RADBOUD UNIVERSITY

Faculty of Social Sciences

How can post-hoc explanation methods such as SHAP and DIFFI enhance the interpretability of Isolation Forest models in the context of bank transaction fraud detection?

April 11, 2025

Justin Snelders
S1054709

Contents

Chapter 1

Introduction

The once niche research area of Artificial Intelligence (AI) is rapidly becoming a foundational technology for modern day applications. Significant increases of AI usage have been reported for both commercial and academic purposes. McKinsey & Company's 2024 AI review reported that 78% of respondent companies used AI in at least one business function. This number was 72% in 2023 and 55% in 2022, showing a significant and persistent increase. (Singla et al., 2025) In the scientific domain, a similar trend can be observed. In a 2023 bibliometric analysis, the European Commission reported a significant growth of AI publications. The commission even observed that research on AI applications in science is growing faster than the AI field as a whole. (Arranz et al., 2023)

As society increasingly adopts more AI systems into workflows, the significance of trust in AI from a user perspective becomes more apparent. Trust is the essential component for humans to accept AI technology adoption. (Afroogh et al., 2024) The factors affecting trust in AI systems can be categorized in technical factors and axiological factors. Transparency, explainability and performance of the AI system are amongst the most important technical factors for trustworthy AI. Most AI methods, however, are based on complex opaque models like deep neural networks. For such "black-box" models, even the developers have limited access to the mechanisms of the model that dictate the input processing. This lack of transparency and explainability could greatly damage the trustworthiness of the AI system.

There is an increasing demand for a new generation of explainable AI (XAI) technology to completely understand how AI models make predictions. (Minh et al., 2022) Understanding how a model works significantly increases transparency and explainability, which is crucial for AI adoption in high-stake sectors like healthcare. In such sectors, AI predictions can have significant consequences for end users like patients.

Chapter 2

Research Question

To address the growing need for XAI in high-stake sectors, this paper investigates the use of explainable anomaly detection in the financial sector. A machine learning model is trained to detect bank transaction outliers and discover possibly fraudulent transactions. This detection is done using Isolation Forest (IF), which is an unsupervised machine learning technique for anomaly detection. (Liu et al., 2008) Two post-hoc explanation methods are applied to the IF model. The first method is a model-agnostic explanation method called SHAP. The second method is a model-specific explanation method to IF called DIFFI.

This is captured by the following research question: **How can post-hoc explanation methods such as SHAP and DIFFI enhance the interpretability of Isolation Forest models in the context of bank transaction fraud detection?** This research aims to explore and compare the effectiveness of these two explanation techniques in providing insights into model behaviour and feature importance. Additionally, potentiall pitfalls and limitations are discussed.

Chapter 3

Methods

3.1 Codebase

The full code for this research is available on GitHub, including installation instructions. (Snelders, 2025)

3.2 Data

This research uses the publicly available Kaggle dataset "Bank Transaction Dataset for Fraud Detection", curated by Vala Khorasani. (Khorasani, 2024) The dataset contains 2512 samples of financial transaction data, covering customer demographics and usage patterns. The dataset is intended for data scientists researching or developing predictive models for financial security applications. Therefore, it closely aligns with this research question.

3.3 Data Preprocessing

The data contains numerous attributes that can be beneficial for tracking and analyzing usage patterns of specific users over time. However, this research is mainly focused on detecting potentially fraudulent transactions based on statistically relevant data inherently linked to a single transaction. Therefore, several attributes related to identification, location and time are discarded. Figure 3.2 shows the final representation of the data.

TransactionAmount	TransactionType	Channel	TransactionDuration	LoginAttempts	AccountBalance
14.09	Debit	ATM	81	1	5112.21
376.24	Debit	ATM	141	1	13758.91
126.29	Debit	Online	56	1	1122.35
184.50	Debit	Online	25	1	8569.06
13.45	Credit	Online	198	1	7429.40
...	...	...	...	...	...
856.21	Credit	Branch	109	1	12690.79
251.54	Debit	Branch	177	1	254.75
28.63	Debit	Branch	146	1	3382.91
185.97	Debit	Online	19	1	1776.91
243.08	Credit	Online	93	1	131.25

Figure 3.1: Structure of the financial transaction data.

Before the data is fed to the IF model, the numerical variables are standardized. The categorical variables are one-hot encoded. Both operations are performed using scikit-learn's machine learning library.

3.4 Isolation Forest

To detect anomalies in the transaction data, an Isolation Forest (IF) model is used. IF is an unsupervised machine learning algorithm specifically designed for anomaly detection. (Liu et al., 2008) Unlike traditional classification models, IF solely focuses on detecting anomalies rather than classifying normal data points. The model assumes that anomalies are sparse and different, which makes them easier to isolate from the rest of the data.

IF builds an ensemble of binary trees. A specified number of subsamples is taken from the data. For each subsample, a binary isolation tree is built by selecting a random feature and taking a random threshold to partition the data. This continues until no further splitting is possible, and the instances are isolated. After all instances are isolated, the average path length from each instance to the root is calculated and an anomaly score can be computed. The anomaly score for a data point x is defined as:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}$$

where $E(h(x))$ is the average path length of x across all isolation trees, and $c(n)$ is the average path length of unsuccessful searches in a Binary Search Tree for sample size n. This is often approximated.

3.5 Depth-based Isolation Forest Feature Importance

The first explanation method this research will examine is Depth-based Isolation Forest Feature Importance (DIFFI). DIFFI is a post-hoc method specifically designed to provide feature importance scores for IF models. (Carletti et al., 2023) The method provides global feature importance scores, and additionally has a local version "Local-DIFFI" for local explanations. This research will restrict itself to global DIFFI.

DIFFI takes each isolation tree and classifies in-bag data into predicted inliers and predicted outliers, determined by a threshold on the anomaly score. Next, the Induced Imbalance Coefficient (IIC) is calculated for each internal node in the tree. This measures how unevenly a feature splits the data. The Cumulative Feature Importance scores (CFIs) in the tree are updated and weighted based on how early the feature appears and how unbalanced the split is. This yields I_O, C_O, I_I, C_I which are the importance and count for outliers and inliers respectively. After this is aggregated across all trees, the final Global Feature Importance score (GFI) is calculated using the following formula:

$$\text{GFI} = \frac{I_O/C_O}{I_I/C_I}$$

The following diagram visualizes the DIFFI pipeline:

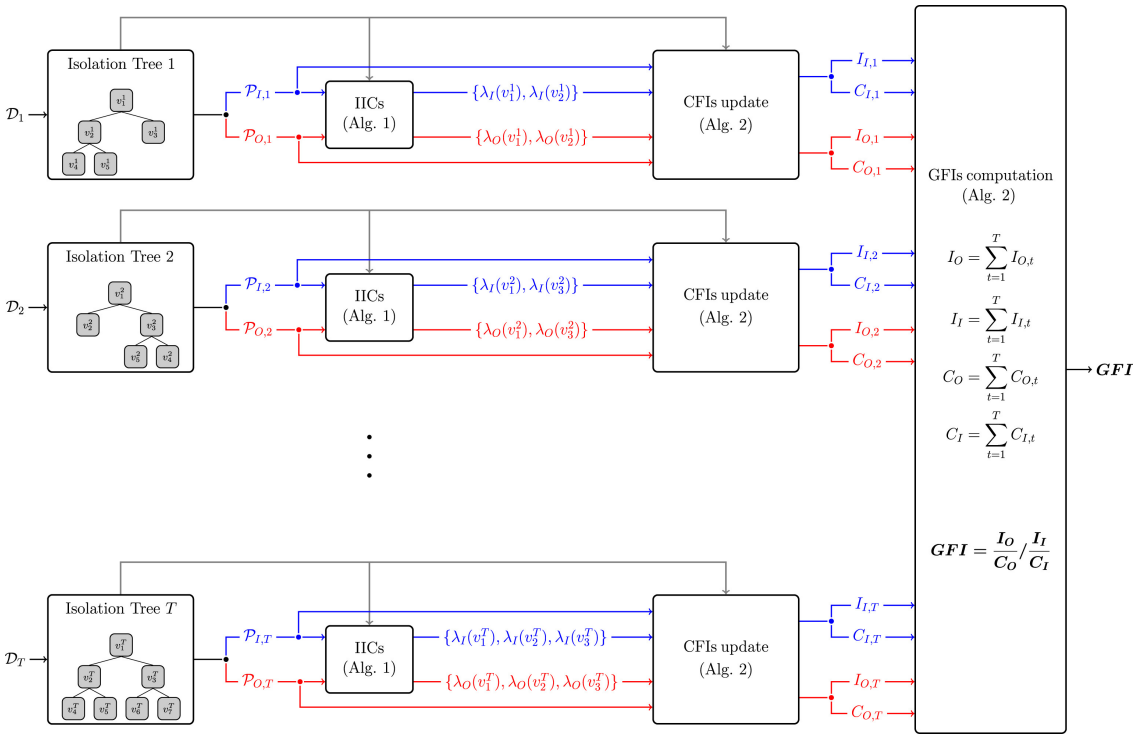

Figure 3.2: Visualization of the DIFFI pipeline. Predicted inliers and outliers are taken from each isolation tree. The IIC is calculated for each internal node, and the CFIs are updated. The resulting importance and counts for the inliers and outliers are used to calculate the GFI. Diagram is retrieved from the paper by Carletti et al.

3.6 SHapley Additive exPlanations

The second explanation model in this research is SHapley Additive exPlanations (SHAP). SHAP is a model-agnostic explanation method providing both global and local explanations for machine learning models. (Lundberg and Lee, 2017) SHAP values are based on Shapley values from game theory, and indicate how much each feature contributes to the final prediction.

Each model is treated as a game, where the features are the players and the prediction is the outcome of cooperation. The SHAP value for a feature is its fair share of the prediction. This indicates how much each feature has contributed to the model prediction. SHAP values are computed as an average of a feature's marginal contributions across all possible feature combinations:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f(S \cup \{i\}) - f(S) \right]$$

where F is the full feature set, S is a subset of features without i and $f(S)$ is the model prediction when only features in S are used.

Chapter 4

Results

4.1 Isolation Forest Anomaly Detection

The IF model detected 126 potential fraudulent transactions. By plotting several features against each other, the results of the anomaly detection can be visualized:

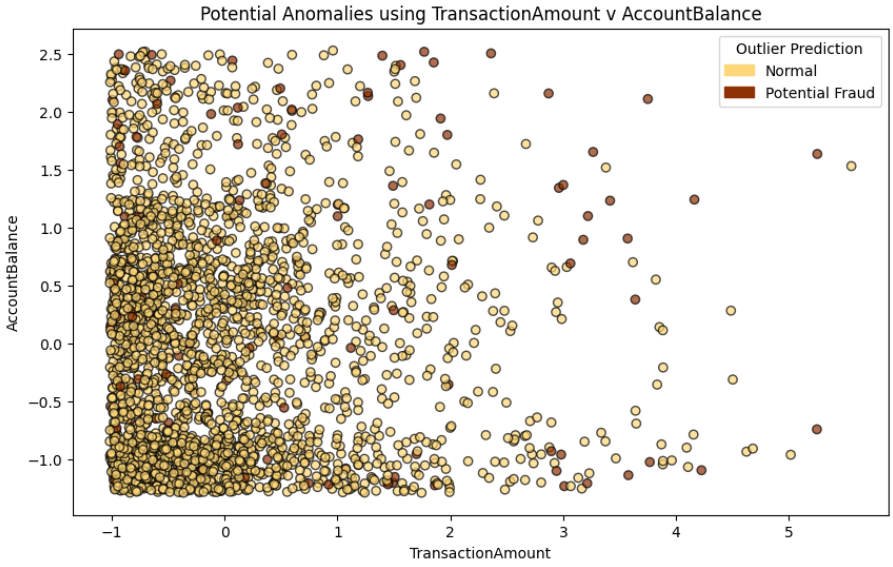

Figure 4.1: Scatterplot of the features TransactionAmount and AccountBalance.

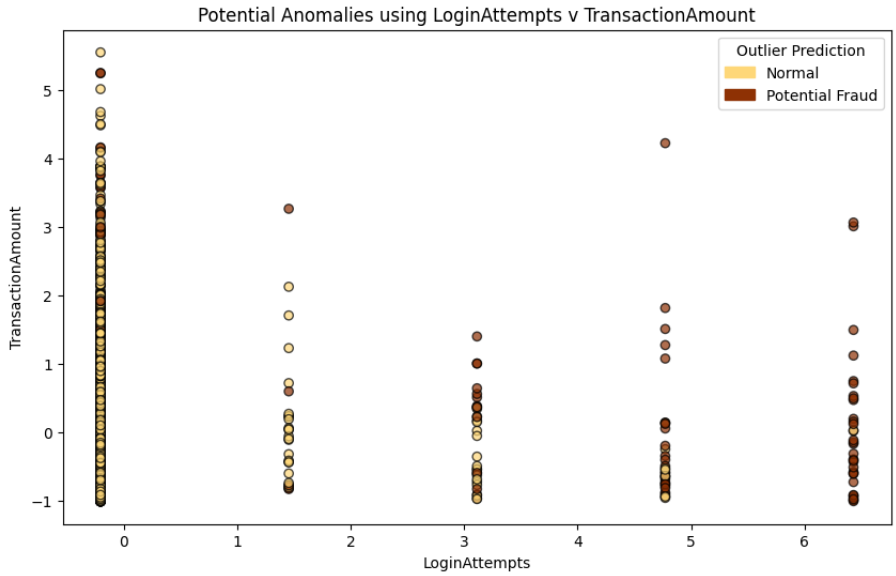

Figure 4.2: Scatterplot of the features LoginAttempts and TransactionAmount.

4.2 DIFFI Scores

The following plot visualizes the DIFFI score for each feature:

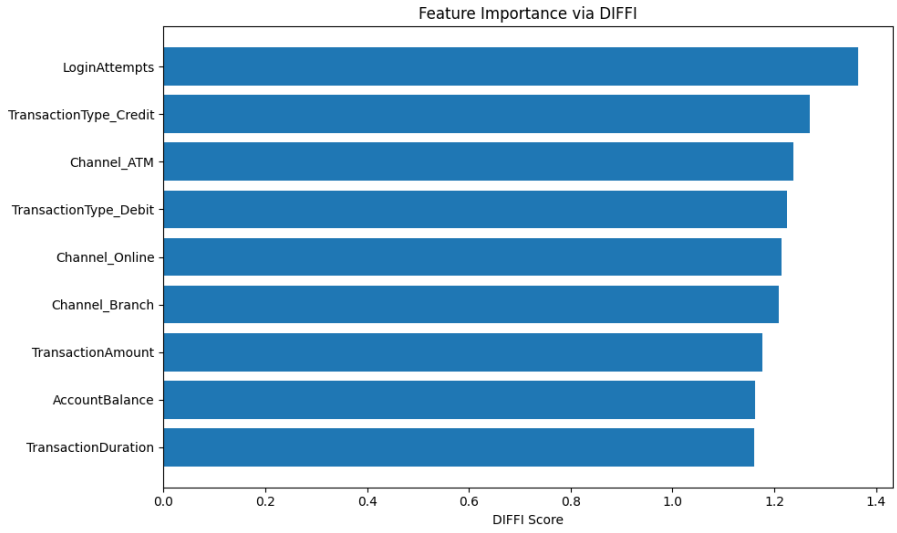

Figure 4.3: DIFFI scores for all features, indicating feature importance via the DIFFI metric.

4.3 SHAP Values

The following plots give a global SHAP explanation of the IF results:

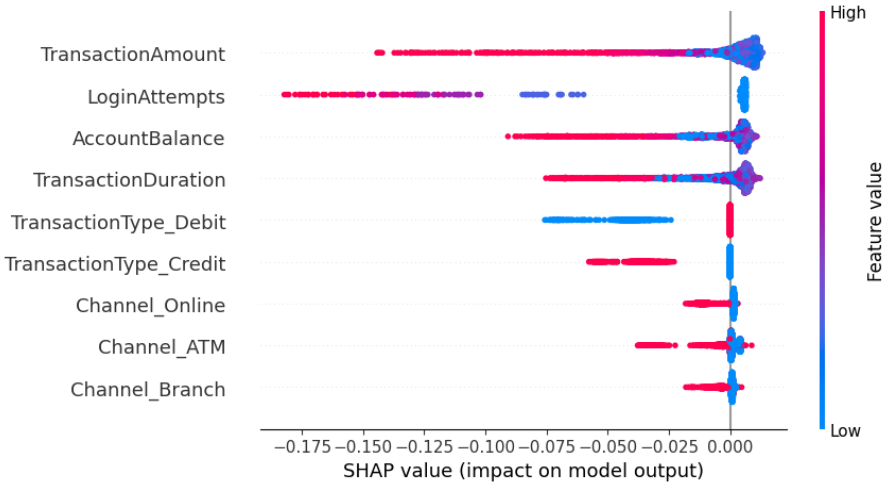

Figure 4.4: SHAP beeswarm plot for the IF results.

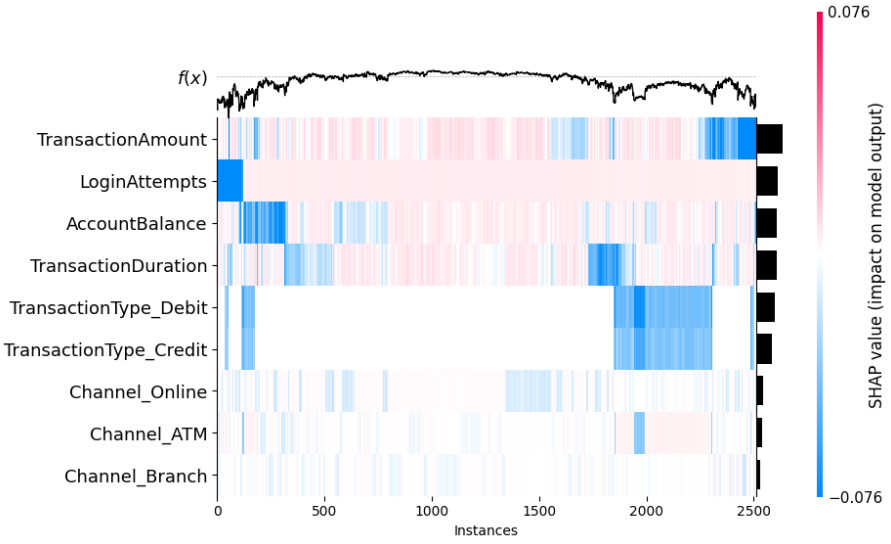

Figure 4.5: SHAP heatmap for the IF results.

Additionally, two local waterfall plots are included for entries detected as anomalies:

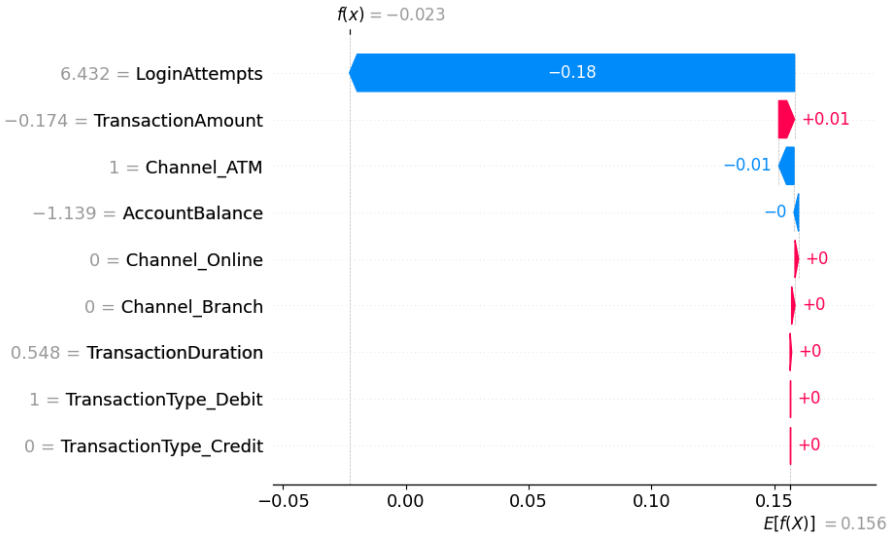

Figure 4.6: SHAP waterfall plot for observation $i = 26$.

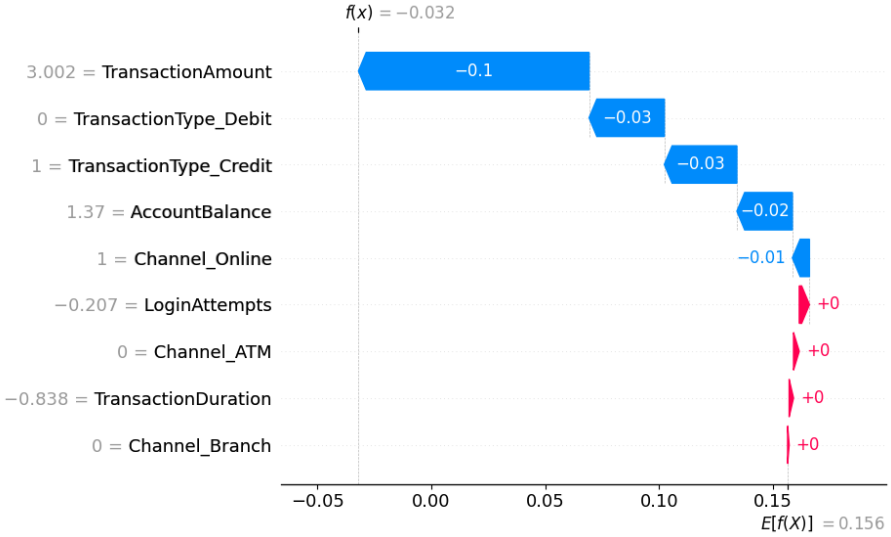

Figure 4.7: SHAP waterfall plot for observation $i = 2380$.

Chapter 5

Discussion

Both DIFFI and SHAP indicated that number of login attempts is crucial for fraudulent transaction detections. The feature has the most effect on the prediction according to SHAP, and has the highest GFI according to DIFFI. However, the DIFFI plot does not clearly show significant differences between feature importances. While this could be true, this seems highly unlikely as the SHAP model did manage to find such differences. High numbers in transaction amount, login attempts, account balance and transaction duration all significantly increase the risk of a fraudulent transaction, which has a logical basis. High amounts of cash being withdrawn could indicate an account plundering, where richer accounts are a more appealing target.

While SHAP improved the interpretability of the IF results (both globally and locally), the DIFFI plot provided no significant explanation. One possible reason for this could be that the implementation was inaccurate. A notable pitfall for using DIFFI is that there exists no official Python implementation for the method at the moment of writing. For this research, the algorithm was developed using the DIFFI paper's pseudocode. (Carletti et al., 2023) The implementation might not entirely adhere to the intended algorithm. Future research could explore developing a sophisticated Python implementation of the DIFFI method, compatible with the sci-kit learn library for IF.

In conclusion, this research explored explainable anomaly detection for bank fraud detection using two explanation methods: DIFFI and SHAP. The research provided insights into the explanation mechanisms and implemented the methods on a real-world use case in a high-stake domain.

Bibliography

Afroogh, S., Akbari, A., Malone, E., Kargar, M., and Alambeigi, H. (2024). Trust in ai: progress, challenges, and future directions. *Humanities and Social Sciences Communications*, 11(1):1–30.

Arranz, D., Bianchini, S., Di Girolamo, V., and Ravet, J. (2023). *Trends in the use of AI in science – A bibliometric analysis*. Publications Office of the European Union.

Carletti, M., Terzi, M., and Susto, G. A. (2023). Interpretable anomaly detection with diffi: Depth-based feature importance of isolation forest. *Engineering Applications of Artificial Intelligence*, 119:105730.

Khorasani, V. (2024). Bank transaction dataset for fraud detection. `https://www.kaggle.com/datasets/valakhorasani/bank-transaction-dataset-for-fraud-detection/data`.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Minh, D., Wang, H. X., Li, Y. F., and Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, pages 1–66.

Singla, A., Sukharevsky, A., Yee, L., Chui, M., and Hall, B. (2025). The state of ai: How organizations are rewiring to capture value. *McKinsey & Company*. `https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai#/`.

Snelders, J. (2025). Explainable anomaly detection with diff and shap. `https://github.com/JustSnelders/XAD-with-DIFF-and-SHAP`.