Attributable Decomposition of Deep Learning Anomaly Detection in Financial Statements

Aakar Tripathi

[a] *tripathiaakar@gmail.com*

Abstract

This paper introduces a novel methodology, Reconstruction Error SHapley Additive exPlanations Ext (RESHAPE), for enhancing the explainability of deep learning models used in the detection of anomalies within financial statement audits. Specifically, RESHAPE provides explanations at an aggregated attribute level for anomalies identified by Autoencoder Neural Networks (AENNs). Furthermore, the paper proposes an evaluation framework to benchmark the efficacy of various explainable AI (XAI) techniques in the context of financial auditing. Empirical results demonstrate RESHAPE's capacity to generate versatile and insightful explanations compared to state-of-the-art baseline methods, thereby contributing to greater transparency in automated audit processes.

Explainable Artificial Intelligence (XAI), Autoencoder Neural Networks (AENNs), Financial Auditing, Anomaly Detection, SHAP, Journal Entries, ERP Systems

1. Introduction

Financial auditing plays a pivotal role in verifying the accuracy and reliability of corporate financial statements. External auditors are tasked with providing assurance that the presented financial information is free from material misstatements caused by either fraud or unintentional errors. The outcomes of these audits directly influence stakeholders, investors, and the broader economy, highlighting the importance of precision and credibility in the auditing process.

With the digital transformation of accounting systems, modern Enterprise Resource Planning (ERP) platforms now log comprehensive and detailed records for every business transaction. This shift has led to an explosion in data volume and complexity, significantly changing how audits are conducted. International auditing standards (e.g., ISA 240 SAS 99 emphasize the necessity for auditors to examine granular accounting data—specifically journal entries (JEs)—to uncover potential red flags indicative of misconduct or systemic errors.

Traditional computer-assisted audit techniques (CAATs), though effective to a degree, are often limited in scope. They typically focus on isolated statistical analysis of specific JE attributes and are not well-equipped to detect complex, multi-attribute anomalies . This limitation has spurred the adoption of AI-driven methodologies, especially those based on deep learning (DL), for their capability to identify subtle, high-dimensional patterns across large datasets .

Autoencoder Neural Networks (AENNs), a type of unsupervised DL model, have gained traction in the auditing domain due to their effectiveness in identifying deviations from learned data distributions. These networks reconstruct input data and assess anomalies through reconstruction error metrics (8; 4). However, despite their performance, AENNs are often seen as 'black boxes' due to the lack of interpretability—particularly problematic in the context of audits, where accountability and transparency are non-negotiable.

To address this, post-hoc explainability tools like SHAP have been employed to provide local feature attributions. Yet, these tools often fall short in unsupervised settings. Existing SHAP adaptations, such as LossSHAP and A-SHAP, either provide overly abstract instance-level explanations or overly granular encoding-level outputs that are difficult for human auditors to interpret effectively (14; 16).

In response to these challenges, we propose RESHAPE—a novel extension of SHAP that leverages reconstruction errors to produce clear, attribute-level explanations for anomalies detected via AENNs. This approach enhances both transparency and usability for auditors. Furthermore, we introduce a domain-specific evaluation framework focused on fidelity, stability, and robustness—key qualities that explanations must meet to be actionable in real-world audits.

Our key contributions are:

- A novel SHAP-based approach, RESHAPE, that explains AENN anomalies on a meaningful attribute level.
- An audit-specific evaluation framework tailored to assess explanation quality across fidelity, stability, and robustness.
- Comprehensive empirical benchmarking on two synthetic and one real-world dataset, demonstrating the practical advantages of RESHAPE over existing methods.

The remainder of this paper is organized as follows: Section II reviews related work in anomaly detection and explainability within audits. Section III details the RESHAPE methodology. Section IV outlines the evaluation framework. Section V presents the experimental design. Section VI discusses results. Finally, Section VII concludes the paper with insights and implications for future work.

2. Related Work

In this section, we provide a detailed overview of the literature relevant to our study, which lies at the intersection of accounting anomaly detection, explainable artificial intelligence (XAI), and model interpretability evaluation. The review is categorized into three thematic areas: (i) the application of Autoencoder Neural Networks (AENNs) for accounting anomaly detection, (ii) the use of XAI techniques in unsupervised anomaly detection, and (iii) existing frameworks for evaluating the effectiveness and quality of explanations.

2.1. Anomaly Detection in Accounting Using AENNs

Anomaly detection has long been a research focus in both data mining and auditing domains. Classical anomaly detection methods—including distance-based and density-based techniques—have been systematically reviewed by researchers such as Ahmed et al. (1), Chandola et al. (2), and Chalapathy and Chawla (3). With the growing complexity of financial data, machine learning, and particularly deep learning models, have demonstrated considerable promise in identifying irregularities.

Among these, Autoencoder Neural Networks (AENNs) have emerged as a prominent approach. Initially explored for anomaly detection by Hawkins et al. (4), AENNs have since been widely adopted for modeling high-dimensional, unlabeled datasets. In fraud detection, for instance, they have been successfully applied to credit card transactions to detect unusual behavior patterns (5; 6).

Specifically in auditing, Schreyer et al. (7) demonstrated how AENNs can detect irregularities in journal entry (JE) datasets by learning typical transaction structures and identifying deviations through reconstruction error. Subsequent work by Schreyer et al. (8) and Schultz and Tropmann-Frick (9) compared AENN outputs with professional auditor judgments, reinforcing their practical relevance. Other studies expanded AENN utility by incorporating temporal structures through LSTM-based AENNs for capturing sequential anomalies . Additionally, Nonnenmacher et al and Schreyer et al. (10) used AENNs to enhance the audit sampling process, while Schreyer et al. (11) employed contrastive self-supervised learning to improve downstream audit tasks.

Despite their strong anomaly detection performance, the interpretability of AENNs remains a bottleneck, particularly in regulated domains such as financial auditing, where explainability is a legal and ethical imperative.

2.2. Explainability in Unsupervised Anomaly Detection

Given that most deep learning models are inherently opaque, the field of Explainable AI (XAI) has produced a variety of tools designed to make model outputs more interpretable. Among the most widely used are LIME (17) and SHAP both of which provide local, post-hoc explanations of model predictions.

SHAP, in particular, leverages the concept of Shapley values from cooperative game theory (24), assigning each input feature an importance score with respect to the model's prediction. While SHAP has been employed in anomaly detection settings—such as explaining PCA-based detection models (16)—its adaptation to unsupervised learning has been more complex due to the lack of clear labels or outputs.

Notably, Antwarg et al. (14) applied SHAP to AENNs by treating the reconstruction error as the output to be explained. Roshan and Zafar (15) used similar techniques for interpreting anomalies in computer networks. These studies typically provide explanations at the instance or encoding level. LossSHAP, for example, interprets the total reconstruction loss of a JE as a whole, while A-SHAP attempts to explain each latent dimension's contribution.

However, these methods fall short in producing actionable insights for auditors. Instance-level explanations often lack granularity, obscuring which specific attributes contributed most to the anomaly. Conversely, encoding-level explanations are excessively fine-grained and scattered across latent dimensions, making them difficult for humans to interpret.

Recent auditing-focused XAI efforts include Rebstadt et al.'s (12) proposal for a personalized XAI role model and a user study by Gnoss et al. (13), which confirmed the usability of SHAP explanations when applied to surrogate models trained on labeled data. However, these approaches sidestep the challenge of directly explaining AENN outputs in their native, unsupervised form.

2.3. Evaluation Frameworks for XAI

Evaluating the quality of explanations remains an open and evolving research area. Doshi-Velez and Kim (20) call for a rigorous science of interpretability, outlining the difficulty in defining and measuring explanation effectiveness. Nguyen and Martínez (21) highlight that no universal benchmark exists, as the utility of an explanation is inherently context-dependent.

Nonetheless, several frameworks and metrics have been proposed. Alvarez-Melis and Jaakkola (22) investigate robustness of interpretability methods, while Mishra et al. (23) survey explanation stability. Amparore et al. (18) developed the LEAF library to evaluate local linear XAI methods through criteria such as local fidelity, conciseness, and prescriptivity. Yang and Kim (19) introduced the BIM framework to compare model interpretability tools quantitatively. Liu et al. created synthetic datasets to measure ground truth faithfulness, monotonicity, ROAR (Remove And Retrain), and infidelity.

Despite these efforts, benchmarking explainability methods remains particularly difficult for unsupervised models like AENNs. Few works offer comprehensive tools for evaluating explanations in auditing-specific contexts—where reproducibility, precision, and attribute-level clarity are critical.

2.4. Motivation for RESHAPE

While prior research has made significant progress in anomaly detection and explanation, a gap persists in methods that offer interpretable, attribute-level explanations specifically tailored to AENNs used in financial auditing. The need for human-understandable, actionable insights is vital to ensure that complex AI systems can be trusted and effectively deployed in high-stakes decision-making environments. Our work, RESHAPE, is designed to bridge this gap by offering a comprehensive,

SHAP-based method optimized for audit settings—delivering explanations that are not only technically sound but practically useful for auditors.

3. Methodology

This section introduces the core methodological framework of RESHAPE (Reconstruction Error SHapley Additive exPlanations Extension), a novel post-hoc explainability method specifically designed to interpret anomalies detected by Autoencoder Neural Networks (AENNs) in financial audit datasets. The section begins with a formal definition of accounting journal entries (JEs), followed by the AENN architecture, a discussion of SHAP as a foundational XAI method, and a comprehensive explanation of how RESHAPE extends SHAP to generate meaningful, attribute-level insights for detected anomalies.

3.1. Formalization of Journal Entries

In financial auditing, the primary data unit under investigation is the journal entry (JE). Let $\mathbb{X}$ represent a dataset comprising N journal entries, formally denoted as $\mathbb{X} = \{x_1, x_2, \ldots, x_N\}$. Each individual journal entry x_i is a structured tuple consisting of a mix of categorical and numerical attributes:

$$x_i = \{a_1^i, a_2^i, \ldots, a_M^i, a_{M+1}^i, \ldots, a_T^i\}$$

where M represents the number of categorical attributes and $K = T - M$ represents the numerical attributes, giving a total of T features per journal entry.

Anomalies in JEs are typically classified into two broad categories:

- **Global Anomalies (Type-A_k)**: These entries contain k attribute values that are completely absent in the training dataset. Such anomalies often reflect rare or unusual account codes, document types, or user IDs, and are thus associated with high risk.
- **Local Anomalies (Type-B_k)**: These entries contain values that are individually common but appear in rare combinations. For example, a valid posting time may be associated with an unlikely user-account pairing. These anomalies suggest potential fraudulent collusion and are also of high audit significance.

3.2. Autoencoder Neural Networks (AENNs)

AENNs are unsupervised neural architectures that learn to reconstruct their input data by passing it through a latent representation. The structure comprises two components: an encoder $f_\theta(\cdot)$ and a decoder $g_\psi(\cdot)$.

Encoder Function:

$$z_i = f_\theta(x_i), \quad z_i \in \mathbb{R}^p$$

Decoder Function:

$$\hat{x}_i = g_\psi(z_i), \quad \hat{x}_i \in \mathbb{R}^n$$

The model is optimized by minimizing the reconstruction loss between input and output:

$$\underset{\theta,\psi}{\operatorname{argmin}} \quad \|x_i - g_\psi(f_\theta(x_i))\| \tag{1}$$

The dimensionality of the bottleneck layer (p) is intentionally constrained such that $\dim(p) < \dim(n)$, forcing the model to learn a compressed representation of the most significant patterns in the data. Any journal entry that deviates substantially from these learned norms will produce a high reconstruction error (RE), defined as:

$$L(x_i; \hat{x}_i)$$

A journal entry x_i is flagged as anomalous if its reconstruction error exceeds a predefined threshold δ, i.e.,

$$L(x_i; \hat{x}_i) > \delta \quad \Rightarrow \quad \tilde{x}_i \text{ is an anomaly}$$

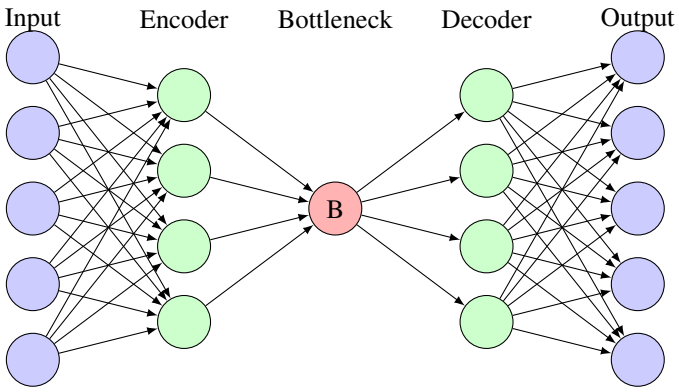


Figure 1: Schematic of an Autoencoder Neural Network used for anomaly detection in accounting records.

3.3. SHapley Additive exPlanations (SHAP)

SHAP provides a unified approach for attributing model outputs to input features based on the concept of Shapley values from cooperative game theory (24). The Shapley value for a feature j represents its average marginal contribution to the model's output across all possible subsets of features:

$$\phi_j = \sum_{S \subseteq K \setminus \{j\}} \frac{|S|!(|K| - |S| - 1)!}{|K|!} \left[\nu(S \cup \{j\}) - \nu(S)\right] \tag{2}$$

Here, K is the set of all features, S is a subset excluding j, and $\nu(\cdot)$ is the model's payoff function.

When applied to AENNs in anomaly detection, existing adaptations include:

- **LossSHAP** (16; 15): Explains the total reconstruction error of a JE by assigning SHAP values to its original features.
- **A-SHAP** (14): Provides attribution at the latent encoding level, offering highly granular explanations.

However, neither of these methods supports audit-level, attribute-based explanations that are both interpretable and diagnostically useful.

3.4. RESHAPE: Attribute-Level SHAP for Anomaly Explanations

RESHAPE extends SHAP to quantify how individual input attributes contribute to the reconstruction error of their corresponding outputs. The approach enables interpretable, attribute-level explanations as follows:

Let $o_m = \{o_{1,m}, o_{2,m}, \ldots, o_{k,m}\}$ represent the k-dimensional encoding of the m-th attribute in a JE, and $\hat{o}_m$ its reconstruction. The attribute-specific reconstruction loss is defined as:

$$L(a_m; \hat{a}_m) = \sum_{k=1}^{K} L(o_{m,k}; \hat{o}_{m,k}) \tag{3}$$

RESHAPE proceeds in three main steps (see Fig. 2):

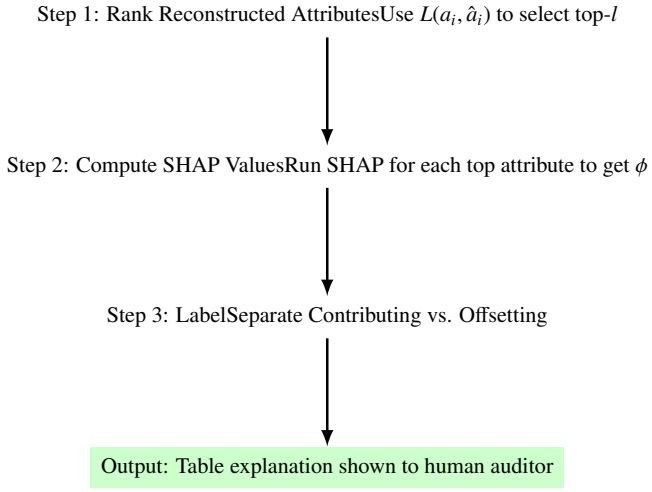


Figure 2: RESHAPE explanation generation process: Attribute-based reconstruction errors are ranked, SHAP values are computed per top-l attribute, and results are presented to auditors.

1. **Ranking Attributes:** The attributes of a JE are sorted in descending order based on their reconstruction error $L(a_m^i; \hat{a}_m^i)$. The top-l attributes are selected as candidates for SHAP-based attribution.
2. **Computing SHAP Values:** For each selected attribute, a separate SHAP run is conducted to evaluate the impact of all input attributes on its reconstruction loss. The SHAP values $\phi_l = \{\phi_{l1}, \phi_{l2}, \ldots, \phi_{lm}\}$ quantify this influence.
3. **Classifying Attributions:** Attributes with positive SHAP values ($\phi_{lj} > 0$) are labeled as *contributing*, while those with negative values ($\phi_{lj} < 0$) are deemed *offsetting*. The final output is a tabulated explanation showing which attributes most significantly contributed to or mitigated the anomaly.

This structured approach provides a high-level yet interpretable view of why a JE was flagged as anomalous, enabling auditors to diagnose the issue at an attribute level rather than parsing dense encodings or black-box outputs.

4. Evaluation Framework

To ensure that the explanations generated by RESHAPE are suitable for practical auditing scenarios, it is necessary to assess their quality using well-defined and interpretable metrics. This section introduces an evaluation framework tailored specifically for financial audits. It emphasizes three essential properties that explanations must exhibit to be actionable by human auditors: **fidelity**, **stability**, and **robustness**. Each of these dimensions is associated with specific quantitative metrics that are applied to evaluate the behavior of RESHAPE and comparable XAI methods.

4.1. Explanation Fidelity

Fidelity measures the degree to which an explanation correctly reflects the model's internal decision-making process. In the context of anomaly detection for auditing, high-fidelity explanations indicate that the attributes identified as anomalous genuinely influenced the AENN's reconstruction error. Three fidelity metrics are employed in this work:

4.1.1. Mean Reciprocal Rank (MRR_r)

This metric quantifies how early a relevant attribute appears in an ordered list of explanatory features for a given anomaly. Specifically, it measures the inverse of the rank at which the first truly anomalous attribute is found across all explanations.

$$\text{MRR}_r = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{\text{rank}_i^R} \tag{4}$$

Here, M is the total number of evaluated explanations, $R = \{a_j^i, \ldots, a_r^i\}$ denotes the set of ground truth anomalous attributes, and rank_i^R is the position at which the first relevant attribute appears in the i-th explanation.

4.1.2. Hits@n Metric

The Hits@n metric evaluates whether any relevant attribute is included among the top-n ranked features in the explanation. This is practical for audit applications, as human auditors typically focus on the top few explanations.

$$\text{Hits@}n_i = \begin{cases} 1, & \text{if } \text{rank}_i^R \leq n \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Again, rank_i^R represents the rank of the first correct attribute in explanation i, and n is a pre-defined threshold (e.g., top 5 or 10 features).

4.1.3. Anomaly Score Reduction

This metric measures how efficiently an explanation helps transform an anomalous journal entry into a non-anomalous one. It does so by sequentially modifying the top-ranked explanatory attributes and observing how the reconstruction error changes. Formally, the error reduction percentage after n modifications is defined as:

$$\text{error}\%_i^n = \frac{L(x_i'; \hat{x}_i')^n}{L(x_i; \hat{x}_i)} \tag{6}$$

Here, x_i' is the modified journal entry after changing the top-n attributes, and $L(\cdot)$ denotes the reconstruction loss. A lower value of error$\%_i^n$ indicates that the most explanatory attributes identified by the model were indeed driving the anomaly.

4.2. Explanation Stability

Stability captures the consistency of an explanation when the explanation process is repeated under different random seeds or sampling configurations. In audit settings, where reproducibility is paramount, a reliable explanation should not vary significantly across multiple runs.

To measure stability, K independent explanation runs are conducted for each anomaly. Then, for the top-n ranked attributes, we compute the variance in ranks for each attribute. The overall Stability Index S_n is given by:

$$S_n = \sqrt{\frac{V_1 + V_2 + \cdots + V_n}{n}} \tag{7}$$

Where $V_i = \mathrm{Var}(r_{1i}, r_{2i}, \ldots, r_{Ki})$ denotes the variance in ranks of the i-th attribute across K runs. A smaller S_n implies greater consistency and thus a more stable explanation.

4.3. Explanation Robustness

Robustness evaluates how much attention an explanation method gives to features that are known to be irrelevant or uninformative. In an audit context, highlighting such attributes can mislead the auditor, reducing the credibility of the system.

We define the robustness using the Mean Reciprocal Rank of uninformative attributes, denoted as MRR_u. This metric mirrors Eq. 4, but now focuses on detecting how highly uninformative attributes are ranked.

$$\mathrm{MRR}_u = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{\mathrm{rank}_i^U} \tag{8}$$

Where $U = \{a_j^i, \ldots, a_r^i\}$ is the set of known uninformative attributes, and rank_i^U represents the position of the first such attribute in explanation i. A lower MRR_u indicates better robustness, as irrelevant features are ranked lower and contribute less to the explanation.

4.4. Summary

The combination of these three dimensions—fidelity, stability, and robustness—creates a holistic framework for evaluating XAI methods in sensitive domains such as financial auditing. This multi-metric approach ensures that the explanations generated by RESHAPE are not only correct but also consistent and trustworthy. In the next section, we describe the experimental design used to apply this framework and validate RESHAPE's performance against existing methods.

5. Experimental Setup

This section outlines the experimental design used to evaluate the performance of RESHAPE in comparison to baseline explainability methods. Our experimental process follows a two-phase approach: (1) training AENN models on synthetic and real-world datasets to detect anomalies, and (2) applying various XAI techniques—including RESHAPE—to explain the detected anomalies. The effectiveness of these explanations is then assessed using the evaluation framework introduced in Section IV.

5.1. Datasets and Preprocessing

To ensure both diversity and audit-relevance, we employ two synthetic datasets and one publicly available real-world dataset. Each dataset is encoded using one-hot or ordinal schemes to match the input format expected by AENNs.

- **Synthetic Boolean Dataset**: This dataset includes 15 randomly generated Boolean variables along with 5 variables computed using logical operations (e.g., `OR`, `XOR`). It contains approximately 2.09 million records, into which 75,000 anomalies have been synthetically injected. Each data instance $x_i \in \mathbb{R}^{20}$ after encoding.
- **Synthetic Accounting Dataset**: This dataset mimics journal entries (JEs) from enterprise ERP systems such as SAP. It contains 533,091 transactions with 7 categorical and 2 numerical features. A total of 280 synthetic anomalies were introduced. The encoded dimensionality per entry is $x_i \in \mathbb{R}^{704}$.
- **Real-World Payment Dataset**: Sourced from a city government's public finance portal, this dataset includes 238,894 vendor payments with 10 categorical and 1 numerical attribute. An additional 300 synthetic anomalies were manually inserted. The high cardinality of categorical fields results in an encoded dimensionality of $x_i \in \mathbb{R}^{8,565}$.

Note: For anonymization and compliance with double-blind review protocols, the URLs for dataset access are temporarily redacted.

Table 1: Overview of datasets and dimensionality after encoding

Dataset	# Records	Anomalies	Encoded Dim.
Boolean (Synthetic)	2.09M	75,000	20
Accounting (Synthetic)	533,091	280	704
Payment (Real-World)	238,894	300	8,565

5.2. Model Architecture and Training Protocol

To train effective AENNs for each dataset, we define dataset-specific network architectures with fully connected layers and Leaky ReLU activation functions ($\alpha = 0.4$). The encoder and decoder are symmetrically designed, with dimensionality decreasing to a latent bottleneck layer and expanding back to the original shape.

Table 2: Architectural details for AENNs across datasets

Dataset	Layer Structure (Encoder-Decoder)
Boolean	20-18-16-15-16-18-20
Accounting	704-512-256-128-[...]-3-4-8-[...]-256-512-704
Payment	6,358-5,096-2,048-[...]-3-4-8-[...]-2,048-5,096-6,358

The loss function used for optimization is the binary cross-entropy loss:

$$L_{BCE}(x_i; \hat{x}_i) = \frac{1}{N} \sum_{i=1}^{N} [x_i \log(\hat{x}_i) + (1 - x_i) \log(1 - \hat{x}_i)] \quad (9)$$

We use the Adam optimizer with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate of 0.0001. Models were trained using mini-batches of size $\rho = 128$ and early stopping was applied once loss convergence was observed.

Table 3: Background set sizes (η) used for SHAP explanations

Dataset	Background Set Size (η)
Boolean (Synthetic)	500
Accounting (Synthetic)	500
Payment (Real-World)	250

5.3. Explanation Setup

After training, the AENNs are used to identify anomalies based on reconstruction error thresholds. These flagged anomalies are subsequently explained using multiple XAI approaches:

- **RESHAPE (ours)**: Attribute-level explanations derived from reconstruction error SHAP extensions.
- **LossSHAP** (16): Explains instance-level reconstruction loss.
- **A-SHAP** (14): Explains encoding-level features.
- **Random**: Attributes ranked randomly to serve as a baseline.

For each explanation method, three separate runs are conducted using different random seeds to account for stochastic variance in background set sampling and SHAP estimation. For computational efficiency, the background set size for SHAP (η) is limited for the high-dimensional Payment dataset.

—

6. Experimental Results

In this section, we present the empirical outcomes of applying RESHAPE and baseline XAI methods to anomalies identified by AENNs. We analyze the results using the three primary evaluation dimensions outlined in Section IV: **fidelity**, **stability**, and **robustness**. Performance comparisons are conducted across the synthetic Boolean, synthetic Accounting, and real-world Payment datasets.

6.1. Explanation Fidelity

6.1.1. Mean Reciprocal Rank (MRR_r)

To assess the fidelity of different XAI approaches, we first examine how effectively each method ranks relevant anomalous attributes. Table 4 summarizes the MRR_r results for Boolean AND and OR logic anomalies, using 100 randomly injected anomaly instances per test case.

Table 4: MRR_r on Boolean Dataset (Higher is Better)

Method	Independent	Dependent	Output Attribute
	Boolean AND Dependency		
Random	0.33 ± 0.29	0.27 ± 0.28	0.16 ± 0.18
LossSHAP (16)	0.93 ± 0.19	0.43 ± 0.25	0.81 ± 0.33
A-SHAP (14)	**1.00 ± 0.00**	**0.98 ± 0.12**	0.17 ± 0.16
RESHAPE (Ours)	**1.00 ± 0.00**	0.61 ± 0.21	**0.86 ± 0.28**
	Boolean OR Dependency		
Random	0.33 ± 0.30	0.26 ± 0.26	0.17 ± 0.21
LossSHAP	0.93 ± 0.18	0.49 ± 0.24	0.79 ± 0.35
A-SHAP	**1.00 ± 0.00**	**1.00 ± 0.00**	0.20 ± 0.03
RESHAPE	**1.00 ± 0.00**	0.64 ± 0.22	**0.82 ± 0.30**

RESHAPE consistently achieves high MRR_r scores, particularly in identifying output features responsible for the anomaly. While A-SHAP performs well on independent variables, its performance drops on output attributes, which are critical in auditing.

6.1.2. Hits@n Accuracy

Table 5 presents the Area Under the Curve (AUC) for Hits@n results across the synthetic accounting and payment datasets. These results validate whether relevant anomalous attributes appear among the top-n explanatory features.

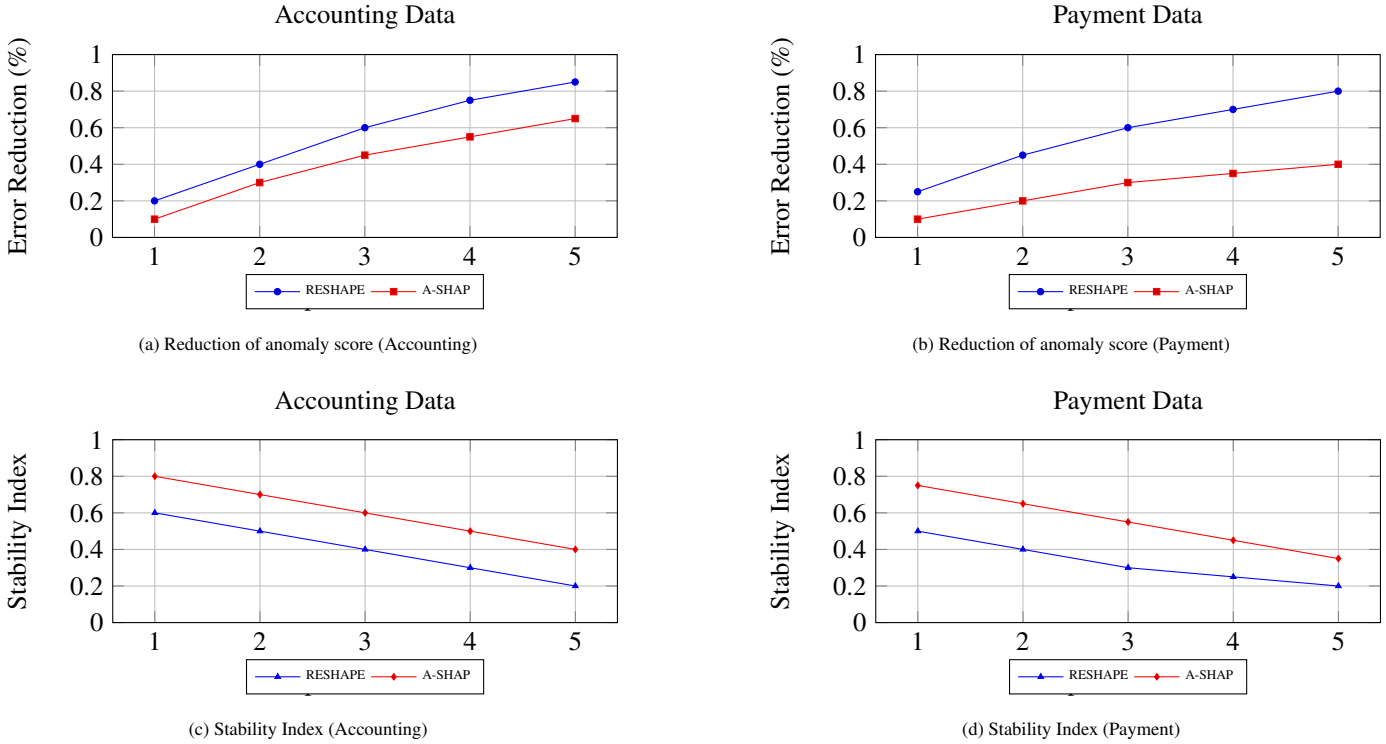
Table 5: AUC for Hits@n (Higher is Better)

XAI Method	Accounting Data	Payment Data
Random	4.11 ± 2.04	5.15 ± 2.91
LossSHAP	5.25 ± 2.50	6.21 ± 2.78
A-SHAP	2.94 ± 2.26	3.41 ± 2.96
RESHAPE (Ours)	**6.61 ± 1.15**	**7.37 ± 2.92**

RESHAPE outperforms all baselines across both datasets, indicating that it more consistently ranks relevant anomalies among the top features — a highly desirable property for interpretability in audits.

6.1.3. Anomaly Score Reduction

Figure **??** shows the effectiveness of each method in reducing anomaly scores by modifying top-ranked attributes. Both RESHAPE and LossSHAP demonstrate a sharp drop in reconstruction error, confirming that their explanations identify attributes causally linked to the anomaly.



(a) Reduction of anomaly score (Accounting)

(b) Reduction of anomaly score (Payment)

(c) Stability Index (Accounting)

(d) Stability Index (Payment)

Figure 3: (a, b) Anomaly score reduction using top explanatory attributes. (c, d) Explanation stability index over multiple SHAP runs.

6.2. *Explanation Stability*

Figure **??** compares the stability index of each method across multiple explanation runs. RESHAPE and LossSHAP both show lower variance in the ranking of top explanatory attributes compared to A-SHAP, indicating higher consistency.

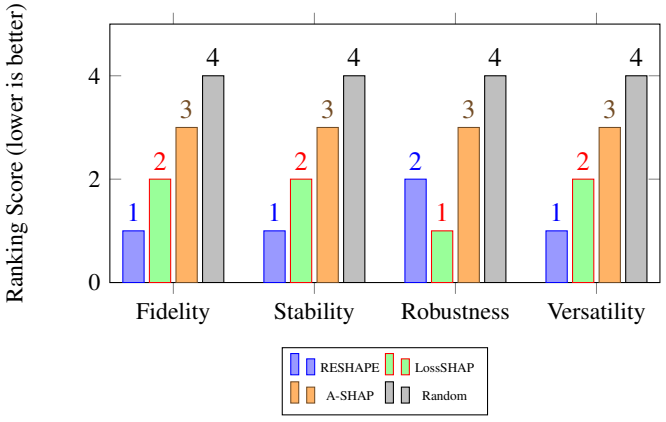


Figure 4: Ranking of XAI methods across evaluation criteria. Lower scores indicate better performance. RESHAPE consistently ranks highest across all metrics.

6.3. *Explanation Robustness*

Robustness is assessed by examining the importance attributed to known uninformative features. Table 6 presents the MRR_u scores. RESHAPE assigns less importance to irrelevant features compared to A-SHAP and random baselines, and performs closely to LossSHAP.

Table 6: Robustness Evaluation: MRR_u (Lower is Better)

XAI Method	Accounting Data	Payment Data
Random	0.32 ± 0.14	0.34 ± 0.11
LossSHAP	**0.17 ± 0.02**	**0.14 ± 0.02**
A-SHAP	0.21 ± 0.08	0.21 ± 0.08
RESHAPE (Ours)	0.18 ± 0.03	0.17 ± 0.04

6.4. *Summary of Results*

Figure **??** summarizes the overall performance of each method across the evaluation dimensions. RESHAPE consistently ranks either first or second in all metrics, showcasing its versatility and suitability for audit applications.

7. Conclusion

This paper introduced RESHAPE, a novel explainability method that extends SHapley Additive exPlanations (SHAP) to produce attribute-level insights for anomalies identified by Autoencoder Neural Networks (AENNs). RESHAPE addresses a critical challenge in financial auditing—interpreting the output of complex, unsupervised deep learning models in a transparent and actionable manner.

We proposed an evaluation framework centered on three essential dimensions—fidelity, stability, and robustness—specifically designed for auditing applications. Through comprehensive experiments on synthetic and real-world datasets, we demonstrated that RESHAPE outperforms existing XAI baselines such

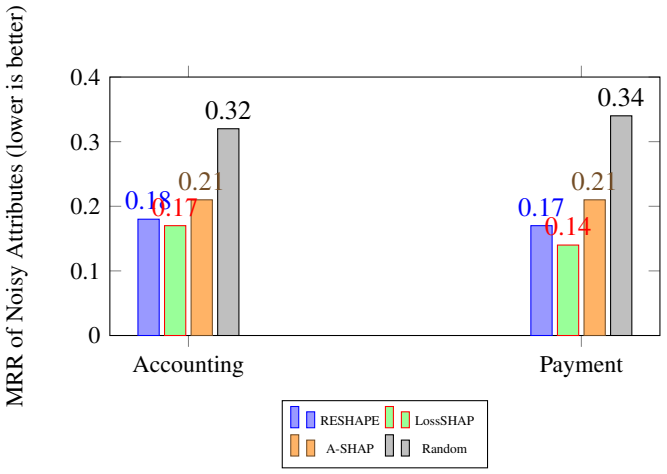


Figure 5: Robustness comparison based on MRR of uninformative (noisy) attributes. Lower values indicate better robustness.

as LossSHAP and A-SHAP. RESHAPE consistently delivered higher-quality explanations that not only aligned with model behavior but also provided interpretable justifications that auditors could use confidently in their assessments.

Our results suggest that RESHAPE can help bridge the gap between high-performing anomaly detection models and their practical deployment in high-stakes domains like finance and regulation. By enhancing the interpretability of AENN-based audit models, RESHAPE contributes to greater trust, transparency, and adoption of AI in critical auditing workflows.

Future work may extend this approach to incorporate temporal dependencies, multi-modal data inputs, or hybrid models that combine supervised and unsupervised learning strategies. Additionally, integrating RESHAPE into interactive audit support systems could further empower auditors with real-time, trustworthy explanations during investigations.

References

[1] M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," *Journal of Network and Computer Applications*, vol. 60, pp. 19–31, 2016.

[2] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, Jul. 2009.

[3] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," *arXiv preprint arXiv:1901.03407*, 2019.

[4] S. Hawkins, H. He, G. Williams, and R. Baxter, "Outlier detection using replicator neural networks," in *Proc. Int. Conf. on Data Warehousing and Knowledge Discovery*, 2002, pp. 170–180.

[5] Z. Kazemi and H. Zarrabi, "Using deep networks for fraud detection in the credit card transactions," in *Proc. IEEE Int. Conf. on Knowledge-Based Engineering and Innovation (KBEI)*, 2017, pp. 630–633.

[6] A. Pumsirirat and L. Yan, "Credit card fraud detection using deep learning based on auto-encoder and restricted Boltzmann machine," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 1, pp. 18–25, 2018.

[7] M. Schreyer, T. Sattarov, D. Borth, A. Dengel, and B. Reimer, "Detection of anomalies in large scale accounting data using deep autoencoder networks," *arXiv preprint arXiv:1709.05254*, 2017.

[8] M. Schreyer, T. Sattarov, C. Schulze, B. Reimer, and D. Borth, "Detection of accounting anomalies in the latent space using adversarial autoencoder neural networks," in *Proc. KDD Workshop on Anomaly Detection in Finance*, 2019.

[9] M. Schultz and M. Tropmann-Frick, "Autoencoder neural networks versus external auditors: Detecting unusual journal entries in financial statement audits," in *Proc. 53rd Hawaii Int. Conf. on System Sciences (HICSS)*, 2020.

[10] M. Schreyer, T. Sattarov, A. Gierbl, B. Reimer, and D. Borth, "Learning sampling in financial statement audits using vector quantised variational autoencoder neural networks," in *Proc. ACM Int. Conf. on AI in Finance (ICAIF)*, 2020.

[11] M. Schreyer, T. Sattarov, and D. Borth, "Multi-view contrastive self-supervised learning of accounting data representations for downstream audit tasks," in *Proc. ACM ICAIF*, 2021.

[12] J. Rebstadt, F. Remark, P. Fukas, P. Meier, and O. Thomas, "Towards personalized explanations for AI systems: Designing a role model for explainable AI in auditing," in *Wirtschaftsinformatik 2022 Proceedings*, 2022.

[13] N. Gnoss, M. Schultz, and M. Tropmann-Frick, "XAI in the audit domain – Explaining an autoencoder model for anomaly detection," in *Proc. 17th Int. Conf. on Wirtschaftsinformatik*, 2022.

[14] L. Antwarg, R. M. Miller, B. Shapira, and L. Rokach, "Explaining anomalies detected by autoencoders using SHAP," *arXiv preprint arXiv:1903.02407*, 2019.

[15] K. Roshan and A. Zafar, "Utilizing XAI technique to improve autoencoder based model for computer network anomaly detection with SHAP," *arXiv preprint arXiv:2112.08442*, 2021.

[16] N. Takeishi, "Shapley values of reconstruction errors of PCA for explaining anomaly detection," in *Proc. IEEE Int. Conf. on Data Mining Workshops (ICDMW)*, 2019, pp. 793–798.

[17] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[18] E. Amparore, A. Perotti, and P. Bajardi, "To trust or not to trust an explanation: Using LEAF to evaluate local linear XAI methods," *PeerJ Computer Science*, vol. 7, pp. e484, 2021.

[19] J. Yang and B. Kim, "Benchmark interpretability methods (BIM)," *arXiv preprint arXiv:2211.04572*, 2022.

[20] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[21] A.-P. Nguyen and M. R. Martínez, "On quantitative aspects of model interpretability," *arXiv preprint arXiv:2007.07584*, 2020.

[22] D. Alvarez-Melis and T. Jaakkola, "On the robustness of interpretability methods," *arXiv preprint arXiv:1806.08049*, 2018.

[23] S. Mishra, S. Dutta, J. Long, and D. Magazzeni, "A survey on the robustness of feature importance and counterfactual explanations," *arXiv preprint arXiv:2111.00358*, 2021.

[24] L. S. Shapley, "Notes on the n-person game: The value of an n-person game," *RAND Corporation*, 1951.