



# Explainable AI (XAI) in Deep Learning Models for Credit Card Fraud Detection

Thong Minh Lai

Student ID: U2259343

Department of Computer Science  
University of Huddersfield

Dr. Hyunkook Lee  
Professor and Supervisor

April 2025

# Contents

---

1. Introduction

2. Background

3. Methods

4. Results

5. Discussion

6. Conclusions

# **Introduction**

---

# Introduction

---

- Credit card fraud is a massive problem, with losses in the UK alone reaching £1.3 billion in 2021.<sup>1</sup>
- Deep Learning (DL) models are powerful for fraud detection, but their “black box” nature makes them hard to trust in high-stakes environments.
- My project tackles this by integrating Explainable AI (XAI) techniques into state-of-the-art DL models, focusing on local interpretability.
- All experiments are conducted on the Sparkov synthetic dataset, which is ideal for benchmarking fraud detection systems.<sup>2</sup>

## Project Aims

- Compare the explainability of multiple deep learning architectures for fraud detection.
- Integrate and evaluate XAI methods (SHAP, LIME, Anchors) for local explanations.
- Develop robust evaluation metrics for generated explanations.

<sup>1</sup>UK Finance, 2022.

<sup>2</sup>Grover et al., 2023.

## **Background**

---

# Background and Motivation

---

- Traditional fraud detection relied on hand-crafted rules, but fraudsters adapt quickly.<sup>3</sup>
- Deep Learning models (CNN, LSTM) can spot subtle, non-linear patterns in transaction data.
- However, financial institutions demand transparency for regulatory and trust reasons.<sup>4</sup>
- XAI methods help open up these black boxes, making model decisions understandable to humans.

## Why Local Explanations?

Local explanations help analysts understand *why* a specific transaction was flagged as fraud, which is crucial for real-world deployment.

---

<sup>3</sup>Sundararamaiah et al., 2024.

<sup>4</sup>Gilpin et al., 2018.

## **Methods**

---

# Dataset and Preprocessing

---

- **Dataset:** Sparkov synthetic data<sup>5</sup>, 1.2M transactions, 22 features, fraud rate  $\approx$  0.58%.
- **Preprocessing:**
  - Feature engineering (e.g., Haversine distance, age groups, temporal features).
  - Standardisation and encoding of categorical variables.
  - SMOTE<sup>6</sup> for balancing the highly imbalanced dataset.

---

<sup>5</sup>Harris, n.d.

<sup>6</sup>Bowyer et al., 2011.

# Model Architectures

---

## CNN Architecture

- Input: (15, 1) feature vector.
- Two Conv1D layers (64, 32 filters),  
batch norm, dropout.
- Dense layers with ReLU, final sigmoid  
for binary classification.
- **Parameters:** 60,065.

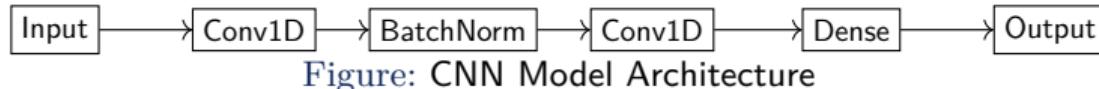


Figure: CNN Model Architecture

# Model Architectures

## CNN Architecture

- Input: (15, 1) feature vector.
- Two Conv1D layers (64, 32 filters), batch norm, dropout.
- Dense layers with ReLU, final sigmoid for binary classification.
- **Parameters:** 60,065.

## LSTM with Attention

- Input: (15, 1) feature vector.
- Lambda layer to expand dimensions.
- Two LSTM layers (50 units each) with dropout (0.3) and recurrent dropout (0.2).
- Custom attention mechanism.
- Dense output with sigmoid for binary classification.
- **Parameters:** 33,502.



Figure: LSTM Model Architecture<sup>7</sup>

<sup>7</sup>Ibtissam et al., 2021.

# Explainable AI (XAI) Techniques and Metrics

---

- **SHAP:** SHapley Additive exPlanations, provides both global and local feature attributions.
- **LIME:** Local Interpretable Model-agnostic Explanations, explains individual predictions with local surrogate models.
- **Anchors:** Rule-based, high-precision explanations for individual predictions.

## Evaluation Metrics

- **Faithfulness:** How well explanations reflect the model's true decision process.
- **Monotonicity:** Whether increasing a feature's value increases its importance.
- **Completeness:** How much of the model's behaviour is captured by the explanation.

# XAI Methods: SHAP

---

## How SHAP values are computed:

- Background dataset represents "average" feature values.
- Establishes baseline prediction for comparison.
- Measures feature impact by swapping actual values with background values.

---

## Algorithm SHAP for Credit Card Detection

---

```
1: Input: Trained model  $f$ , transaction  $x$ 
2: for each feature  $i$  in  $x$  do
3:   Initialize SHAP value  $\phi_i = 0$ 
4:   for each subset  $S$  of features not containing  $i$  do
5:     Create two samples:  $x_{S \cup \{i\}}$  and  $x_S$ 
6:     Compute marginal contribution:  $f(x_{S \cup \{i\}}) - f(x_S)$ 
7:     Weight the contribution based on subset size
8:     Add weighted contribution to  $\phi_i$ 
9:   end for
10: end for
11: Output: SHAP values  $\phi_1, \phi_2, \dots, \phi_n$  showing each feature's contribution to  $f(x)$ 
```

---

# XAI Methods: LIME

---

## How LIME Works

- Creates perturbed samples around the original transaction
- Weights samples by proximity to the original based on a kernel value
- Fits an interpretable model locally
- Shows feature contributions with confidence intervals

---

### Algorithm LIME for Credit Card Detection

---

- 1: **Input:** Trained model  $f$ , transaction  $x$
  - 2: Generate  $N$  perturbed samples around  $x$  by randomly changing feature values
  - 3: **for** each perturbed sample  $x'$  **do**
  - 4:     Predict  $f(x')$
  - 5:     Compute similarity between  $x$  and  $x'$
  - 6: **end for**
  - 7: Fit a simple interpretable model  $g$  (e.g., linear model) to predict  $f(x')$  using the perturbed samples, weighted by similarity
  - 8: **Output:** Coefficients of  $g$  as explanations for  $f(x)$ 's prediction
-

# XAI Methods: Anchors

---

## What are Anchors?

- Highly precise "IF-THEN" rules explaining model decisions
- Focus on minimum conditions needed to maintain prediction
- Trade precision for coverage (fewer cases explained)
- Easily understood by non-technical stakeholders

---

### Algorithm Anchors for Credit Card Detection

---

- 1: **Input:** Trained model  $f$ , transaction  $x$
  - 2: Initialize anchors  $A = \emptyset$
  - 3: **while** precision of  $A$  (fraction of perturbed samples where  $f$  predicts same as  $f(x)$ ) < threshold **do**
  - 4:     For each candidate feature not in  $A$ :
  - 5:         Add feature to  $A$  and estimate new precision
  - 6:         Add feature that increases precision the most to  $A$
  - 7: **end while**
  - 8: **Output:** Anchors  $A$ : set of feature-value rules that "guarantee"  $f(x)$ 's prediction with high precision
-

# XAI Evaluation: Faithfulness Metric

---

Evaluate an explanation  
with Faithfulness  
metric:

- Measures the correlation between feature importance and changes in the model's predictions when features are altered
- Feature importance correlation analysis with sequential feature removal

---

## Algorithm Faithfulness Metric for XAI Explanations

---

- 1: **Input:** Model  $f$ , sample  $x$ , explanation scores  $E$  for features
  - 2: **for** each feature  $i$  in  $x$  **do**
  - 3:     Remove or mask feature  $i$  in  $x$  to get  $x_{-i}$
  - 4:     Compute prediction difference:  $\Delta_i = |f(x) - f(x_{-i})|$
  - 5: **end for**
  - 6: Compute correlation between  $\Delta_i$  and  $E_i$  across all features
  - 7: **Output:** Faithfulness score (e.g., correlation coefficient)
-

# XAI Evaluation: Monotonicity Metric

---

## Evaluate an explanation with Monotonicity metric:

- Evaluates whether removing features causes consistent changes in the model's predictions
- Sequential feature removal testing with prediction tracking

---

### Algorithm Monotonicity Metric for XAI Explanations

---

```
1: Input: Model  $f$ , sample  $x$ , explanation scores  $E$  for features
2: for each feature  $i$  in  $x$  do
3:   Change feature  $i$ 's value to get  $x_{+i}$ 
4:   Compute prediction change:  $\Delta_i = f(x_{+i}) - f(x)$ 
5:   if  $E_i > 0$  then
6:     Check if  $\Delta_i > 0$  (prediction increases)
7:   else if  $E_i < 0$  then
8:     Check if  $\Delta_i < 0$  (prediction decreases)
9:   end if
10: end for
11: Calculate fraction of features where explanation and prediction
    change agree
12: Output: Monotonicity score (agreement ratio)
```

---

# XAI Evaluation: Completeness Metric

---

Evaluate an explanation with Completeness metric:

- Assesses how much of the model's prediction is captured by the explanation
- Coverage measurement comparing explained variance to total variance

---

## Algorithm Completeness Metric for XAI Explanations

---

- 1: **Input:** Model  $f$ , sample  $x$ , explanation scores  $E$  for features
  - 2: Compute model prediction:  $y = f(x)$
  - 3: Compute baseline prediction:  $y_{base} = f(\text{baseline input})$
  - 4: Sum explanation scores:  $S = \sum_i E_i$
  - 5: Compute completeness error:  $|S - (y - y_{base})|$
  - 6: **Output:** Completeness score (lower error means higher completeness)
-

# Balancing with Stratified Sampling

---

- **Class Imbalance Challenge:**
  - Only 0.58% of transactions are fraudulent in training data
  - Risk of model bias towards predicting legitimate transactions
- **Confidence-based Stratified Sampling:** For comprehensive XAI evaluation

## Confidence Bins for Stratified Evaluation

- **Very Low** (0.0-0.2) — Strong contradiction to model classification
- **Low** (0.2-0.4) — Weak patterns contradicting classification
- **Borderline** (0.4-0.6) — Ambiguous cases with mixed signals
- **High** (0.6-0.8) — Strong but not conclusive patterns
- **Very High** (0.8-1.0) — Clear fraud/non-fraud patterns

# Balancing with Stratified Sampling

---

- **Class Imbalance Challenge:**
  - Only 0.58% of transactions are fraudulent in training data
  - Risk of model bias towards predicting legitimate transactions
- **Confidence-based Stratified Sampling:** For comprehensive XAI evaluation

## Key Benefits

- Ensures comprehensive XAI evaluation across confidence levels
- Enables fair comparison between SHAP, LIME and Anchors

# Edge Case Handling

---

- **Edge Case Identification:**

- Z-score calculation to identify feature value deviations
- Selection of transactions with extreme feature values
- Analysis of unusual transaction patterns

## Selection Process

- **Calculate z-scores:**  $z = \frac{|x-\mu|}{\sigma}$  for important features
- **Maximum deviation:** Find samples with largest z-scores across features
- **Combine with stratified samples:** Ensures both typical and extreme cases
- **Special handling:** Edge cases receive additional manual review

# System Architecture

- 1. Data Collection & Model Design:**  
Obtain synthetic transaction data and design fraud detection models.
- 2. Data Preprocessing:** Raw transaction data is cleaned, normalised, and balanced using the SMOTE method.
- 3. Model Training:** Separate pipelines are implemented for training models.
- 4. XAI Integration:** Use XAI methods to generate explanations for model predictions.
- 5. Performance Evaluation:** XAI evaluation metrics (**Faithfulness**, **Monotonicity**, **Completeness**).

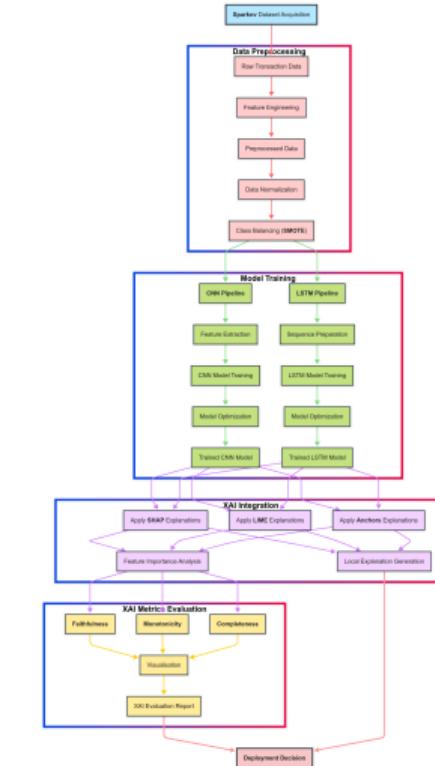


Figure: System architecture illustrative diagram

## **Results**

---

# Model Performance: CNN vs LSTM

---

- I compared two deep learning architectures: a Convolutional Neural Network (CNN) and a Long Short-Term Memory (LSTM) model with attention.
- Both models were trained and evaluated on the Sparkov synthetic dataset, which is highly imbalanced and mimics real-world credit card transaction patterns.

Table: Performance Metrics for CNN and LSTM Models

Model	Accuracy	ROC AUC	Precision (Fraud)	Recall (Fraud)
CNN	98.66%	0.994	21.34%	91.84%
LSTM	97.58%	0.971	11.80%	81.49%

## CNN Model

- True Negatives: 546,311
- False Positives: 7,263
- False Negatives: 175
- True Positives: 1,970

## LSTM Model

- True Negatives: 540,512
- False Positives: 13,062
- False Negatives: 397
- True Positives: 1,748

# Feature Patterns using Statistical Analysis



Figure: Features Distributions in the Sparkov Train Dataset

Explainable AI in Deep Learning for Fraud Detection

# Feature Patterns using Statistical Analysis

---

## Key Statistical Patterns for Fraud Detection (without XAI methods)

- **Amount Anomalies:** Transactions exceeding 500 showed 3.7x higher fraud probability
- **Temporal Patterns:** 23:00-04:00 transactions had 2.9x increased fraud risk
- **Geographical Anomalies:** Transactions  $>75\text{km}$  from cardholder location showed strong fraud indicators
- **Demographic Patterns:** Highest fraud concentration in 50-60 age group (1,443 cases)

# SHAP Summary Plot and Global Importance

---

- **Feature Impact Analysis:**

- Summary plot shows feature importance across all test samples
- Color represents feature value (red = high, blue = low)
- Position shows impact on prediction (right = toward fraud)

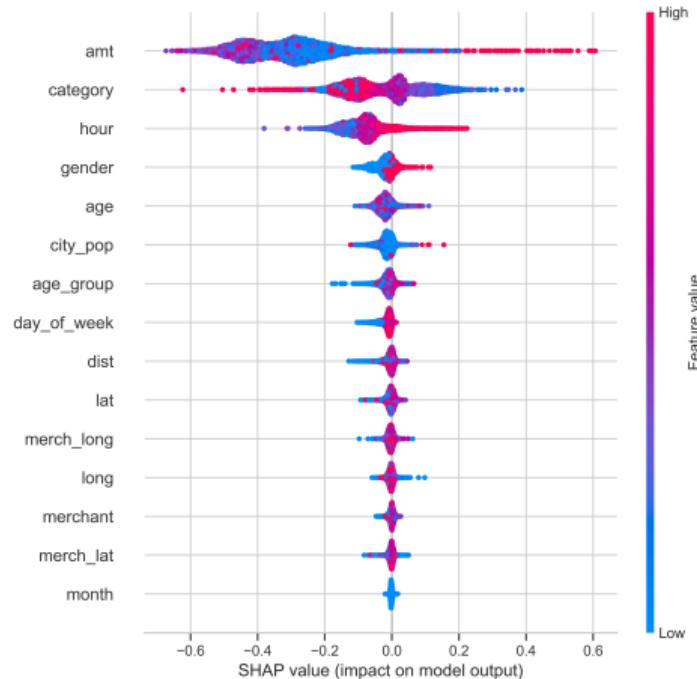


Figure: SHAP Summary Plot: Feature Importance for CNN Model

# Feature Importance: What Drives Fraud Predictions?

- Using SHAP, I identified the most influential features for the CNN model.
- Top features:**
  - Transaction Amount:** Higher values are a strong fraud indicator.
  - Merchant Category**
  - Hour of Transaction:** Transactions between 23:00 and 04:00 are riskier.
  - Gender:** and **Age** Moderate influence.
  - Distance to Merchant:** Unusual distances often signal fraud.
  - Latitude/Longitude:** Minimal impact.

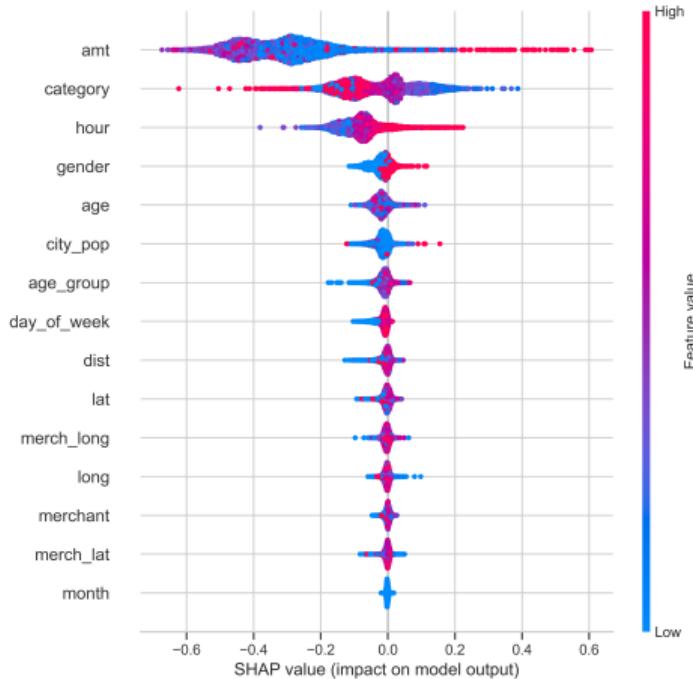


Figure: SHAP Summary Plot: Feature Importance for CNN Model

# SHAP Results Local Visualization

## SHAP Waterfall Plot

- Shows step-by-step impact
- Starting from base value (0.511)
- Category, hour, and age\_group decrease fraud probability
- Final prediction: 0.989 (98.9% fraud confidence)

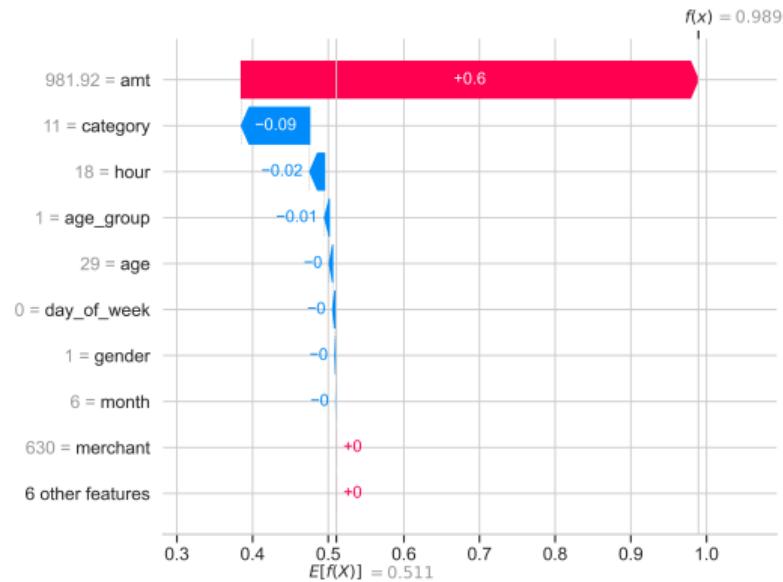


Figure: SHAP Waterfall Plot for the LSTM model's prediction (98.92%) at sample index 1044

# SHAP Results Local Visualization

## SHAP Force Plot

- Shows feature contributions pushing prediction from the base value
- Red = pushing toward fraud
- Blue = pushing toward legitimate
- Example: Transaction amount (+0.6) strongly indicates fraud



Figure: SHAP Force Plot for the LSTM model's prediction (98.92%) at sample index 1044

# LIME Explanation

## Example Interpretation

- Green bars support fraud signals
- Red bars indicate legitimate prediction
- Width shows contribution magnitude
- **Result:** 99.48% confidence in fraud

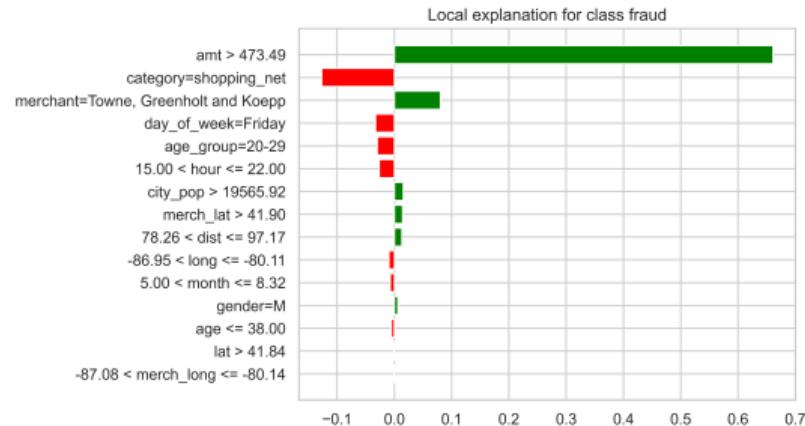


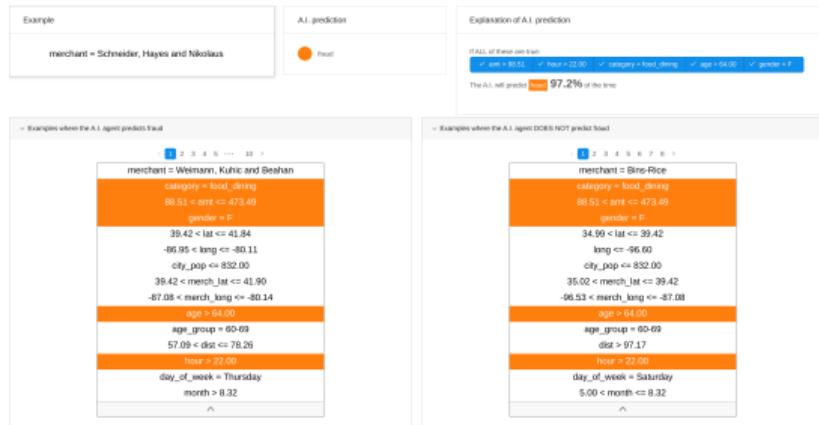
Figure: LIME explanation showing feature contributions for the CNN model's prediction (99.48%) at sample index 1044

# Anchors Rule-Based Explanations

## Rule-Based Explanations:

1. amt > 88.51
2. hour > 22.00
3. category = food\_dining
4. age > 64.00
5. gender = F

**Precision:** 97.2% of cases matching this rule are correctly predicted as fraud  
**Coverage:** 8.4% of all fraud cases are covered by this rule



**Figure:** Anchors Explanation Interactive Observation in Notebook for the LSTM model's prediction (79.83%) at sample index 2025

# XAI Effectiveness Across Confidence Levels

- The effectiveness of the XAI method depends on the model's prediction confidence.
- Very High Confidence ( $>0.8$ ):** SHAP explanations are most reliable.
- High Confidence (0.6–0.8):** SHAP and LIME together provide a balanced view.
- Borderline (0.4–0.6):** A multi-method approach (SHAP, LIME, Anchors) is best.
- Low Confidence ( $<0.4$ ):** SHAP plus human review is recommended.

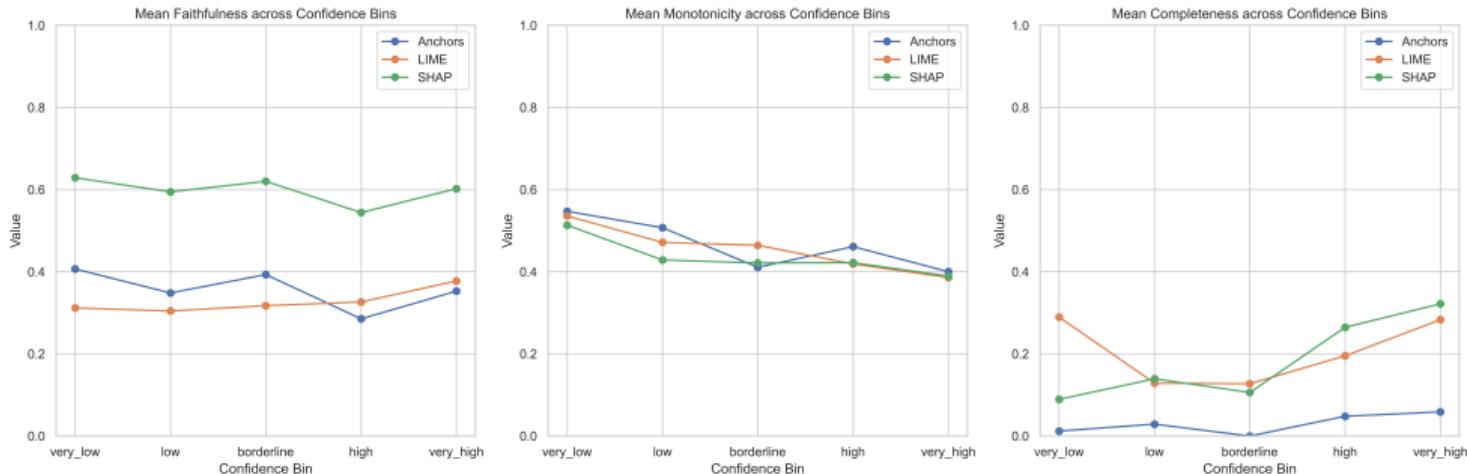


Figure: XAI Methods Performance Across Confidence Bins

# XAI Methods: Performance Comparison Between Models

---

## Key Findings:

- SHAP achieved the highest faithfulness for both models:
  - **LSTM**: 0.761 (significantly higher)
  - **CNN**: 0.443
- LSTM model consistently showed better faithfulness scores across all XAI methods
- Monotonicity and Completeness showed less variation between models

Model & Method	Faith.	Mono.	Comp.
<b>CNN + SHAP</b>	0.443	0.464	0.208
CNN + LIME	0.254	0.488	<b>0.295</b>
CNN + Anchors	0.315	0.486	0.033
<b>LSTM + SHAP</b>	<b>0.761</b>	0.429	0.134
LSTM + LIME	0.396	0.445	0.139
LSTM + Anchors	0.412	<b>0.469</b>	0.022

Table: XAI Methods Performance Metrics

# Summary of Key Findings

---

- **SHAP is the most faithful XAI method**, especially for high-confidence predictions, providing explanations that align closely with the model's actual decision process.
- **LIME** offers balanced completeness and is particularly useful for local, case-by-case explanations.
- **Anchors** delivers the most human-interpretable rules, though with limited coverage.
- **Key global fraud indicators** across all models: transaction amount, category, and transaction hour.
- **XAI method effectiveness** varies with prediction confidence, SHAP is robust across all levels, while LIME achieves the best completeness. Anchors can give an easy-to-understand explanation.

## Takeaway

Integrating XAI with deep learning models not only boosts trust and transparency but also provides actionable insights for fraud analysts and operational teams.

## **Discussion**

---

# Practical Implications and Challenges

---

- **Real-world deployment:** Both models process single transactions in under 5 seconds, and generate explanations in less than 3 minutes, making them suitable for real-time fraud detection.
- **Computational overhead:** SHAP explanations are computationally intensive (up to 45 minutes for 100 explanations), requiring careful resource management.
- **Stakeholder needs:**
  - SHAP is best for data scientists needing technical depth.
  - LIME is ideal for fraud analysts seeking intuitive, local explanations.
  - Anchors are valuable for operational teams needing clear, actionable rules.
- **Access to real data:** As changing fraud patterns in the real world would be a problem

## **Conclusions**

---

# Conclusions

---

- Successfully integrated XAI methods (SHAP, LIME, Anchors) with deep learning models for credit card fraud detection.
- Developed a novel confidence-based evaluation framework for XAI effectiveness.
- Selecting XAI methods based on the prediction confidence and end user needs.

## Final Thought

XAI is not just a technical add-on, it is essential for building trust, meeting regulatory requirements, and empowering analysts in the fight against fraud.

# Future Work

---

- **Expand model diversity:** Explore more advanced Deep Learning models and their explainability for fraud detection.
- **User studies:** Conduct usability studies with fraud analysts to assess the practical value of XAI explanations.
- **Optimise XAI computation:** Investigate faster, scalable XAI methods for real-time deployment.
- **More advanced DL-based explainers:** DeepLIFT<sup>8</sup> (Deep Learning Important FeaTures), X-NeSyL<sup>9</sup> (eXplainable Neural Symbolic Learning).

---

<sup>8</sup>Shrikumar et al., 2017.

<sup>9</sup>Díaz-Rodríguez et al., 2022.

# Acknowledgements

---

- I would like to thank my supervisor, Dr. Hyunkook Lee, for his guidance, expertise, and encouragement throughout this project.
- A grateful attitude to the pioneering researchers behind the previous brilliant works on XAI concepts and Deep Learning architecture designs, Sparkov synthetic data in credit card fraud detection, which made this project possible.

# Poster Presentation



## Explainable AI (XAI) in Deep Learning Models for Credit Card Fraud Detection

Thong Minh Lai<sup>1</sup> Supervisor: Dr. Hyunkook Lee<sup>2</sup>

<sup>1</sup>University of Huddersfield, Department of Computer Science

### Introduction & Motivation

Credit card fraud remains a persistent crime issue, with it losses reaching £1.3 billion in 2021 (UK Finance, 2022). While deep learning models have shown promise for fraud detection, their "black box" nature makes them difficult to trust and deploy in real-world financial systems. This project aims to bridge that gap by integrating Explainable AI (XAI) techniques, making model decisions transparent and understandable for analysts and stakeholders.

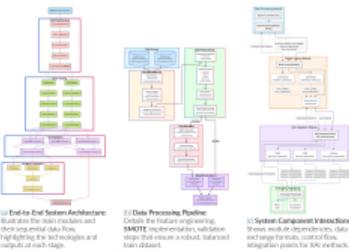
#### Project Objectives

- Develop and compare models explaining (CNN, LSTM with attention) for fraud detection.
- Integrate XAI methods (SHAP, LIME, Anchors) for local interpretability.
- Evaluate the explainability using XAI metrics (Fairfulness, Monotonicity, Completeness).
- Provide insights for financial institutions implementing transparent AI.

#### Why This Matters

Understanding fraud detection models through explainable AI is essential in today's financial landscape. Financial institutions face increasing regulatory pressure for transparency in automated decision-making. By integrating XAI into our models, we can ensure that they are not only accurate but also transparent and accountable. This research will demonstrate how explainable AI can be used to detect sophisticated fraud techniques (Goh et al., 2023). Explainable Fraud Detection creates a crucial hedge between complex AI systems and human oversight, enabling compliance officers to validate regulatory adherence and fraud analysts to verify and refine AI-flagged transactions with their domain expertise. This research will also highlight the importance of explainability in AI, showing that fraud decisions are based on identifiable, reasonable patterns that can be communicated to customers, auditors, and management. As financial fraud becomes more sophisticated, the human AI partnership must remain vigilant, leveraging the unique capabilities of neural networks with human judgment and regulatory compliance requirements.

### System Architecture: End-to-End Workflow and Modular Design



1. **Data Collection & Model Design:** Gather synthetic transaction data (labels, etc.) and design model architectures (CNN, LSTM). Data is loaded and preprocessed.
2. **Data Preprocessing:** Applies feature engineering (temporal, spatial, demographic), handles class imbalance using SMOTE (Bouyer et al., 2011) and standardizes/encodes features to ensure model interpretability.
3. **Model Training (CNN & LSTM):** Train and optimize for fraud detection. Each model is tuned and validated independently.
4. **XAI Integration:** Implements SHAP, LIME, and Anchors for local interpretability, providing visualizations and explanations.
5. **Explaining Performance Evaluation:** Comprehensive XAI evaluation metrics (Fairness, Monotonicity, Completeness), with visual dashboards for comparative analysis.

smminh@gmail.com | 1234567890@hudd.ac.uk

### Predictive Model Architectures: CNN and LSTM



Figure 1. CNN Model Layers.



Figure 2. LSTM Model Layers with Attention Mechanism.

CNN:  
• Conv1D layers (4A, 32 filters)  
• Batch Normalization, Dropout, Dense layers  
• 59,937 trainable parameters

LSTM with Attention:  
• Two LSTM layers (50 units each)  
• Attention mechanism, Dense output layer  
• 33,502 trainable parameters

### Explanations from XAI Methods: SHAP, LIME, and Anchors



Figure 3. SHAP Waterfall Plot: Contributions for a CNN prediction.

Figure 4. SHAP Force Plot: Cumulative impact from base value.



Figure 5. LIME Explanation Chart: For visualization of fraud impacts.



Figure 6. Anchors visualizations: present clear IF-THEN rules that specify the exact conditions leading to a specific prediction. Each rule lists thresholds for relevant features (e.g., amount = 84.51, year = 22.00, gender = male, age = 30, ...), and its weight (e.g., 0.001), making it straightforward to understand the criteria that trigger a model prediction. The absolute precision percentage indicates the reliability of the rule for similar tasks.

### XAI Evaluation Metrics and Insights

Metric	CNN	LSTM
Fairfulness	0.809 (1.00% Deviation)	0.807 (0.00% Deviation)
Monotonicity	0.529 (0.00%)	0.499 (0.364)
Completeness	0.620 (0.00%)	0.208 (0.255)

Note: Higher values indicate better performance across all metrics.

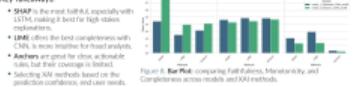


Figure 7. Bar Plot: comparing Fairfulness, Monotonicity and Completeness across models and XAI methods.

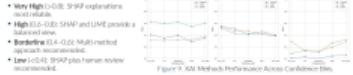


Figure 8. Line Plot: XAI Effectiveness by Confidence Level.



Figure 9. Line Plot: XAI Methods Performance Across Confidence Bits.

(a) Fairfulness Comparison (b) Monotonicity Comparison (c) Completeness Comparison

Figure 10. Radar Charts: Comparison of All Average Metrics Under XAI Methods and Models.

### Practical Implications

- Real-time Detection:** A transaction's explanations can be predicted in under 3 seconds, making it suitable for real-time deployment.
- Stakeholder Satisfaction:**
  - Data Scientists: Technical insights and visualizations help them interpret model decisions.
  - Business Analysts: Clear visualizations support legitimate transactions.
- Computational Overhead:** SHAP is highly efficient, allowing for quick identification of the most influential factors.

### Conclusions

This project successfully integrated and evaluated the XAI methods SHAP, LIME, and Anchors with state-of-the-art deep learning models (CNN and LSTM) for credit card fraud detection. The approach addressed the "black box" challenge of deep learning, making model decisions more transparent and attributable for a range of stakeholders. This work demonstrated that XAI can significantly enhance the trustworthiness of AI systems, particularly in high-stakes applications where selective trust on one model is often required. Future directions include optimizing XAI computation for real-time use, expanding to more diverse datasets, and combining XAI methods with AI audits to further refine the system.

**References**

Bouyer, S., et al. (2011). Synthetic data for imbalanced classification: An experimental study. *Machine Learning*, 85(1), 129–152.

Goh, J., et al. (2023). Explainable Fraud Detection: A Survey. *IEEE Transactions on Dependable and Secure Computing*, 19(1), 1–16.

UK Finance. (2022). *Annual Report and Accounts 2021*. London, UK: UK Finance.

Wang, Y., et al. (2020). Explainable AI for Credit Card Fraud Detection: A Survey. *Journal of Computer Information Systems*, 60(4), 1020–1038.

Wang, Y., et al. (2021). Explainable AI for Credit Card Fraud Detection: A Survey. *Journal of Computer Information Systems*, 61(1), 1020–1038.

Wang, Y., et al. (2022). Explainable AI for Credit Card Fraud Detection: A Survey. *Journal of Computer Information Systems*, 62(1), 1020–1038.

# References

---

- Bowyer, K. W., Chawla, N. V., Hall, L. O., & Kegelmeyer, W. P. (2011).SMOTE: synthetic minority over-sampling technique. *CoRR*, *abs/1106.1813*.  
<http://arxiv.org/abs/1106.1813>
- Díaz-Rodríguez, N., Lamas, A., Sanchez, J., Franchi, G., Donadello, I., Tabik, S., Filliat, D., Cruz, P., Montes, R., & Herrera, F. (2022).Explainable neural-symbolic learning (x-nesyl) methodology to fuse deep learning representations with expert knowledge graphs: The monumai cultural heritage use case. *Information Fusion*, *79*, 58–83. [https://doi.org/https://doi.org/10.1016/j.inffus.2021.09.022](https://doi.org/10.1016/j.inffus.2021.09.022)
- Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018).Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 80–89. <https://doi.org/10.1109/DSAA.2018.00018>
- Grover, P., Xu, J., Tittelfitz, J., Cheng, A., Li, Z., Zablocki, J., Liu, J., & Zhou, H. (2023). Fraud dataset benchmark and applications.  
<https://arxiv.org/abs/2208.14417>
- Harris, B. (n.d.). *Generate fake credit card transaction data, including fraudulent transactions*. GitHub. Retrieved April 23, 2025, from  
[https://github.com/namebrandon/Sparkov\\_Data\\_Generation](https://github.com/namebrandon/Sparkov_Data_Generation)
- Ibtissam, B., Samira, D., Bouabid, E. O., & Jaafar, J. (2021).Enhanced credit card fraud detection based on attention mechanism and lstm deep model. *Journal of Big Data*, *8*(151). <https://doi.org/10.1186/s40537-021-00541-8>
- Shrikumar, A., Greenside, P., & Kundaje, A. (2017, August). Learning important features through propagating activation differences. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (pp. 3145–3153, Vol. 70). PMLR.  
<https://proceedings.mlr.press/v70/shrikumar17a.html>
- Sundararamaiah, M., Nagarajan, S. K. S., Mudunuru, K. R., & Remala, R. (2024).Unifying ai and rule-based models for financial fraud detection. *International Journal of Computer Trends and Technology*, *72*, 61–68. <https://doi.org/10.14445/22312803/IJCTT-V72I12P107>
- UK Finance. (2022). *Annual fraud report 2022*. UK Finance. Retrieved April 23, 2025, from  
<https://www.ukfinance.org.uk/policy-and-guidance/reports-and-publications/annual-fraud-report-2022>

# Thank you!

U2259343@unimail.hud.ac.uk

Github: ThongLai/Credit-Card-Transaction-Fraud-Detection-Using-Explainable-AI

## Invitation

*I welcome any questions, feedback, or collaboration ideas for making fraud detection smarter and more transparent!*