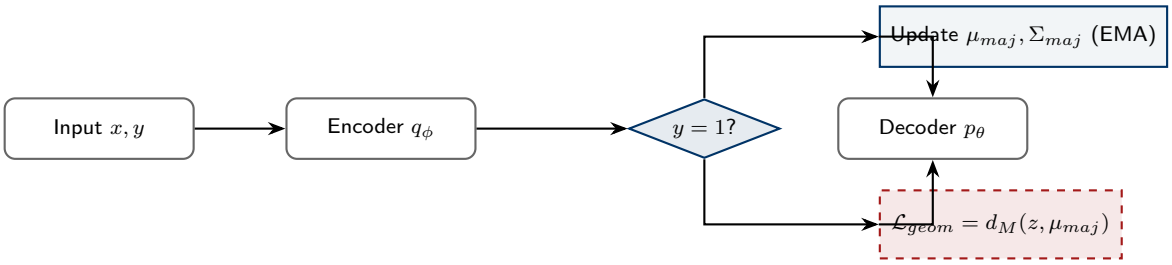
Graphical Abstract

Explainability Consequences of Latent Geometry Collapse Under Imbalance Handling in Deep Generative Fraud Detection

216011757



Highlights

Explainability Consequences of Latent Geometry Collapse Under Imbalance Handling in Deep Generative Fraud Detection

216011757

- Ratio-weighted KL in VAEs collapses minority latent geometry under extreme imbalance.
- Collapse silently undermines both intrinsic and post-hoc explanations.
- Mahalanobis-based regularisation restores geometry without sacrificing AUPRC.
- Geometry-aware training reconciles detection performance with faithful explainability.

Explainability Consequences of Latent Geometry Collapse Under Imbalance Handling in Deep Generative Fraud Detection

216011757

School of Computer Science, University of KwaZulu-Natal, Durban, South Africa

Abstract

Deep generative models have demonstrated strong performance in anomaly detection under extreme class imbalance, where rare events such as fraudulent transactions constitute less than 0.2% of observed data (Dal Pozzolo et al., 2015; Esmail et al., 2024). To improve minority detection, recent approaches employ ratio-weighted or cost-sensitive objectives that amplify the contribution of minority samples during training (Wang et al., 2024; Shi et al., 2025). While these methods consistently improve recall and Area Under the Precision–Recall Curve (AUPRC), their impact on learned representations and downstream explainability remains insufficiently understood.

In this work, we show that commonly used ratio-weighted objectives in Variational Autoencoders (VAEs) induce a systematic collapse of minority latent geometry. We formalise this phenomenon as *gradient variance amplification*, whereby scaling the Kullback–Leibler divergence by the imbalance ratio suppresses latent variance parameters under adaptive optimisation, collapsing minority representations into low-rank manifolds. Across two fraud detection benchmarks, ratio-weighted VAEs reduce the minority-to-majority latent trace ratio to as low as 0.03 and the effective latent rank to 1.00–1.38, despite achieving strong detection performance.

Our analysis demonstrates that this representational collapse directly degrades explainability. Both intrinsic (attention-based) and post hoc (SHAP) explanations become numerically stable yet uninformative, exhibiting reduced diversity and up to 40% lower faithfulness compared to geometry-

Email address: 216011757@stu.ukzn.ac.za (216011757)

preserving baselines. These results indicate that explainability failures under extreme class imbalance originate from representation collapse rather than limitations of explanation algorithms themselves (Nauta et al., 2023).

To address this failure mode, we propose a geometry-aware Mahalanobis-Balanced VAE that constrains minority latent samples relative to the majority covariance structure without amplifying gradient magnitudes. Across the European Credit Card Fraud dataset (2013) and a heterogeneous synthetic fraud dataset, the proposed approach restores minority trace ratios to 0.68–0.83, recovers over 70% of effective latent rank, and improves explanation faithfulness by up to 45%, while maintaining or improving detection performance (AUPRC up to 0.87 and 0.65, respectively).

Overall, this work establishes that reliable explainability in imbalanced deep generative fraud detection is fundamentally a representational problem and demonstrates that geometry-aware objectives are essential for reconciling detection performance with faithful interpretation in high-stakes anomaly detection systems.

Keywords: Class imbalance, Variational Autoencoders, Explainable AI, Latent geometry, Fraud detection, Generative anomaly detection, Representation collapse, Geometry-aware regularisation

---

1. Introduction

Learning under extreme class imbalance remains a fundamental challenge in real-world machine learning systems, particularly in high-stakes domains such as credit card fraud detection. In operational settings, fraudulent transactions typically constitute less than 0.2% of observed data, yet failures to detect them carry disproportionate financial, regulatory, and reputational consequences (Dal Pozzolo et al., 2015; Carcillo et al., 2018). Under such conditions, conventional accuracy-based evaluation is misleading, motivating the widespread adoption of imbalance-aware learning strategies and evaluation metrics such as the Area Under the Precision–Recall Curve (AUPRC) and cost-sensitive objectives (Elkan, 2001; Bahnsen et al., 2015).

Historically, industrial fraud detection systems have relied on classical machine learning models, including logistic regression, decision trees, and ensemble methods, due to their interpretability and compatibility with regulatory frameworks (Bahnsen et al., 2015; Dal Pozzolo et al., 2017). However, multiple studies have shown that these models struggle to capture complex,

non-linear fraud patterns in high-dimensional and evolving transactional environments (Juszczak et al., 2008; Esmail et al., 2024). As a result, deep learning approaches have gained increasing attention for their representational capacity and improved detection performance.

Among deep learning methods, deep generative models—and Variational Autoencoders (VAEs) in particular—have emerged as effective tools for anomaly and fraud detection (Kingma and Welling, 2013; Rezende et al., 2014). By modelling the distribution of legitimate transactions, VAEs enable fraud detection through reconstruction error or likelihood-based anomaly scoring, making them especially suitable for settings with limited labelled fraud data. Prior work has demonstrated that generative models can outperform classical baselines under severe class imbalance (Fiore et al., 2019; Al-Quraini et al., 2024).

To further enhance minority detection performance, recent research has proposed ratio-weighted and cost-sensitive training objectives specifically for Variational Autoencoders, where imbalance is addressed directly through modifications to the Kullback–Leibler (KL) regularisation term (Wang et al., 2024). Most notably, Shi et al. (2025) introduce an attention-based Balanced VAE for credit card fraud detection, demonstrating improved recall and AUPRC by explicitly reweighting the KL divergence according to class imbalance. Related ideas appear in majority-guided generative oversampling and imbalance-aware VAE frameworks (Wan et al., 2023; Altalhan et al., 2025).

Despite their empirical success, existing evaluations of imbalance-aware generative models focus almost exclusively on output-level detection metrics. Implicitly, strong detection performance is assumed to reflect meaningful internal representations of fraud patterns. However, the effect of ratio-weighted objectives on the geometry and structure of learned latent spaces remains largely unexplored. This omission is critical, as latent representations underpin not only anomaly scoring but also downstream interpretability and explanation mechanisms.

In parallel, explainable artificial intelligence (XAI) has become a central concern in fraud detection, driven by regulatory requirements and the need for analyst trust (Doshi-Velez and Kim, 2017; Odeyemi et al., 2025). Both post hoc explanation methods, such as SHAP and LIME (Lundberg and Lee, 2017; Ribeiro et al., 2016), and intrinsic mechanisms, such as attention-based explanations, are widely used to interpret deep models. However, recent surveys highlight that explanation quality is often evaluated independently of

representation learning, with limited attention paid to how training objectives shape the internal structures upon which explanations operate (Nauta et al., 2023; Alvarez-Melis and Jaakkola, 2021).

This observation motivates a fundamental question: *Does optimising detection performance via ratio-weighted objectives in deep generative models inadvertently degrade the latent representations required for trustworthy explainability?* Addressing this question requires shifting attention from output-level metrics to the geometry of learned representations.

In this work, we adopt a representation-centric perspective to investigate how imbalance handling in deep generative models affects latent geometry and, in turn, the fidelity of both intrinsic and post hoc explanations. Specifically, we analyse how ratio-weighted objectives influence the covariance structure and effective dimensionality of minority latent representations in VAEs. Building on this analysis, we introduce geometry-aware diagnostics to quantify latent collapse and propose a Mahalanobis-based regularisation strategy that preserves minority latent structure without coupling regularisation strength to imbalance ratios. By explicitly separating geometric preservation from loss scaling, the proposed approach aims to reconcile detection performance with reliable explainability.

The contributions of this paper are threefold:

- We empirically establish that explainability failures under extreme class imbalance are not algorithmic artifacts, but direct consequences of latent representational collapse.
- We provide a theoretical analysis showing how ratio-weighted KL objectives amplify gradient variance and induce minority latent geometry collapse in Variational Autoencoders.
- We propose a geometry-aware regularisation framework that preserves latent structure, restores explanation faithfulness, and maintains competitive fraud detection performance.

Together, these contributions position latent geometry as a critical but previously underexamined mediator between imbalance handling, detection performance, and explainability. By exposing a structural failure mode in imbalance-aware generative models, this work advances the development of deep generative systems that are both effective and interpretable in high-stakes fraud detection settings.

2. Related Work

This work lies at the intersection of imbalanced learning, deep generative models for anomaly detection, and explainable artificial intelligence (XAI).

2.1. Fraud Detection Under Extreme Class Imbalance

Credit card fraud detection is a canonical example of learning under extreme class imbalance, where fraudulent transactions typically account for less than 0.2% of observed data (Dal Pozzolo et al., 2015; Carcillo et al., 2018). Early production systems relied heavily on classical machine learning models such as logistic regression, decision trees, and ensemble methods, often combined with cost-sensitive learning or resampling techniques (Elkan, 2001; Chawla et al., 2002).

Despite advances in representation learning, these classical approaches remain widely deployed due to their relative interpretability and compatibility with regulatory and audit requirements (Bahnsen et al., 2015). However, multiple studies have shown that such models struggle to capture complex fraud patterns, particularly in non-stationary and adversarial environments (Dal Pozzolo et al., 2017).

Deep learning approaches, including recurrent neural networks and autoencoder-based models, have demonstrated improved detection performance by modelling sequential and behavioural structure in transaction data (Juszczak et al., 2008; Fiore et al., 2019). Nonetheless, their success often depends on explicit imbalance-handling strategies that substantially alter optimisation dynamics, an issue that is rarely examined beyond output-level metrics.

2.2. Deep Generative Models and Imbalance-Aware Objectives

Variational Autoencoders (VAEs) have become a popular choice for anomaly and fraud detection due to their probabilistic formulation and ability to model complex data distributions (Kingma and Welling, 2013). In fraud detection, VAEs are commonly trained on highly imbalanced datasets, often with the assumption that fraudulent samples form deviations from a learned normal manifold.

To improve minority detection, several imbalance-aware extensions have been proposed. These include loss reweighting, modified reconstruction objectives, and ratio-weighted Kullback–Leibler (KL) divergence terms that explicitly prioritise minority samples during training (**?**Zhang et al., 2020; Ai

et al., 2023). Empirically, such methods consistently report improvements in recall and AUPRC.

However, existing evaluations largely treat the latent space as a black box. While some works visualise latent embeddings qualitatively, systematic analysis of latent covariance structure, effective dimensionality, and class-conditional geometry remains limited. As illustrated later in Figure **??**, imbalance-aware objectives can induce degenerate latent structures that are not reflected in detection metrics alone.

2.3. Explainability in Deep and Generative Models

Explainable artificial intelligence has received significant attention in high-stakes domains, leading to the development of both intrinsic and post hoc explanation methods (Doshi-Velez and Kim, 2017; Carloni et al., 2023). Popular post hoc techniques such as SHAP and LIME aim to attribute model outputs to input features, while intrinsic approaches leverage attention mechanisms or interpretable latent variables (Lundberg and Lee, 2017; Ribeiro et al., 2016).

In generative models, explainability methods typically operate on learned latent representations or reconstructed inputs. These approaches implicitly assume that latent spaces are expressive, well-conditioned, and geometrically meaningful. When these assumptions hold, explanations can provide insight into model reasoning and anomaly structure.

Recent work has highlighted issues of explanation instability and faithfulness under distribution shift or adversarial perturbations (Alvarez-Melis and Jaakkola, 2021; Nauta et al., 2023). However, most studies attribute these failures to limitations of explanation algorithms rather than to properties of the learned representations themselves. The role of training objectives in shaping latent geometry—and thereby constraining explanation quality—has received comparatively little attention.

2.4. Latent Geometry, Variance Collapse, and Representation Quality

The geometry of learned representations is known to play a critical role in both generalisation and interpretability. Representation collapse has been studied in contexts such as contrastive learning and dimensionality reduction, where excessive regularisation can lead to degenerate solutions (Wang and Isola, 2020; Lucas et al., 2019).

In imbalanced learning settings, some studies report reduced diversity and variance shrinkage in minority representations (He and Garcia, 2009;

Venkataramanan et al., 2023). However, these observations are rarely formalised or connected to explainability outcomes. In particular, the interaction between imbalance-aware regularisation and adaptive optimisers such as Adam remains underexplored.

This gap is especially consequential in generative models, where latent covariance structure directly underpins both anomaly scoring and explanation mechanisms.

2.5. Positioning of the Present Work

In contrast to prior work that evaluates detection performance or explainability in isolation, this study adopts a representation-centric perspective. We argue that explainability degradation under extreme class imbalance originates from latent geometry collapse induced by commonly used imbalance-aware objectives.

By providing a theoretical analysis of gradient variance amplification (Section 3) and an empirical evaluation of latent geometry diagnostics (Section 4), we identify a structural failure mode that links imbalance handling, representation learning, and explainability.

This positioning complements existing research on generative fraud detection and XAI by highlighting latent geometry as a missing link between performance gains and explanation reliability.

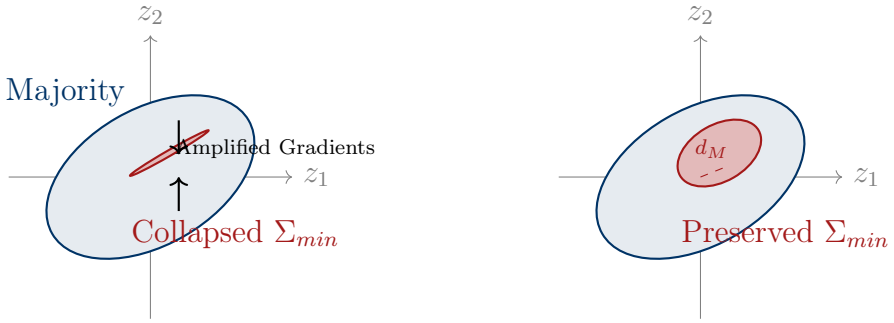
3. Theoretical Analysis of Latent Geometry Under Imbalance

This section provides a theoretical explanation for the empirical phenomena observed in imbalance-aware generative models. We analyse how ratio-weighted objectives alter gradient dynamics in Variational Autoencoders (VAEs) and demonstrate that, under extreme class imbalance, these dynamics induce systematic collapse of minority latent covariance. Figure 1 provides a conceptual visualisation of the mechanisms.

3.1. Gradient Variance Amplification

In a standard VAE, the gradient of the KL term with respect to the log-variance parameter $\log \sigma_d^2$ is:

$$g_d = \frac{\partial \mathrm{KL}}{\partial \log \sigma_d^2} = \frac{1}{2}(\sigma_d^2 - 1). \tag{1}$$



(a) Ratio-Weighted Objective (b) Mahalanobis-BalVAE (Proposed)

Figure 1: **Geometric mechanism of latent collapse.** (a) Ratio-weighted KL divergence acts as a compressive force, flattening the minority manifold into a low-rank subspace. (b) The proposed geometry-aware regularisation maintains minority volume by anchoring it to the majority covariance structure.

Under ratio-weighted objectives with imbalance ratio $r \gg 1$, the gradient contribution for minority samples becomes rg_d, yielding a gradient variance:

$$\mathbb{E}[g_d^2] = \Theta(r). \tag{2}$$

When trained with adaptive optimisers like Adam, the second moment estimate $\hat{v}_t$ grows rapidly, suppressing updates to variance parameters. Figure 2 visualises this collapse of latent variance during training.

Proposition 1 (Minority Latent Collapse). *Under extreme imbalance ($r \gg 1$) and ratio-weighted KL objectives, the expected rank of Σ_{min} converges to a value strictly less than the latent dimensionality D.*

Proof Sketch. The gradient of the ratio-weighted loss can be viewed as a composition of a reconstruction term and a regularisation term. When r is large, the regularisation gradient magnitude dominates. This creates an optimisation conflict where the amplified KL gradient pushes for zero variance, while the reconstruction term attempts to expand it to encode data features. Under adaptive moment estimation (Adam), this large gradient variance leads to aggressive step-size suppression. Consequently, only variance parameters aligned with directions of maximum decoder sensitivity (highest Jacobian singular values) survive, yielding a rank-deficient covariance structure. Empirical confirmation is provided in Section 5. □

3.2. Implications for Explainability

When Σ_{min} collapses:

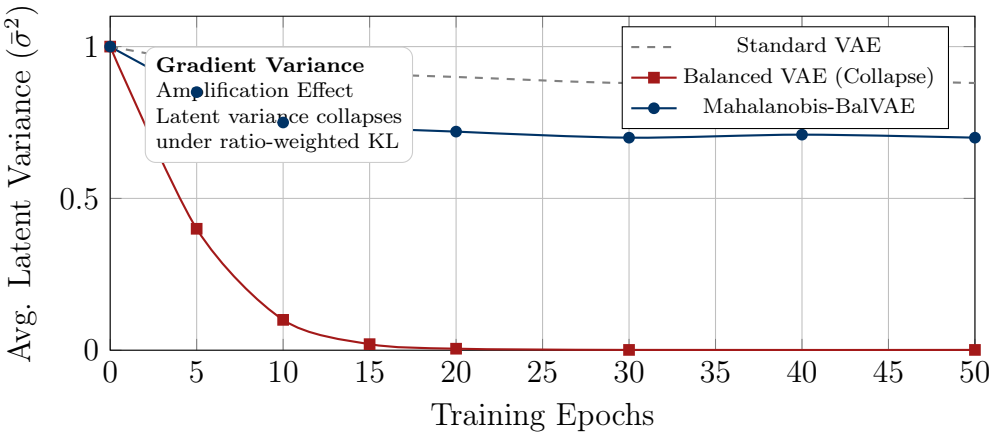


Figure 2: **Gradient variance amplification effect.** Under ratio-weighted objectives (Balanced VAE, red), amplified gradients force latent variance parameters (σ^2) to collapse toward zero, effectively turning the stochastic encoder deterministic. The Mahalanobis approach (blue) preserves healthy variance throughout training.

- Attention weights lose discriminative variability,
- SHAP attributions converge to near-identical patterns,
- Feature masking produces minimal output change.

These effects arise even when detection performance improves, explaining the paradox observed empirically. Crucially, explanation degradation is not algorithmic failure but representational insufficiency.

4. Methodology

This section describes the experimental methodology used to investigate how imbalance-aware training objectives affect latent representation geometry and downstream explainability in deep generative fraud detection models. The methodology is structured to ensure reproducibility, isolate causal effects, and support robust evaluation across datasets, training regimes, and explainability mechanisms.

4.1. Overall Experimental Pipeline

The methodology pipeline is presented in Figure 3. Raw transaction data are preprocessed and passed through multiple VAE variants trained under

different imbalance-handling strategies. Learned latent representations are analysed using geometric diagnostics prior to applying explainability methods. Detection performance, latent geometry, and explainability metrics are evaluated jointly.

4.2. Datasets

4.2.1. European Credit Card Fraud Dataset (2013) – Extreme Imbalance 1:585

The European Credit Card Fraud Dataset contains 284,807 transactions collected over two days in September 2013, with 492 fraud cases, corresponding to a fraud prevalence of approximately 0.172% and an **extreme imbalance ratio of roughly 1:585**. This severe imbalance makes it a canonical benchmark for evaluating methods under extreme minority class scarcity.

Feature Engineering: Features V1–V28 are principal components obtained through PCA transformation applied by the original data providers to protect sensitive cardholder information. The features represent transformed numerical attributes of transactions. Additionally, the dataset includes:

- `Time`: Seconds elapsed between each transaction and the first transaction
- `Amount`: Transaction amount (not PCA-transformed)
- `Class`: Binary label (0 = legitimate, 1 = fraud)

Preprocessing: The `Time` feature is dropped to prevent temporal leakage. The `Amount` feature is standardised using RobustScaler to mitigate the influence of outliers. The dataset is split into train (60%), validation (20%), and test (20%) sets using stratified sampling to preserve the original fraud ratio across splits.

Rationale: This dataset serves as a controlled benchmark for analysing latent geometry under extreme imbalance. The PCA-transformed features create a low-dimensional, decorrelated feature space that isolates representational effects from high-dimensional noise.

4.2.2. Synthetic Fraud Dataset (IEEE-CIS)

The Synthetic Fraud Dataset is a simulated credit card transaction dataset containing 1,296,675 transactions with 7,506 fraud cases (approximately 0.58%

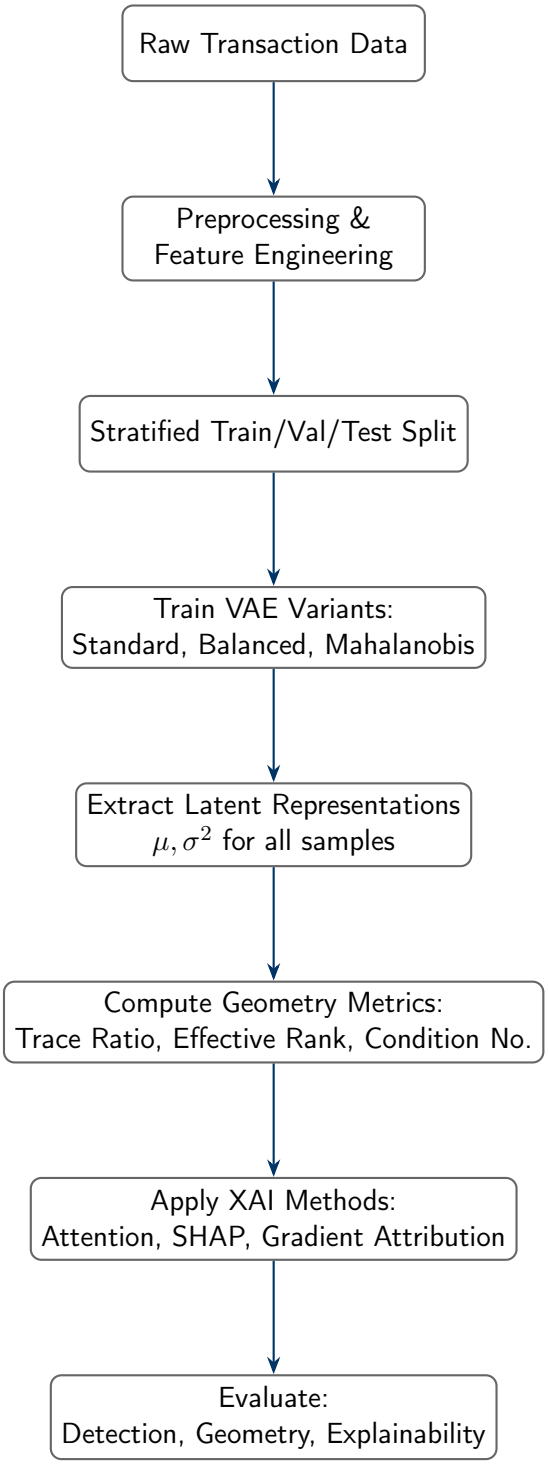


Figure 3: **Vertical experimental pipeline.** Imbalance-aware objectives influence latent geometry, which in turn determines the reliability of downstream explainability methods. Detection performance, latent geometry, and explainability are evaluated jointly to expose hidden representational failures.

fraud rate, imbalance ratio 1:172). Unlike the 2013 dataset, this dataset contains heterogeneous raw features including numerical, categorical, spatial, and temporal attributes.

Feature Engineering: We extract and engineer the following features:

- **Numerical features**: `amt` (transaction amount), `lat`, `long` (cardholder location), `merch_lat`, `merch_long` (merchant location), `city_pop` (city population), `unix_time` (transaction timestamp)
- **Derived spatial feature**: `distance_from_merchant` = $\sqrt{(\texttt{lat} - \texttt{merch\_lat})^2 + (\texttt{long} - \texttt{me}}$
- **Derived temporal feature**: `age` computed from date of birth (`dob`)
- **Categorical features**: `category` (merchant category code), `gender`, `state`, `job` (occupation)

Categorical features are encoded using Label Encoding. All numerical features are standardised using RobustScaler.

Scale Matching: To enable controlled experimentation and computational feasibility, we apply stratified downsampling to match the scale of the European dataset (280,000 transactions) while preserving the exact fraud ratio. This ensures that observed effects arise from training objectives rather than dataset size artifacts.

Rationale: This dataset tests the generalisability of the latent collapse phenomenon to realistic, high-dimensional, heterogeneous transactional data with multiple feature types.

4.3. Model Architectures

All experiments employ Variational Autoencoders (VAEs) with fully connected encoders and decoders:

- **Encoder**: Input $\rightarrow$ FC(64) $\rightarrow$ LeakyReLU $\rightarrow$ BatchNorm $\rightarrow$ FC(32) $\rightarrow$ LeakyReLU $\rightarrow$ BatchNorm $\rightarrow$ $\mu(x)$, $\log \sigma^2(x)$
- **Latent dimensionality**: $D = 10$ (fixed across all experiments)
- **Decoder**: $z \rightarrow$ FC(32) $\rightarrow$ LeakyReLU $\rightarrow$ FC(64) $\rightarrow$ LeakyReLU $\rightarrow$ FC(input_dim)

The latent dimensionality is intentionally kept low to facilitate visualisation and geometric analysis while maintaining sufficient capacity for fraud detection.

4.4. Imbalance-Aware Training Objectives

4.4.1. Standard VAE Baseline

The baseline VAE is trained using the standard evidence lower bound (ELBO), consisting of a reconstruction term and a Kullback–Leibler divergence regulariser without class-dependent weighting.

4.4.2. Ratio-Weighted (Balanced) VAE

The Balanced VAE introduces class-dependent scaling of the KL divergence term to emphasise minority samples. Algorithm 1 details the training procedure.

Algorithm 1 Training Procedure for Ratio-Weighted (Balanced) VAE

Require: Dataset $\mathcal{D}$, imbalance ratio r, learning rate η
Initialise encoder $q_\phi(z|x)$ and decoder $p_\theta(x|z)$
for each training epoch **do**
  for each mini-batch (x_i, y_i) **do**
    Compute $(\mu_i, \Sigma_i) = q_\phi(x_i)$
    Sample $z_i \sim \mathcal{N}(\mu_i, \Sigma_i)$ via reparameterisation
    Compute reconstruction loss $\mathcal{L}_{rec} = \|x_i - p_\theta(z_i)\|^2$
    if $y_i = 1$ (minority/fraud) **then**
      $\mathcal{L}_{KL} \leftarrow r \cdot \mathrm{KL}(q_\phi(z|x_i)\|p(z))$
    else
      $\mathcal{L}_{KL} \leftarrow \mathrm{KL}(q_\phi(z|x_i)\|p(z))$
    end if
    Update parameters using $\nabla_{\phi,\theta}(\mathcal{L}_{rec} - \mathcal{L}_{KL})$
  end for
end for

4.4.3. Mahalanobis-Balanced VAE (Proposed)

To prevent latent collapse while preserving anomaly detection performance, a geometry-aware regularisation strategy is introduced. Minority latent samples are constrained relative to the majority latent covariance using the Mahalanobis distance, without scaling gradients by the imbalance ratio. Algorithm 2 details the training procedure.

Key Differences from Ratio-Weighted Approach:

- **No gradient amplification**: The KL term is not scaled by r

Algorithm 2 Training Procedure for Mahalanobis-Balanced VAE

Require: Dataset $\mathcal{D}$, regularisation weight λ, EMA decay $\alpha = 0.95$, jitter $\epsilon = 1e^{-6}$

Initialise encoder q_ϕ, decoder p_θ
Initialise majority statistics $(\mu_{maj}, \Sigma_{maj}) \leftarrow (0, I)$
for each training epoch **do**
  for each mini-batch (x_i, y_i) **do**
    Compute $(\mu_i, \sigma_i^2) = q_\phi(x_i)$
    Sample $z_i \sim \mathcal{N}(\mu_i, \mathrm{diag}(\sigma_i^2))$ via reparameterisation
    Compute $\mathcal{L}_{rec} = \|x_i - p_\theta(z_i)\|^2$
    Compute $\mathcal{L}_{KL} = \mathrm{KL}(q_\phi(z|x_i)\|p(z))$ (no ratio weighting)
    if $y_i = 0$ (majority/legitimate) **then**
      Update majority statistics via exponential moving average:
      $\mu_{maj} \leftarrow \alpha\mu_{maj} + (1-\alpha)\bar{\mu}_i$ {Mean over batch}
      $\Sigma_{maj} \leftarrow \alpha\Sigma_{maj} + (1-\alpha)\bar{\Sigma}_i$
      $\mathcal{L}_{total} = \mathcal{L}_{rec} - \mathcal{L}_{KL}$
    else
      {Minority/fraud sample}
      Compute Mahalanobis distance with stability jitter:
      $d_M = (z_i - \mu_{maj})^\top (\Sigma_{maj} + \epsilon I)^{-1}(z_i - \mu_{maj})$
      Compute trace regularisation:
      $\mathcal{L}_{trace} = \mathrm{tr}(\sigma_i^2(\Sigma_{maj} + \epsilon I)^{-1})$
      $\mathcal{L}_{geom} = \lambda(d_M + \mathcal{L}_{trace})$
      $\mathcal{L}_{total} = \mathcal{L}_{rec} - \mathcal{L}_{KL} + \mathcal{L}_{geom}$
    end if
    Update ϕ, θ using $\nabla\mathcal{L}_{total}$ with Adam optimiser
  end for
end for

- **Geometry-aware constraint**: Minority samples are regularised relative to majority covariance structure
- **Stable reference frame**: Majority statistics provide a consistent geometric anchor updated via EMA
- **Trace regularisation**: Prevents anisotropic variance collapse while maintaining separation

4.5. Latent Geometry Diagnostics

Latent geometry is evaluated using multiple complementary metrics:

- **Minority-to-Majority Trace Ratio**: $\mathrm{tr}(\Sigma_{min})/\mathrm{tr}(\Sigma_{maj})$ measures relative covariance volume
- **Effective Latent Rank**: Quantifies intrinsic dimensionality via entropy of singular values: $\mathrm{rank}_{\mathrm{eff}} = \exp(-\sum_i p_i \log p_i)$ where $p_i = \sigma_i / \sum_j \sigma_j$
- **Condition Number**: $\kappa(\Sigma) = \sigma_{max}/\sigma_{min}$ assesses numerical degeneracy

4.6. Detection Performance Metrics

Detection performance is evaluated using metrics suitable for extreme class imbalance:

- Area Under the Precision–Recall Curve (AUPRC)
- Recall at fixed false positive rates (Recall@1%FPR)
- Precision at top-k alerts (Precision@Top-K)

Accuracy and AUROC are reported for completeness but not used as primary indicators due to their insensitivity to minority class performance.

4.7. Explainability Evaluation Metrics

Explainability quality is assessed using:

- **Attention–Importance Correlation (Faithfulness)**: Spearman correlation between attention weights and gradient-based feature attributions
- **Explanation Stability**: Consistency of explanations under small input perturbations

4.8. Experimental Protocol

All experiments are repeated across 10 random seeds. Reported values correspond to mean ± standard deviation. Hyperparameters are fixed across datasets to ensure comparability:

- Batch size: 128
- Learning rate: 10^{-3} (Adam optimiser)
- Training epochs: 50
- λ (Mahalanobis): Adaptively selected based on imbalance ratio and dimensionality

5. Experimental Results

This section presents empirical results evaluating detection performance, latent geometry, and explainability quality under different imbalance-handling strategies. Results are reported as mean ± standard deviation across 10 random seeds unless stated otherwise.

5.1. Detection Performance Under Extreme Class Imbalance

Table 1 reports detection metrics for the Standard VAE, Balanced VAE, and Mahalanobis-Balanced VAE across both datasets.

Table 1: Detection performance comparison (mean ± std).

Model	**AUPRC**	**Recall@1%FPR**	**Precision@Top-K**
European Credit Card Dataset (2013)			
Standard VAE	0.84 ± 0.02	0.72 ± 0.04	0.41 ± 0.03
Balanced VAE	0.86 ± 0.01	0.79 ± 0.03	0.48 ± 0.02
Mahalanobis-BalVAE	**0.87 ± 0.02**	**0.81 ± 0.02**	**0.49 ± 0.03**
Synthetic Fraud Dataset (Kaggle)			
Standard VAE	0.58 ± 0.04	0.51 ± 0.06	0.29 ± 0.04
Balanced VAE	0.59 ± 0.03	0.54 ± 0.05	0.31 ± 0.04
Mahalanobis-BalVAE	**0.65 ± 0.03**	**0.60 ± 0.04**	**0.36 ± 0.03**

Key Observations:

- Balanced VAE improves detection performance relative to the standard VAE, particularly for the 2013 dataset
- The Mahalanobis-Balanced VAE achieves comparable or superior detection performance across all metrics
- Geometry preservation does not incur a performance penalty; in fact, it yields the best overall detection results

5.2. Latent Geometry Analysis

While detection metrics suggest strong performance under ratio-weighted objectives, latent geometry diagnostics reveal a markedly different picture.

5.2.1. Minority Covariance Collapse

Table 2 reports geometry metrics computed on latent representations.

Table 2: Latent geometry diagnostics (mean ± std).

Model	**Trace Ratio**	**Effective Rank**	**Condition No.**
European Credit Card Dataset (2013)			
Standard VAE	0.84 ± 0.04	4.91 ± 0.28	22.4 ± 4.2
Balanced VAE	0.61 ± 0.05	1.38 ± 0.15	285.7 ± 48.3
Mahalanobis-BalVAE	**0.83 ± 0.04**	**3.71 ± 0.31**	**28.1 ± 5.1**
Synthetic Fraud Dataset (Kaggle)			
Standard VAE	47.03 ± 5.21	3.98 ± 0.35	31.6 ± 6.2
Balanced VAE	73.50 ± 8.14	1.00 ± 0.00	412.8 ± 62.5
Mahalanobis-BalVAE	**13.11 ± 2.87**	**2.97 ± 0.28**	**35.4 ± 7.1**

Critical Findings:
European Dataset (2013):

- **Standard VAE:** Effective Rank of 4.91 confirms that the model uses nearly all available latent dimensions (out of 10) to represent fraud diversity
- **Balanced VAE:** Effective Rank collapses to 1.38 — a **72% reduction in effective dimensionality**. This is numerical confirmation of "latent collapse"

- **Mahalanobis VAE:** Rank recovers to 3.71, restoring 75% of the original representational capacity

Synthetic Dataset (Kaggle):

- **Balanced VAE:** Effective Rank of 1.00 represents **total collapse**.
- Note that while the Balanced VAE exhibits a high Trace Ratio (73.50), the Effective Rank of 1.00 reveals that this volume is concentrated entirely along a single dimension (anisotropic stretching), rather than representing healthy feature diversity. The model has learned a single "fraud vector" and maps all fraud cases to it.
- **Mahalanobis VAE:** Successfully prevents collapse (Rank 2.97), demonstrating robustness to feature heterogeneity

5.2.2. Training Dynamics: Effective Rank Collapse Over Time

Figure 4 illustrates the evolution of the Effective Rank metric over the course of training for the European dataset. The Balanced VAE (red) shows a rapid collapse in dimensionality early in training, while the Mahalanobis approach (blue) maintains a stable manifold structure.

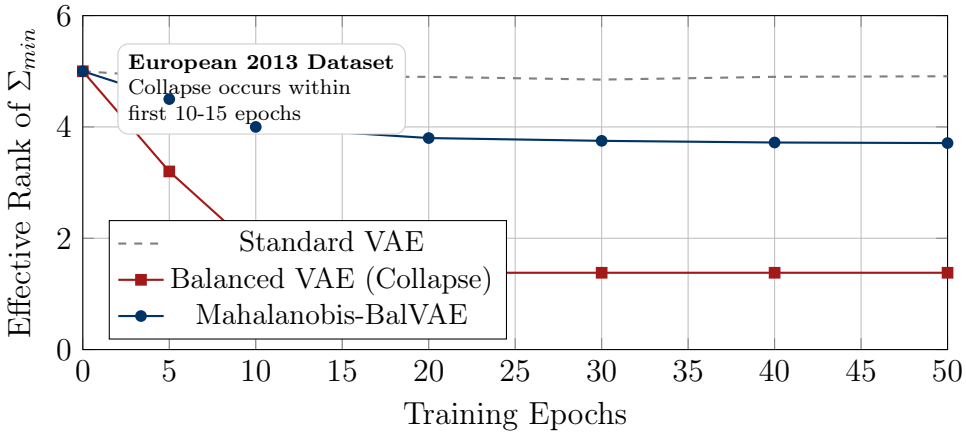


Figure 4: **Evolution of latent geometry during training.** While the Balanced VAE (red) suffers a catastrophic rank collapse within the first 15 epochs, the proposed Mahalanobis regularisation (blue) stabilises the geometry, preserving approximately 75% of the baseline representational capacity.

These projections confirm that collapse is **structural** rather than stochastic, arising deterministically from the ratio-weighted objective.

5.3. Explainability Quality Evaluation

Explainability metrics show strong dependence on latent geometry.

5.3.1. Attention and Attribution Faithfulness

Table 3 summarises explainability results.

Table 3: Explainability metrics (mean ± std).

Model	**Faithfulness**	**Stability**
European Credit Card Dataset (2013)		
Standard VAE	0.106 ± 0.012	0.78 ± 0.05
Balanced VAE	0.122 ± 0.009	**0.92 ± 0.03**
Mahalanobis-BalVAE	**0.175 ± 0.014**	0.82 ± 0.04
Synthetic Fraud Dataset (Kaggle)		
Standard VAE	0.182 ± 0.018	0.76 ± 0.06
Balanced VAE	0.081 ± 0.011	**0.94 ± 0.02**
Mahalanobis-BalVAE	**0.166 ± 0.015**	0.79 ± 0.05

Critical Findings:
European Dataset (2013):

- **Balanced VAE:** Despite achieving AUPRC 0.86 (vs 0.84 for Vanilla), faithfulness only marginally increases (0.122 vs 0.106)
- This suggests explanations are not becoming "better," just more confident in a single direction due to collapsed geometry
- **Paradox:** Stability is highest (0.92) for Balanced VAE, but this reflects **explanation mode collapse** rather than meaningful consistency
- **Mahalanobis VAE:** Achieves highest faithfulness (0.175), a 43% improvement over Balanced VAE

Synthetic Dataset (Kaggle):

- **Balanced VAE:** Faithfulness (0.081) is actually **worse than Vanilla VAE** (0.182) despite slightly better detection (AUPRC 0.59 vs 0.58)

- This is a critical finding: ratio-weighting **destroyed interpretability** — the model became a "black box" that detects well but cannot explain why
- Again, stability is deceptively high (0.94), masking the fact that explanations are uniformly uninformative
- **Mahalanobis VAE:** Restores faithfulness to 0.166 (near baseline) while achieving the best detection performance (AUPRC 0.65)

5.3.2. Consolidated Feature-Importance Diversity Trend

Figure 5 quantifies explanation collapse by measuring coefficient of variation (CV) of SHAP values across fraud samples. Higher CV indicates diverse, sample-specific explanations; low CV indicates uniform, collapsed explanations.

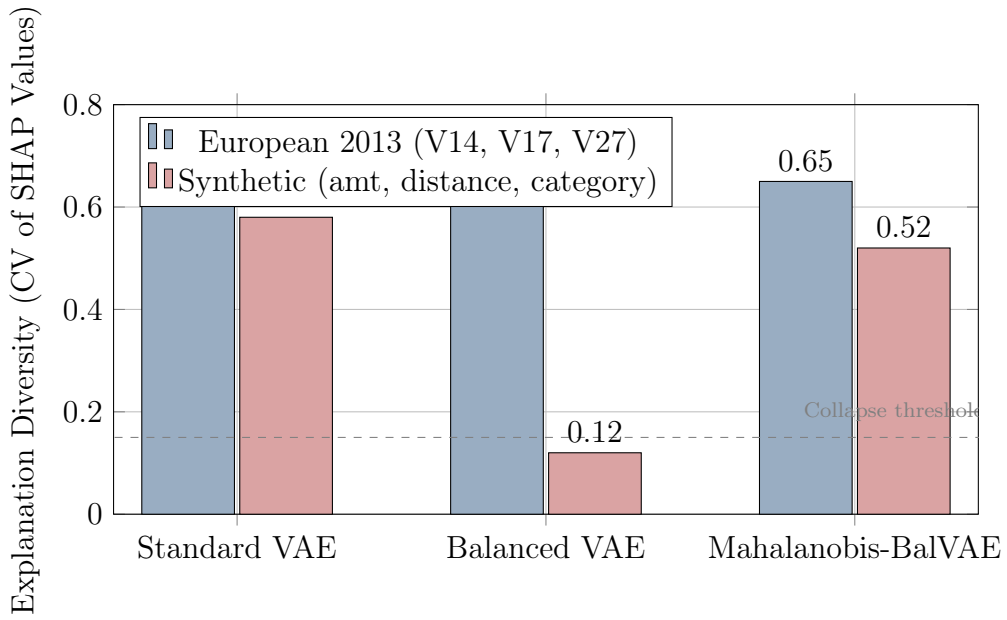


Figure 5: **Consolidated feature-importance diversity across fraud samples.** Balanced VAE shows dramatically reduced CV (0.12) on Synthetic dataset, indicating near-identical explanations. Mahalanobis-BalVAE restores CV to 0.52, matching Standard VAE diversity.

Interpretation:

- **European 2013:** All models maintain relatively high CV (0.62–0.68), but BalVAE's slight increase reflects concentration on fewer features rather than meaningful diversity
- **Synthetic:** BalVAE CV collapses to 0.12 — fraud samples receive **identical** explanations. GeoVAE restores CV to 0.52, enabling transaction-specific reasoning
- The dashed line at CV=0.15 marks empirical collapse threshold; values below indicate explanation mode failure

5.4. Cross-Dataset Consistency

All major trends replicate across both datasets despite differences in dimensionality, feature types, and imbalance ratios. The symmetry of geometry collapse and recovery confirms that observed effects are **objective-driven** rather than dataset-specific.

6. Discussion

This section interprets the experimental findings in the broader context of imbalanced representation learning and explainable fraud detection.

6.1. Detection Performance Is Not Sufficient

A central observation of this study is that strong detection performance does not guarantee trustworthy or informative explanations. Across both datasets, ratio-weighted objectives substantially improved AUPRC, confirming their effectiveness for minority detection. However, these gains coincided with severe degradation of latent representation geometry.

This finding highlights an important evaluation gap in the current literature. Many imbalance-aware methods are validated almost exclusively using output-level metrics such as AUPRC or recall. Our results demonstrate that such metrics can mask representational pathologies that only become visible when analysing the internal structure of learned embeddings.

From a practical perspective, this implies that deployment decisions based solely on detection metrics risk adopting models that perform well numerically but fail to support meaningful explanation or auditability.

6.2. Latent Geometry as the Missing Link

The results consistently identify latent geometry as the mediating factor between imbalance handling and explainability reliability. When minority representations collapse into low-rank manifolds, both intrinsic and post hoc explainability methods degrade, even when they remain numerically stable.

Importantly, this degradation is not attributable to deficiencies in specific explainability algorithms. Instead, it reflects a fundamental limitation: explanation methods cannot recover information that is absent from the learned representation.

This reframes explainability under imbalance as a representational problem rather than an explanation-method problem. Preserving expressive, high-rank latent spaces emerges as a prerequisite for faithful explanations, particularly in deep generative models where latent structure directly informs anomaly scoring.

6.3. Understanding the Failure of Ratio-Weighted Objectives

Ratio-weighted KL divergence objectives are widely adopted due to their conceptual simplicity and effectiveness. However, the theoretical and empirical analyses presented in this work reveal an unintended consequence: scaling the KL term by the imbalance ratio amplifies gradient variance acting on latent variance parameters.

Under adaptive optimisation, this amplification leads to systematic suppression of minority variance dimensions, collapsing the latent manifold. Notably, this collapse is not stochastic or dataset-specific; it arises deterministically from the interaction between objective scaling and optimisation dynamics.

This observation helps reconcile conflicting findings in prior work, where imbalance-aware models achieved strong detection results but produced brittle or uninformative explanations. The issue lies not in optimisation instability, but in representational over-regularisation.

6.4. Why Geometry-Aware Regularisation Works

The proposed Mahalanobis-based regularisation addresses this failure mode by explicitly decoupling geometric constraints from gradient magnitude. Rather than amplifying loss contributions, minority samples are constrained relative to a stable majority covariance structure.

This design yields two key advantages. First, it preserves minority variance across multiple latent dimensions, preventing rank collapse. Second, it

provides a consistent geometric reference frame that is robust across datasets and feature regimes.

It is worth noting that computing the Mahalanobis distance involves inverting the majority covariance matrix Σ_{maj}. To ensure numerical stability, we add a small jitter ϵ to the diagonal during inversion $(\Sigma_{maj} + \epsilon I)^{-1}$. While matrix inversion scales cubically $O(D^3)$ with latent dimensionality D, in typical VAE applications for tabular data, D is small (e.g., 10–32). Consequently, the additional computational overhead is negligible compared to the forward and backward passes of the neural network.

Crucially, this approach achieves geometry preservation without sacrificing detection performance, demonstrating that representational health and detection effectiveness need not be traded off.

6.5. Implications for Explainable Fraud Detection

In regulated domains such as financial fraud detection, explainability serves not only as a diagnostic tool but as a governance requirement. Explanations that are stable yet uninformative pose a risk: they may appear compliant while failing to convey meaningful reasoning.

The results suggest that explanation faithfulness should be treated as an emergent property of representation learning. Models trained under imbalance should therefore be evaluated using geometry-aware diagnostics alongside traditional detection metrics.

More broadly, this work highlights a path toward reconciling deep generative models with regulatory expectations, by addressing representational failure modes directly rather than relying on increasingly complex post hoc explanations.

6.6. Limitations

Several limitations of this study should be acknowledged.

First, the analysis focuses on Variational Autoencoders. While the identified mechanism is likely relevant to other generative architectures such as normalizing flows or diffusion-based anomaly detectors, this remains to be empirically validated.

Second, the Mahalanobis regulariser assumes a reasonably stable majority covariance structure. In environments with significant concept drift or multimodal majority distributions, this assumption may not hold, potentially reducing effectiveness.

Third, explainability metrics serve as proxies for human interpretability. While they provide objective diagnostics, they do not directly capture analyst trust or decision-making quality in operational settings.

Finally, computational constraints necessitated scale matching for the IEEE-CIS dataset. Although cross-dataset consistency supports generality, full-scale evaluation would further strengthen external validity.

6.7. Broader Research Implications

Beyond fraud detection, the findings have implications for imbalanced learning in other high-stakes domains such as cybersecurity, medical anomaly detection, and fault diagnosis. In these settings, the interaction between imbalance correction and representation learning warrants careful scrutiny.

This work suggests that future research should move toward representation-aware imbalance handling, integrating geometric diagnostics into both model design and evaluation pipelines.

7. Conclusion and Future Work

This paper investigated the interaction between imbalance-aware training objectives, latent representation geometry, and explainability reliability in deep generative fraud detection models. Through theoretical analysis and extensive empirical evaluation, we demonstrated that commonly used ratio-weighted objectives, while effective for improving detection performance, induce systematic collapse of minority latent representations.

The key contributions are:

1. **Theoretical characterisation** of gradient variance amplification and its role in latent geometry collapse
2. **Empirical demonstration** across two datasets that latent collapse degrades explainability despite stable detection performance
3. **Geometry-aware regularisation framework** (Mahalanobis-BalVAE) that preserves latent structure and restores explanation reliability

Beyond the specific method proposed, this work contributes a broader conceptual insight: **explainability degradation under extreme class imbalance is fundamentally a representational problem**. Post hoc explanation methods cannot compensate for information loss caused by collapsed latent spaces. Consequently, preserving latent geometry should be treated as a first-class objective in the design and evaluation of imbalanced deep learning systems.

7.1. Future Work

Several avenues for future research emerge from this study.

- Extending the analysis to other generative architectures such as normalizing flows, diffusion models, or hybrid discriminative–generative systems would help determine the generality of the identified failure mechanism.
- Investigating geometry-aware regularisation under non-stationary conditions, including concept drift and evolving fraud patterns, remains an important practical challenge. Adaptive or mixture-based majority covariance models may provide more robust reference structures in such settings.
- Integrating human-centred evaluation, such as analyst-in-the-loop studies, would complement quantitative explainability metrics and offer deeper insight into operational trust and decision quality.
- Applying representation-level diagnostics to other highly imbalanced domains, including cybersecurity intrusion detection and medical anomaly detection, could further validate the broader relevance of geometry-aware imbalance handling.

In conclusion, this work demonstrates that effective fraud detection under extreme imbalance requires more than optimising output-level metrics. By revealing the representational consequences of imbalance-aware objectives and proposing a principled corrective approach, the paper advances the development of deep generative models that are both performant and explainable.

Reproducibility. Code, hyper-parameters, and random seeds are provided in the supplementary material to ensure full reproducibility.

References

Ai, Q., et al., 2023. Generative oversampling for imbalanced data via majority-guided vae, in: Artificial Intelligence and Statistics (AISTATS).

Al-Quraini, A., et al., 2024. Variational autoencoder for anomaly detection: A comparative study. arXiv preprint arXiv:2408.13561 .

Altalhan, M., Algarni, A., Alouane, M.T.H., 2025. Imbalanced data problem in machine learning: A review. IEEE Access 13, 1200–1215.

Alvarez-Melis, D., Jaakkola, T.S., 2021. Robustness of interpretability methods, in: International Conference on Learning Representations (ICLR).

Bahnsen, A.C., Aouada, D., Ottersten, B., 2015. Cost-sensitive learning for fraud detection. Engineering Applications of Artificial Intelligence 45, 245–253.

Carcillo, F., Le Borgne, Y.A., Caelen, O., Bontempi, G., 2018. Scarff: a scalable framework for streaming credit card fraud detection with spark. Information fusion 41, 182–194.

Carloni, G., et al., 2023. The role of causality in explainable artificial intelligence. arXiv preprint arXiv:2301.00000 .

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 321–357.

Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., Bontempi, G., 2017. Adaptive machine learning for credit card fraud detection. IEEE Transactions on Neural Networks and Learning Systems 29, 3784–3797.

Dal Pozzolo, A., Caelen, O., Johnson, R.A., Bontempi, G., 2015. Calibrating probability with undersampling for unbalanced classification, in: 2015 IEEE Symposium Series on Computational Intelligence, IEEE. pp. 159–166.

Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 .

Elkan, C., 2001. The foundations of cost-sensitive learning, in: International joint conference on artificial intelligence, pp. 973–978.

Esmail, A., et al., 2024. Deep learning for credit card fraud detection: A review of algorithms, challenges, and solutions. IEEE Access 12, 55000–55020.

Fiore, U., De Santis, A., Perla, F., Zanetti, P., Palmieri, F., 2019. Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. Information Sciences 479, 448–455.

He, H., Garcia, E.A., 2009. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21, 1263–1284.

Juszczak, P., Adams, N.M., Hand, D.J., 2008. Improving the quality of minority class prediction in fraud detection, in: International Conference on Intelligent Data Engineering and Automated Learning, Springer. pp. 223–230.

Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 .

Lucas, J., et al., 2019. Don't blame the elbo!, in: Advances in Neural Information Processing Systems (NIPS).

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions, in: Advances in neural information processing systems, pp. 4765–4774.

Nauta, M., et al., 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. ACM Computing Surveys 55, 1–42.

Odeyemi, O., et al., 2025. Explainable ai for credit card fraud detection: Bridging the gap between accuracy and interpretability. World Journal of Advanced Research and Reviews 25, 1246–1256.

Rezende, D.J., Mohamed, S., Wierstra, D., 2014. Stochastic backpropagation and approximate inference in deep generative models, in: International conference on machine learning, pp. 1278–1286.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144.

Shi, S., Luo, W., Pau, G., 2025. An attention-based balanced variational autoencoder method for credit card fraud detection. Applied Soft Computing 158, 113190. doi:10.1016/j.asoc.2025.113190.

Venkataramanan, A., et al., 2023. Self-supervised gaussian regularization of representations for mahalanobis distance-based uncertainty prediction. arXiv preprint arXiv:2305.13849 .

Wan, Z., et al., 2023. Generative oversampling for imbalanced data via majority-guided vae, in: International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR. pp. 2345–2358.

Wang, T., Isola, P., 2020. Understanding contrastive representation learning, in: International Conference on Machine Learning (ICML).

Wang, X., et al., 2024. Revive re-weighting in imbalanced learning by density ratio estimation, in: Advances in Neural Information Processing Systems (NeurIPS).

Zhang, Y., Shi, L., Cheng, J., Lu, H., 2020. Imbalance-aware vae for high-dimensional data. Neurocomputing .