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

Thong Minh Lai [1] Supervisor: Dr. Hyunkook Lee [2]

[1]University of Huddersfield, Department of Computer Science

Introduction & Motivation

Credit card fraud remains a persistent and costly issue, with UK 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)** methods, making model decisions transparent, actionable for analysts and stakeholders.

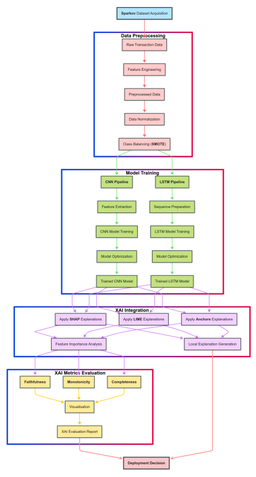
Project Objectives:

- Develop and compare models explainability (**CNN**, **LSTM with attention**) for fraud detection.
- Integrate XAI methods (**SHAP**, **LIME**, **Anchors**) for local interpretability.
- Evaluate the explainability using XAI metrics (**Faithfulness**, **Monotonicity**, **Completeness**).
- Insights for financial institutions to implement transparent AI.

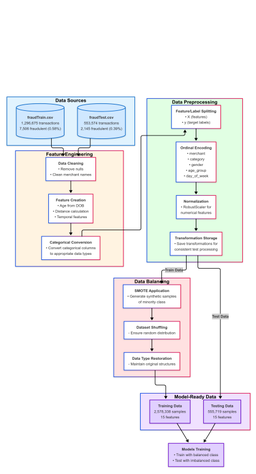
Why This Matters

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 (Anang et al., 2024), while simultaneously needing to improve model performance to combat sophisticated fraud techniques (Galla et al., 2023). Explainable fraud detection creates a crucial bridge 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 (Greenwood & Van Buren III, 2010). This transparency builds stakeholder trust by demonstrating that fraud decisions are based on identifiable, reasonable patterns that can be communicated to customers, auditors, and management. As financial fraud becomes more sophisticated, this human-AI partnership represents the most effective defence, combining the pattern-recognition capabilities of neural networks with human judgment and regulatory compliance requirements.

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



(a) **End-to-End System Architecture:** Illustrates the main modules and their sequential data flow, highlighting the technologies and outputs at each stage.



(b) **Data Processing Pipeline:** Details the feature engineering, **SMOTE** implementation, validation steps that ensure a robust, balanced train dataset.

(c) **System Component Interactions:** Shows module dependencies, data exchange formats, control flow, and integration points for XAI methods.

Figure 1. **System Architecture and Workflow.**

1. **Data Collection & Model Design:** Gathers **Sparkov** synthetic transaction data (Harris, n.d.), designs model architectures (**CNN**, **LSTM**). Data is loaded and versioned for reproducibility.
2. **Data Preprocessing:** Applies feature engineering (temporal, spatial, demographic), handles class imbalance using **SMOTE** (Bowyer et al., 2011), and standardises/encodes features to ensure high-quality model input.
3. **Model Training (CNN & LSTM):** Trains and optimised for fraud detection. Each model is tuned and validated independently.
4. **XAI Integration:** Implements SHAP, LIME, and Anchors for local interpretability, providing instance-level explanations for model predictions.
5. **Explainability Performance Evaluation:** Comprehensive XAI evaluation metrics (**Faithfulness**, **Monotonicity**, **Completeness**), with visual dashboards for comparative analysis.

Predictive Model Architectures: CNN and LSTM

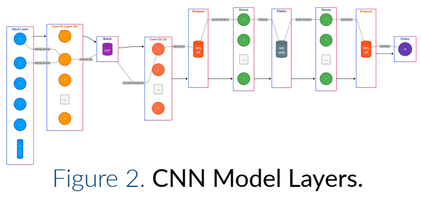


Figure 2. **CNN Model Layers.**

- **Conv1D** layers (64, 32 filters)
- Batch Normalisation, Dropout, Dense layers
- **59,937** trainable parameters

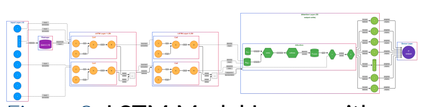


Figure 3. **LSTM Model Layers with Attention Mechanism.**

- Two **LSTM** layers (50 units)
- **Attention mechanism**, Dense output layer
- **33,502** trainable parameters

XAI Methods Explanations: SHAP, LIME, Anchors

SHAP (SHapley Additive exPlanations) visualisations, including the waterfall and force plots, illustrate how individual features influence the model’s prediction. In the waterfall plot, each bar represents the contribution of a feature, starting from a base value (e.g., 0.511). Red bars indicate features that increase the fraud probability, while blue bars show features that decrease it.

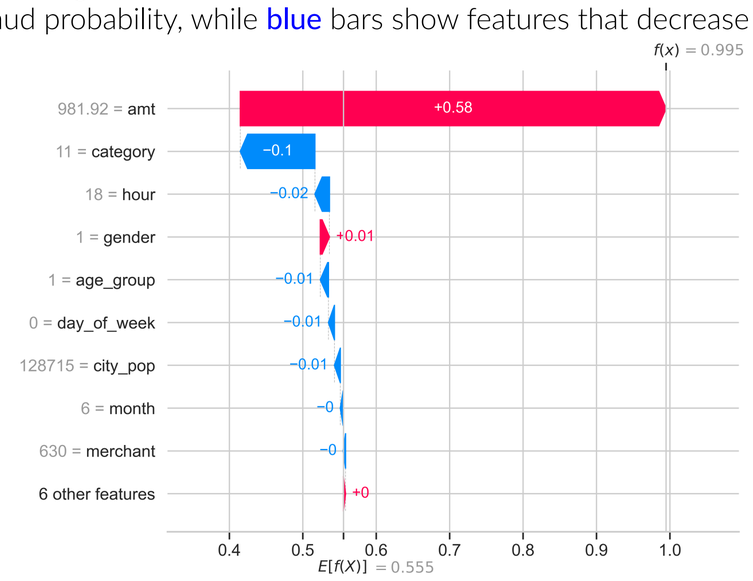


Figure 4. **SHAP Waterfall Plot:** Contributions for a CNN prediction.

The force plot provides a complementary view, displaying the cumulative effect of all features, with the width of each bar indicating the magnitude of its impact. Features are arranged by their influence, making it easy to identify which factors most strongly push the prediction towards fraud or legitimacy.

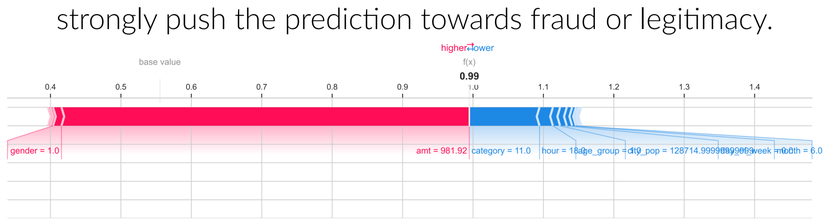


Figure 5. **SHAP Force Plot:** Cumulative impact from base value.

LIME (Local Interpretable Model-agnostic Explanations) visualisations use a bar chart to display feature importance and directionality for a specific prediction. Green bars represent features suggesting fraud, while red bars indicate features supporting legitimate transactions. The length of each bar corresponds to the magnitude of the feature’s impact, and features are ordered by their absolute importance, allowing for quick identification of the most influential factors.

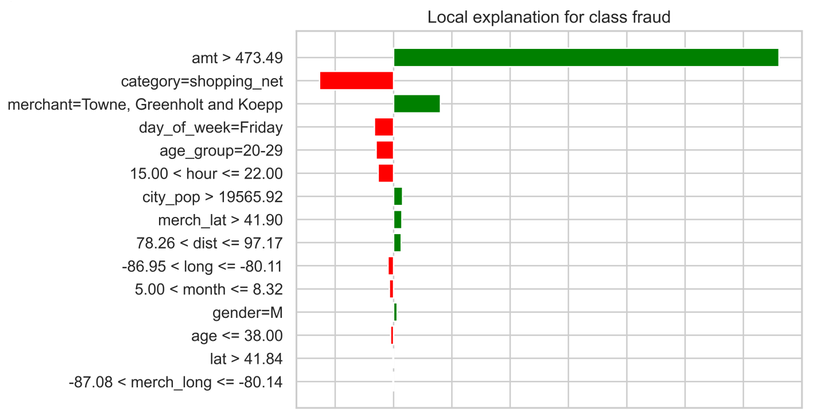


Figure 6. **LIME Explanation Chart:** Bar visualisation of feature impact.

Anchors visualisations present clear **IF-THEN** rules that specify the exact conditions leading to a prediction. Each rule lists threshold values for relevant features (e.g., *amount > 88.51*, *hour > 22.00*, *category = food_dining*, *age > 64.00*, *gender = F*), making it straightforward to understand the criteria that trigger a fraud prediction. The associated precision percentage indicates the reliability of the rule for similar cases.

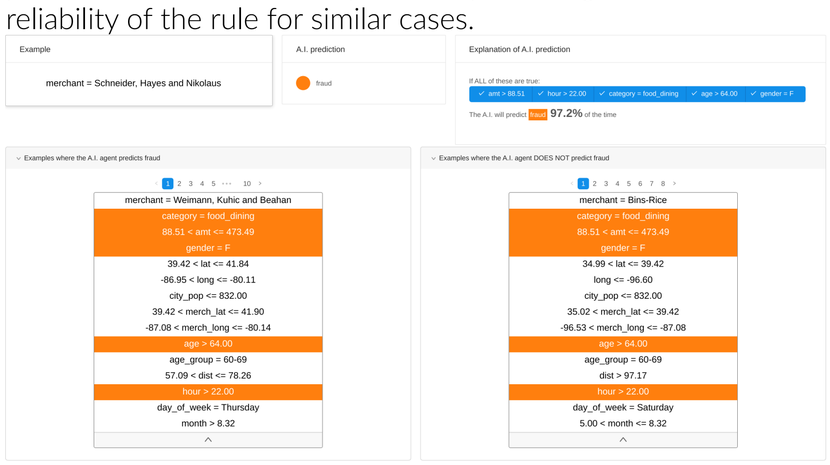


Figure 7. **Anchors Rule-based:** Conditions for a prediction.

XAI Evaluation Metrics and Insights

XAI Metrics:

- **Faithfulness:** How well explanations reflect model logic (higher scores indicate better alignment with model behaviour)
- **Monotonicity:** Consistency of feature importance across different inputs
- **Completeness:** Coverage of the model’s decision process by the explanation

Table 1. Evaluation Metrics Comparison for XAI Methods on CNN, LSTM Models. Higher values are better for all metrics.

Metric	LSTM			CNN		
	SHAP	LIME	Anchors	SHAP	LIME	Anchors
Faithfulness	**0.761**	0.396	0.412	0.443	0.254	0.315
Monotonicity	0.429	0.445	**0.469**	0.464	**0.488**	0.486
Completeness	0.134	**0.139**	0.022	0.208	**0.295**	0.033

Key Takeaways:

- **SHAP** is the most faithful, especially with LSTM, making it best for high-stakes explanations.
- **LIME** offers the best completeness with CNN, and is more intuitive for fraud analysts.
- **Anchors** are great for clear, actionable rules, but their coverage is limited.
- Selecting XAI methods based on the prediction confidence and end user needs.

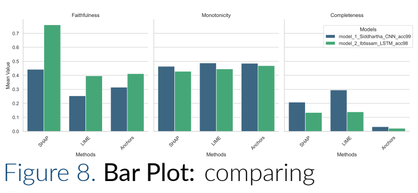


Figure 8. **Bar Plot:** comparing Faithfulness, Monotonicity, and Completeness across models and XAI methods.

XAI Effectiveness by Confidence Level:

- **Very High** (>0.8): SHAP explanations most reliable.
- **High** (0.6–0.8): SHAP and LIME provide a balanced view.
- **Borderline** (0.4–0.6): Multi-method approach recommended.
- **Low** (<0.4): SHAP plus human review recommended.

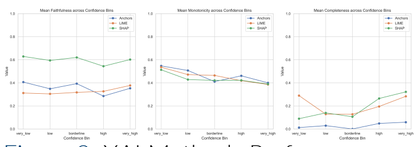
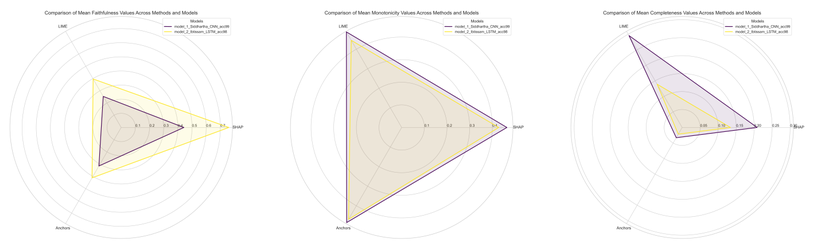


Figure 9. XAI Methods Performance Across Confidence Bins.



(a) **Faithfulness Comparison** (b) **Monotonicity Comparison** (c) **Completeness Comparison**

Figure 10. **Radar Charts:** Comparison of All Average Metrics Values Across Methods and Models

Practical Implications Limitations & Future Work

- **Real-time Detection:** A transaction’s explanations can be produced in under 3 minutes, suitable for operational deployment.
- **Stakeholder Suitability:**
  - **SHAP:** Data scientists (technical depth)
  - **LIME:** Fraud analysts (intuitive, visual)
  - **Anchors:** Operations teams (clear rules)
- **Computational Overhead:** **SHAP** is resource intensive, requiring careful resource management optimisation.
- **Data:** Reliance on synthetic data (**Sparkov**) may limit real-world generalisability.
- **User Experience:** XAI outputs can be complex for non-technical users.
- **Next Steps:**
  - Optimise XAI computation for real-time use.
  - Further research on other advanced explainers (**DeepLIFT**, **X-NeSyL**, etc,).
  - Conduct usability studies with fraud analysts.
  - Incorporate more diverse, real-world datasets.

Conclusions

This project successfully integrated and evaluated three leading 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 actionable for a range of stakeholders. This work demonstrates a path for deploying explainable AI in real-world fraud detection, with practical recommendations for method selection based on use case and confidence level. Future directions include optimising XAI computation for real-time use, expanding to more advanced explainers, and conducting usability studies with fraud analysts to further refine the system.

References

Anang, A. N., Ajewumi, O., Sonubi, T., Nwafor, K., Arogundade, J., & Akinbi, I. (2024).Explainable ai in financial technologies: Balancing innovation with regulatory compliance. *International Journal of Science and Research Archive*, *13*, 1793–1806. https://doi.org/10.30574/ijsra.2024.13.1.1870

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

Galla, E. P., Gollangi, H. K., Boddapati, V. N., Sarisa, M., Polimetla, K., Rajaram, S. K., & Reddy, M. S. (2023).Enhancing performance of financial fraud detection through machine learning model. *J Contemp Edu Theo Artific Intel: JCETAI-101*. https://doi.org/10.2139/ssrn.4993827

Greenwood, M., & Van Buren III, H. J. (2010).Trust and stakeholder theory: Trustworthiness in the organisation–stakeholder relationship. *Journal of Business Ethics*, *95*, 425–438. https://doi.org/10.1007/s10551-010-0414-4

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

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

laiminhthong1@gmail.com | U2259343@unimail.hud.ac.uk (BSc) Computer Science with Artificial Intelligence https://github.com/ThongLai/Credit-Card-Transaction-Fraud-Detection-Using-Explainable-AI