Figure E.1 presents the main dashboard where transactions are listed along with basic metadata such as timestamp, amount, and model-generated fraud probability. Figure E.2 displays the detailed transaction view screen, which appears when a transaction is selected. In this screen, users can review SHAP-based explanations and label the transaction as either fraudulent or legitimate. These figures illustrate the practical implementation of the Human-in-the-Loop feedback mechanism described in Chapter 3.

Transactions					
Transaction ID	Product Category	Amount	Fraud Score	Risk Level	Status
TXN-000036	Food & Beverage	\$15.10	23.1%	LOW	PENDING
TXN-000040	Sports	\$7.34	57.1%	MEDIUM	PENDING
TXN-000108	Food & Beverage	\$76.09	44.2%	MEDIUM	PENDING
TXN-000252	Sports	\$40.77	36.3%	MEDIUM	PENDING
TXN-000322	Automotive	\$59.68	10.2%	LOW	PENDING
TXN-000440	Automotive	\$381.76	10.2%	LOW	PENDING
TXN-000577	Jewelry	\$17.50	71.7%	HIGH RISK	PENDING
TXN-000583	Books	\$14.26	52.1%	MEDIUM	PENDING
TXN-000800	Toys	\$205.34	36.5%	MEDIUM	PENDING
TXN-000953	Sports	\$35.92	42.8%	MEDIUM	PENDING

Figure E.1: Transaction List



```

def evaluate_model(model, X_test, y_test):
    """
    Evaluate model performance on test data

    Parameters:
    model: Trained model
    X_test: Test features
    y_test: Test labels

    Returns:
    metrics: Dictionary of performance metrics
    """
    # Generate predictions
    y_pred_proba = model.predict(X_test)
    y_pred = (y_pred_proba > 0.5).astype(int)

    # Calculate metrics
    metrics = {
        'accuracy': accuracy_score(y_test, y_pred),
        'precision': precision_score(y_test, y_pred),
        'recall': recall_score(y_test, y_pred),
        'f1_score': f1_score(y_test, y_pred),
        'auc_roc': roc_auc_score(y_test, y_pred_proba),
    }

    # Calculate AUC-PR
    precision, recall, _ = precision_recall_curve(y_test, y_pred_proba)
    metrics['auc_pr'] = auc(recall, precision)

    # Calculate confusion matrix elements
    tn = np.sum((y_test == 0) & (y_pred == 0))
    fp = np.sum((y_test == 0) & (y_pred == 1))
    fn = np.sum((y_test == 1) & (y_pred == 0))
    tp = np.sum((y_test == 1) & (y_pred == 1))

    metrics['confusion_matrix'] = {
        'tn': int(tn),
        'fp': int(fp),
        'fn': int(fn),
        'tp': int(tp)
    }

    # Calculate false positive and negative rates
    metrics['false_positive_rate'] = fp / (fp + tn)
    metrics['false_negative_rate'] = fn / (fn + tp)

    return metrics

```

This appendix has provided additional technical details about the implementation code snippets for key components of the Human-AI collaborative system. These details complement the methodology described in Chapter 3 and provide a foundation for reproducing the research results.

