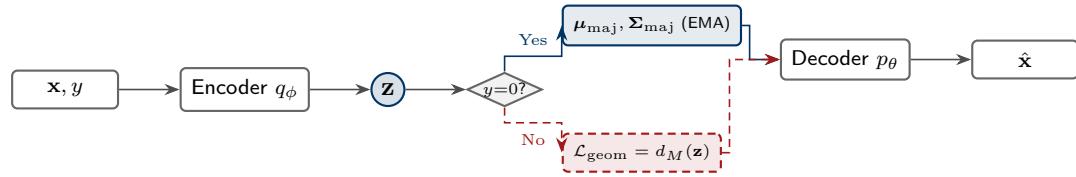


# Graphical Abstract

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

Minenhle Mpulo



## Highlights

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

Minenhle Mpulo

- Ratio-weighted KL in VAEs collapses minority latent geometry under extreme imbalance (effective rank  $\approx 1.0\text{--}2.2$ ).
- Collapse reduces explanation diversity, with Balanced VAE showing 18% reduction on European dataset.
- Mahalanobis-based regularisation restores geometry (rank  $\approx 2.2\text{--}4.4$ ) without sacrificing AUPRC.
- Geometry-aware training consistently achieves highest explanation diversity across datasets.

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

Minenhle Mpulo

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

---

## Abstract

Deep generative models have demonstrated strong anomaly detection performance under extreme class imbalance, where fraudulent transactions typically constitute less than 0.2% of observed data [1, 2]. Recent work introduces ratio-weighted training objectives to amplify minority-class contributions during optimisation, consistently improving recall and Area Under the Precision–Recall Curve (AUPRC) for Credit Card Fraud Datasets [3, 4]. However, the impact of such objectives on learned latent representations and downstream explainability remains unexplored [5].

This work shows that ratio-weighted objectives in Variational Autoencoders (VAEs) induce systematic latent geometry collapse under extreme imbalance. Specifically, scaling the Kullback–Leibler divergence by the imbalance ratio interacts with adaptive optimization to suppress latent variance parameters, concentrating minority representations into low-dimensional subspaces. Across both European and IEEE-CIS datasets with imbalance ratios ranging from 172:1 to 578:1, ratio-weighted VAEs reduce the effective latent rank by up to 70% relative to standard baselines, despite competitive detection performance ( $\text{AUPRC} \approx 0.64\text{--}0.83$ ). Our analysis demonstrates that this representational collapse directly impacts explainability. Evaluating explanation quality using faithfulness, stability, and diversity metrics adapted from [5], we find that geometry collapse reduces explanation diversity as fraud instances receive increasingly similar attributions. On the European dataset, the ratio-weighted VAE reduces diversity by 18% compared to the standard baseline, while on the Synthetic dataset, effective rank collapses to near-unity despite competitive detection.

To address this, we propose a geometry-aware regularization strategy based on Mahalanobis distance to the majority latent covariance. By constraining minority representations relative to a stable reference manifold rather than scaling loss terms, this approach preserves latent dimensionality (2.17–4.41) without amplifying gradients. Experiments demonstrate that geometry-aware training restores representational structure and consistently achieves the highest explanation diversity across both datasets while maintaining detection performance, establishing that imbalance handling and explainability need not be competing objectives.

*Keywords:* Class imbalance, Variational Autoencoders, Explainable AI, Latent geometry, Fraud detection, Deep generative models, Representation collapse, Geometry-aware regularization, Credit Card Fraud

---

## 1. Introduction

Credit card fraud detection operates under an extreme class imbalance, with fraudulent transactions typically forming less than 0.2% of observed data [1, 6], yet it is heavily dependent on this minority data to precisely detect fraud. This imbalance is not merely a statistical inconvenience; it fundamentally shapes how models learn, what representations they form, and ultimately, how reliably they can be explained. In regulated financial environments, where explainability is a compliance requirement rather than an optional feature, understanding these interactions is essential [7].

The challenge of learning under such imbalance has driven substantial methodological innovation. Early approaches relied on data-level resampling and cost-sensitive learning [8, 9, 10]. More recently, deep generative models, particularly Variational Autoencoders (VAEs), have emerged as effective tools for fraud detection, modelling legitimate transaction distributions and identifying anomalies through reconstruction error [11, 12]. To further improve minority detection, researchers have introduced ratio-weighted training objectives that scale the Kullback–Leibler (KL) regularisation term by the class imbalance ratio [4, 3]. These approaches consistently improve detection metrics such as recall and Area Under the Precision–Recall Curve (AUPRC).

However, a critical question remains unexamined: what happens to the internal representations when we apply such aggressive imbalance correction? Existing evaluations focus almost exclusively on output-level detection performance, implicitly assuming that strong detection results reflect meaningful representations [13, 14]. This assumption is problematic since latent representations underpin not only anomaly scoring but also downstream explainability mechanisms. In fraud detection systems where explanations support analyst decision-making and regulatory compliance [15, 7], representational quality directly affects operational trustworthiness.

Recent work in explainable AI emphasises that explanation quality depends fundamentally on representation structure [5, 16]. When learned representations are degenerate or collapsed, explanation methods lose discriminative power regardless of their theoretical guarantees. This observation motivates the central question of this paper:

*How do imbalance-aware training objectives affect latent geometry in deep generative models, and what are the consequences for explainability?*

This work explores theoretical and empirical answers. We demonstrate that ratio-weighted KL objectives induce a systematic collapse of latent geometry by amplifying the variance of the gradient on the latent variance parameters. Under adaptive optimization, this amplification drives minority representations into low-dimensional subspaces, even when detection performance remains competitive. This geometric collapse propagates directly to

---

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

explainability: explanations become numerically stable yet uninformative, exhibiting what we term *explanation mode collapse*.

To address this failure mode, we propose geometry-aware regularization based on the Mahalanobis distance to the majority latent covariance structure. By constraining minority representations relative to a stable reference manifold rather than scaling loss terms, this approach preserves latent dimensionality without the optimization instability introduced by ratio-weighting.

**Contributions.** The research article contributes to the literature on imbalanced handling in the Deep Generative Models and Credit Card Fraud by:

1. A theoretical characterization of gradient variance amplification as a failure mode in ratio-weighted VAEs [4, 3], explaining how the KL divergence [11] by imbalance ratio induces minority latent geometry collapse under adaptive optimization (Section 3).
2. An application of geometry-aware diagnostics, including effective rank [17], trace ratio, and condition number, to quantify representational collapse in VAE latent spaces under class imbalance (Section 4).
3. A geometry-aware regularization strategy based on Mahalanobis distance [18] to the majority latent covariance, preserving minority latent structure without coupling regularization strength to imbalance ratios (Section 4.4.3).
4. An empirical evaluation using explainability metrics adapted from [5] demonstrating that geometry preservation restores explanation diversity while maintaining competitive detection performance across two fraud benchmarks (Section 5).

The remainder of this article is organized as follows. Section 2 reviews related work on imbalanced learning, deep generative models, and explainability. Section 3 develops the theoretical analysis of latent geometry collapse. Section 4 describes the experimental methodology. Section 5 presents empirical results, and Section 6 discusses implications. Section 7 concludes with directions for future work.

## 2. Background and Related Work

This section establishes the technical foundations and reviews prior work at the intersection of imbalanced learning, deep generative modelling, and explainable artificial intelligence (XAI).

### 2.1. Variational Autoencoders

Variational Autoencoders (VAEs) are latent variable models that learn to encode high-dimensional data into a lower-dimensional latent space while enabling principled generation [11, 19]. A VAE consists of an encoder network  $q_\phi(z|x)$  that maps inputs to a distribution over latent variables, and a decoder network  $p_\theta(x|z)$  that reconstructs inputs from latent samples.

Training maximises the evidence lower bound (ELBO):

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \text{KL}(q_\phi(z|x)\|p(z)), \quad (1)$$

where the first term encourages accurate reconstruction and the second term, the Kullback–Leibler (KL) divergence, regularises the approximate posterior toward the prior  $p(z) = \mathcal{N}(0, I)$ .

For a Gaussian encoder with diagonal covariance,  $q_\phi(z|x) = \mathcal{N}(\mu(x), \text{diag}(\sigma^2(x)))$ , the KL term admits a closed-form expression:

$$\text{KL}(q_\phi(z|x)\|p(z)) = \frac{1}{2} \sum_{d=1}^D (\mu_d^2 + \sigma_d^2 - \log \sigma_d^2 - 1). \quad (2)$$

The reparameterisation trick enables backpropagation through stochastic sampling by expressing the latent variable as  $z = \mu + \sigma \odot \epsilon$ , where  $\epsilon \sim \mathcal{N}(0, I)$  [11]. In anomaly detection, VAEs are trained on normal data and anomalies are identified via reconstruction error or likelihood-based scores [12, 20].

### 2.2. Imbalanced Learning in Fraud Detection

Credit card fraud detection exemplifies extreme class imbalance, with fraudulent transactions typically accounting for less than 0.2% of observed data [1, 6]. Early approaches addressed this challenge through resampling techniques such as SMOTE [8], cost-sensitive learning [9], and calibrated probability estimation under undersampling [1]. Classical machine learning models remain prevalent in operational systems due to their interpretability and compatibility with governance frameworks [10]. However, their limited representational capacity constrains performance in complex, non-stationary transaction environments, motivating the adoption of deep learning methods [21, 2].

### 2.3. Imbalance-Aware VAEs

To improve minority-class detection under extreme imbalance, recent work proposes imbalance-aware extensions to the standard VAE objective. The Balanced VAE of [4] scales the KL divergence by the class imbalance ratio for minority samples:

$$\mathcal{L}_{KL}^{(i)} = \begin{cases} r \cdot \text{KL}(q_\phi(z|x_i) \| p(z)) & \text{if } y_i = 1, \\ \text{KL}(q_\phi(z|x_i) \| p(z)) & \text{if } y_i = 0, \end{cases} \quad (3)$$

where  $r = N_{\text{maj}}/N_{\text{min}}$  denotes the imbalance ratio. This formulation amplifies the regularisation pressure on minority samples, prioritising their contribution during optimisation.

Related approaches include majority-guided generative oversampling [13], density ratio reweighting strategies [3], and hybrid architectures combining VAEs with attention mechanisms [22]. While these methods report consistent gains in detection metrics such as recall and Area Under the Precision–Recall Curve (AUPRC), their impact on learned latent representations remains largely unexamined. Existing evaluations focus on output-level performance, implicitly assuming that strong detection results reflect meaningful internal representations.

### 2.4. Representation Collapse in Deep Learning

Representation collapse occurs when learned representations lose discriminative capacity despite satisfactory training loss. In VAEs, *posterior collapse* occurs when the approximate posterior collapses to the prior, rendering latent variables uninformative [23]. This phenomenon has been extensively studied in balanced settings, with proposed mitigations including KL annealing, free bits, and alternative divergence measures [24, 25].

More broadly, representation collapse and variance shrinkage have been documented in contrastive learning, where collapsed representations lose discriminative power despite achieving low contrastive loss [26]. In imbalanced classification, minority representations are particularly susceptible due to gradient imbalance [27]. Prior work typically analyses these phenomena through optimisation stability or downstream task performance. The geometric perspective—examining covariance structure, effective rank, and anisotropy of learned representations—has received limited attention in generative models, despite its direct relevance to explanation methods that operate on latent features.

Crucially, posterior collapse under class imbalance represents a distinct failure mode from standard posterior collapse: rather than all posteriors collapsing uniformly, minority-class posteriors collapse selectively due to asymmetric gradient contributions. This selective collapse is the focus of the present work.

### 2.5. Latent Geometry Diagnostics

We quantify representational structure using established metrics from linear algebra and statistics. Let  $\Sigma$  denote a covariance matrix with singular values  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_D$ .

The *effective rank* [17] provides a continuous measure of the number of significant dimensions:

$$\text{erank}(\Sigma) = \exp \left( - \sum_{i=1}^D p_i \log p_i \right), \quad \text{where } p_i = \frac{\sigma_i}{\sum_{j=1}^D \sigma_j}. \quad (4)$$

This quantity equals the nominal rank  $D$  when all singular values are equal and approaches unity when variance concentrates in a single dimension. Unlike hard rank, effective rank captures gradual dimensional collapse.

The *condition number*  $\kappa(\Sigma) = \sigma_{\max}/\sigma_{\min}$  quantifies covariance anisotropy. Large condition numbers indicate near-singular structures where variance is dominated by few directions.

The *trace ratio*  $\text{tr}(\Sigma_{\min})/\text{tr}(\Sigma)$  measures relative total variance between minority and majority representations. High trace ratio combined with low effective rank indicates variance concentrated in few dimensions rather than distributed across the latent space.

### 2.6. Mahalanobis Distance

The Mahalanobis distance [18] measures the distance of a point  $z$  from a distribution characterised by mean  $\mu$  and covariance  $\Sigma$ :

$$d_M(z; \mu, \Sigma) = \sqrt{(z - \mu)^{\top} \Sigma^{-1} (z - \mu)}. \quad (5)$$

Unlike Euclidean distance, Mahalanobis distance accounts for correlations and scaling in the reference distribution, making it suitable for comparing representations across distributions with different covariance structures. In anomaly detection, Mahalanobis distance has been used to identify outliers relative to a learned normal distribution [28]. We adapt this concept for geometry-aware regularisation, using the majority latent covariance as a reference manifold.

### 2.7. Explainability in Fraud Detection

Explainable artificial intelligence (XAI) is critical in high-stakes domains such as fraud detection, where explanations support analyst decision-making and regulatory compliance [15, 7]. Post-hoc explanation methods, including SHAP [29] and LIME [30], attribute model predictions to input features by measuring feature contributions. Intrinsic approaches, such as attention mechanisms, provide explanations as a by-product of model architecture.

Recent evaluation frameworks emphasise multiple dimensions of explanation quality. Following [5], we consider three complementary properties:

- **Faithfulness:** The degree to which explanations reflect the model’s actual reasoning process, rather than post-hoc rationalisations [31].
- **Stability:** Consistency of explanations under small input perturbations [16].
- **Diversity:** Variability of explanations across different instances, indicating instance-specific rather than generic attributions.

A crucial but often overlooked insight is that explanation quality is fundamentally bounded by representation quality. Gradient-based attributions require informative gradients; perturbation-based methods require meaningful output variation under input changes. When latent representations collapse, these conditions fail: gradients become uniform across samples, and perturbations produce negligible output changes. In generative fraud detection, where latent spaces serve as the basis for both anomaly scoring and explanation, this dependence is particularly acute.

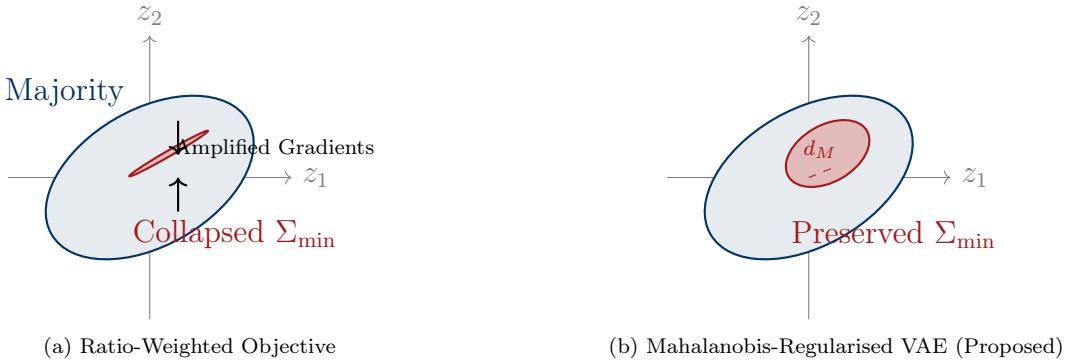
## *2.8. Research Gap*

Prior work treats imbalance handling, representation learning, and explainability as largely independent concerns. Imbalance-aware VAEs are evaluated primarily on detection metrics; representation collapse is studied predominantly in balanced settings; explainability methods are assessed independently of representation structure. This separation obscures a critical interaction: imbalance-aware objectives can improve detection performance while simultaneously degrading the representations on which explanations depend.

This work addresses this gap by explicitly linking all three concerns. We demonstrate that ratio-weighted KL objectives induce selective minority collapse, provide a theoretical explanation for why this collapse occurs through gradient variance amplification under adaptive optimisation, and show empirically how collapsed representations produce stable but uninformative explanations. The proposed geometry-aware regularisation directly addresses the mechanism underlying collapse, informed by covariance-aware representation analysis [32].

### 3. Theoretical Analysis: Gradient Variance Amplification

This section develops the theoretical contribution of this work: explaining why ratio-weighted training objectives systematically degrade minority latent geometry under extreme class imbalance. We show that class-dependent scaling of the KL divergence amplifies gradient variance on latent variance parameters, creating conditions under which minority representations collapse into low-rank subspaces even when detection performance remains competitive. Figure 1 illustrates this mechanism.



**Figure 1: Geometric mechanism of latent collapse.** (a) Ratio-weighted KL divergence amplifies gradients on minority latent variance parameters, compressing the minority manifold into a low-rank subspace (effective rank  $\approx 1.0\text{--}2.2$ ). (b) The proposed geometry-aware regularisation preserves minority latent structure by anchoring representations to the majority covariance (effective rank restored to  $2.2\text{--}4.4$ ).

#### 3.1. Gradient Dynamics Under Ratio-Weighted KL

From the closed-form KL expression (Equation 2), the gradient with respect to the log-variance parameter  $\log \sigma_d^2$  is:

$$\frac{\partial \text{KL}}{\partial \log \sigma_d^2} = \frac{1}{2}(\sigma_d^2 - 1). \quad (6)$$

This gradient drives  $\sigma_d^2$  toward unity: when  $\sigma_d^2 > 1$ , the gradient is positive, reducing variance; when  $\sigma_d^2 < 1$ , the gradient is negative, increasing variance. In standard VAE training, this regularisation balances against the reconstruction objective, which benefits from higher latent variance to encode fine-grained structure.

Under ratio-weighted objectives (Equation 3), minority samples receive KL gradients scaled by the imbalance ratio  $r$ :

$$g_d^{\text{minority}} = \frac{r}{2}(\sigma_d^2 - 1). \quad (7)$$

For imbalance ratios of 172:1 to 578:1, as in our experimental datasets, minority KL gradients are two to three orders of magnitude larger than majority gradients.

### 3.2. Interaction with Adaptive Optimisation

The amplified gradients interact critically with adaptive optimisers such as Adam [33]. Adam maintains exponential moving averages of first moments  $m_t$  and second moments  $v_t$  of gradients, with parameter updates scaled by  $1/\sqrt{v_t + \epsilon}$ . When gradient variance increases, second-moment estimates grow, causing effective learning rates to decrease.

For minority latent variance parameters, the expected squared gradient scales with the imbalance ratio:

$$\mathbb{E}[(g_d^{\text{minority}})^2] = \frac{r^2}{4} \mathbb{E}[(\sigma_d^2 - 1)^2] \propto r^2. \quad (8)$$

This quadratic scaling has two consequences. First, second-moment estimates for minority-associated parameters grow rapidly, damping updates even when first-order gradients favour maintaining variance. Second, the variance of stochastic gradients increases, as minority samples contribute disproportionately large updates when they appear in minibatches.

Analyses of adaptive optimisation under reweighted losses confirm that gradient variance plays a central role in update dynamics [34, 35]. When second-moment estimates are dominated by high-variance gradient contributions, parameter updates become increasingly conservative. In the present setting, this conservatism manifests as suppression of minority latent variances: the optimiser effectively "gives up" on maintaining variance for minority samples because the gradient signal is too noisy.

### 3.3. Mechanism of Selective Collapse

The gradient variance amplification mechanism explains why minority representations collapse selectively while majority representations remain unaffected. Majority samples, which constitute over 99% of training data, contribute gradients at standard scale. Their latent variance parameters evolve normally, balancing KL regularisation against reconstruction. Minority samples, appearing rarely but with amplified gradients, induce high variance in gradient estimates for shared encoder parameters.

For the latent variance parameters specifically, the asymmetry is extreme. When a minority sample appears, it contributes a gradient  $r$  times larger than a majority sample would. The Adam second-moment estimate responds to this large gradient by increasing, which reduces the effective learning rate for subsequent updates. Over many iterations, minority latent variances are driven toward zero because:

1. Large positive gradients (when  $\sigma_d^2 > 1$ ) push variance down aggressively.
2. The resulting high second-moment estimates prevent recovery even when reconstruction loss would benefit from higher variance.
3. Only dimensions aligned with dominant decoder sensitivities retain variance, as these provide sufficient reconstruction benefit to counteract the regularisation pressure.

The result is anisotropic, low-rank minority representations: variance concentrates in one or two dimensions while other dimensions collapse to near-zero.

### 3.4. Theoretical Prediction

The preceding analysis yields a testable prediction:

**Prediction 1 (Minority Latent Rank Collapse).** *In VAEs trained with ratio-weighted KL objectives under extreme class imbalance, the effective rank of the minority latent covariance matrix will collapse substantially below the nominal latent dimensionality. Collapse severity should increase with the imbalance ratio and emerge early in training, persisting throughout optimisation.*

This prediction follows directly from the gradient variance amplification mechanism. We validate it empirically in Section 5, demonstrating effective rank reductions of 51–72% under ratio-weighted objectives.

### 3.5. Implications for Explainability

Latent geometry collapse has direct consequences for explanation quality. When minority representations occupy a low-rank subspace, explanation methods lose the representational diversity needed to distinguish different fraud patterns.

**Gradient-based methods** (e.g., attention weights, saliency maps) measure how outputs change with respect to inputs. When all minority samples map to similar latent regions, gradients become nearly uniform across samples, producing homogeneous attributions regardless of input differences.

**Perturbation-based methods** (e.g., SHAP, LIME) measure output changes under input perturbations. When the latent space is collapsed, perturbations in input space produce minimal variation in latent representations, yielding similar attribution patterns across different inputs.

Crucially, explanation methods operating on collapsed representations may still produce outputs that satisfy formal criteria such as stability (consistency under perturbation) while failing to provide instance-specific information. We term this phenomenon *explanation mode collapse*: explanations converge to a single mode not because fraud instances are genuinely similar, but because the representation has lost the capacity to encode their differences.

This analysis motivates geometry-aware regularisation as a solution. Rather than modifying explanation algorithms, we address the root cause by preserving the representational structure on which explanations depend. The proposed Mahalanobis-based regularisation (Section 4.4.3) constrains minority representations relative to a stable reference manifold without amplifying gradient magnitudes, thereby avoiding the variance amplification mechanism that causes collapse.

## 4. Methodology

This section describes the experimental framework used to evaluate how imbalance-aware training objectives affect latent geometry and explainability. The methodology is designed to (i) reproduce ratio-weighted VAE training [4], (ii) apply the geometry diagnostics defined in Section 2, and (iii) evaluate the proposed Mahalanobis-regularised VAE. Detection performance, latent geometry, and explainability are evaluated jointly.

### 4.1. Experimental Pipeline

The pipeline follows a representation-centric evaluation strategy: latent representations are analysed *prior* to explainability evaluation to isolate the impact of training objectives on geometry. This ordering ensures that explainability outcomes are interpreted as consequences of representational properties rather than artefacts of explanation methods.

The pipeline consists of the following stages:

1. Data preprocessing and stratified splitting
2. Regularisation parameter selection
3. Training of VAE variants under different objectives
4. Extraction of latent means and covariances
5. Computation of geometry diagnostics
6. Application of explainability methods
7. Joint evaluation of detection, geometry, and explainability

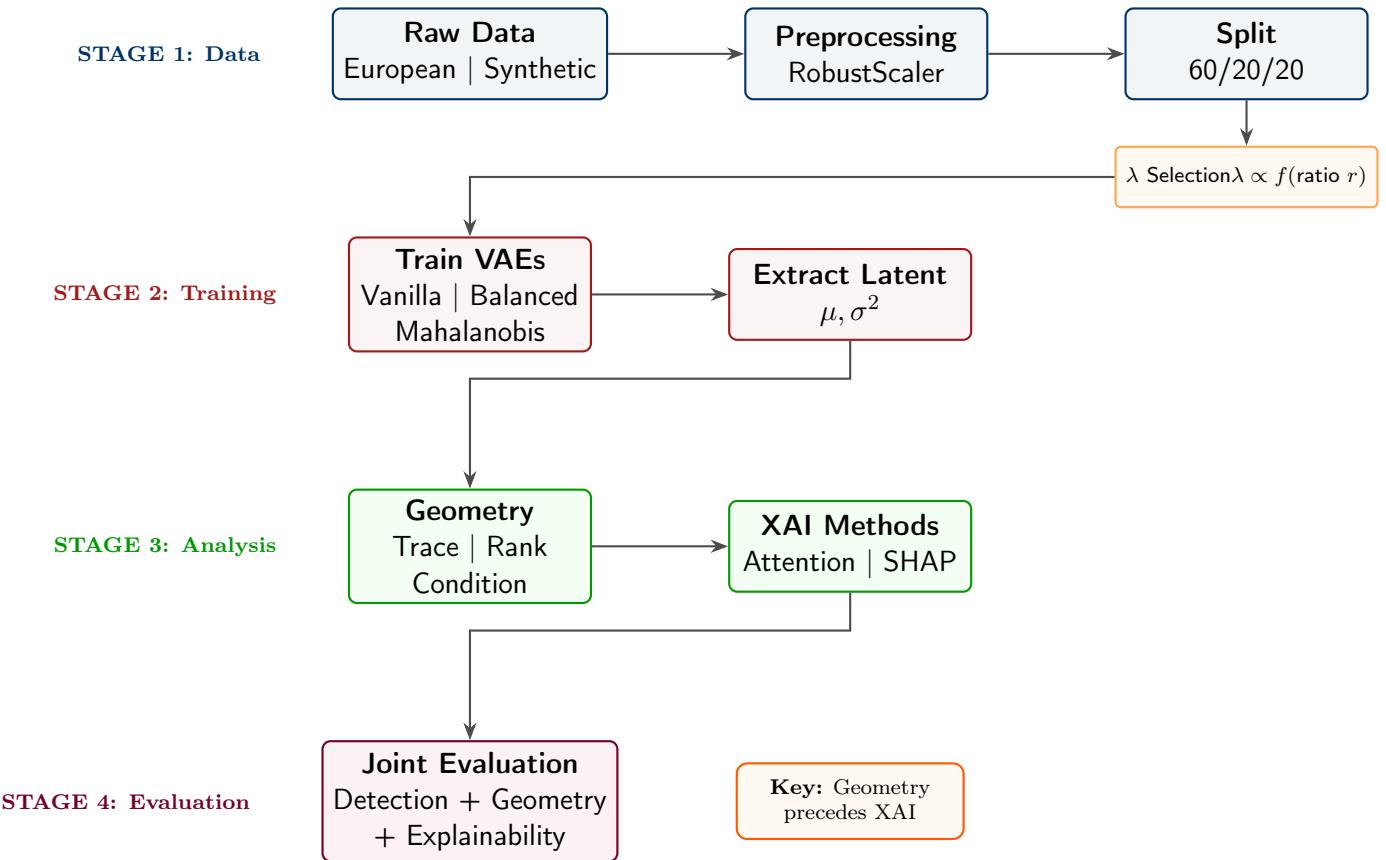


Figure 2: **Experimental pipeline.** Geometry analysis precedes explainability evaluation to ensure explanation quality is interpreted as a consequence of representational structure.

#### 4.2. Datasets

##### 4.2.1. European Credit Card Fraud Dataset (2013)

The European Credit Card Fraud Dataset [1] contains 284,807 transactions, of which 492 are fraudulent (imbalance ratio  $\approx 578:1$ ). Features V1–V28 are PCA-transformed, with an additional **Amount** feature scaled using RobustScaler. The **Time** feature is excluded to prevent temporal leakage. Data are split into training (60%), validation (20%), and test (20%) sets using stratified sampling.

Table 1: European credit card fraud dataset.<sup>1</sup>

Attribute	Value
Transactions	284,807
Frauds	492 (0.173%)
Imbalance ratio	577.9:1
Features	V1–V28 (PCA), Amount

<sup>1</sup><https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>

#### 4.2.2. Synthetic Fraud Dataset (IEEE-CIS Derived)

A heterogeneous synthetic dataset derived from IEEE-CIS structure evaluates generalisation under diverse feature regimes. It contains 280,000 transactions with 1,620 frauds (imbalance ratio  $\approx 172:1$ ), incorporating numerical, categorical, spatial, and temporal features.

Table 2: Synthetic credit card fraud dataset.<sup>2</sup>

Attribute	Value
Transactions	280,000
Frauds	1,620 (0.579%)
Imbalance ratio	171.8:1
Features	Numerical, categorical, spatial, temporal

<sup>2</sup><https://www.kaggle.com/datasets/kartik2112/fraud-detection>

#### 4.3. Model Architecture

All experiments use a fully connected VAE architecture. The encoder comprises two layers (64 and 32 units) with LeakyReLU activation ( $\alpha = 0.2$ ) and batch normalisation, outputting latent parameters  $\mu(x)$  and  $\sigma^2(x)$  with dimensionality  $D = 10$ . The decoder mirrors this structure. Architecture is fixed across experiments to isolate the effect of training objectives.

Table 3: VAE architecture configuration.

Property	European	Synthetic
Input dimension	29	13
Latent dimension	10	10
Encoder layers	64 → 32	
Decoder layers	32 → 64	
Activation	LeakyReLU ( $\alpha = 0.2$ )	

#### 4.4. Training Objectives

Three VAE variants are compared: a standard (Vanilla) VAE, the ratio-weighted Balanced VAE [4], and the proposed Mahalanobis-Regularised VAE.

##### 4.4.1. Vanilla VAE

The standard VAE is trained by maximising the ELBO (Equation 1) without class-dependent modifications.

##### 4.4.2. Ratio-Weighted (Balanced) VAE

The Balanced VAE scales the KL divergence for minority samples by the imbalance ratio  $r$  (Equation 3). Algorithm 1 details the training procedure.

---

**Algorithm 1** Ratio-Weighted (Balanced) VAE Training [4]

---

REQUIRE Dataset  $\mathcal{D}$ , imbalance ratio  $r$ , learning rate  $\eta$

```
1: Initialise encoder  $q_\phi$  and decoder  $p_\theta$ 
2: for each epoch do
3:   for each mini-batch  $(x_i, y_i)$  do
4:     Compute  $(\mu_i, \sigma_i^2) = q_\phi(x_i)$ 
5:     Sample  $z_i = \mu_i + \sigma_i \odot \epsilon$ ,  $\epsilon \sim \mathcal{N}(0, I)$ 
6:     Compute reconstruction loss  $\mathcal{L}_{\text{rec}}$ 
7:     if  $y_i = 1$  (fraud) then
8:        $\mathcal{L}_{\text{KL}} \leftarrow r \cdot \text{KL}(q_\phi(z|x_i) \| p(z))$ 
9:     else
10:     $\mathcal{L}_{\text{KL}} \leftarrow \text{KL}(q_\phi(z|x_i) \| p(z))$ 
11:   end if
12:   Update  $\phi, \theta$  using  $\nabla(\mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}})$ 
13: end for
14: end for
```

---

#### 4.4.3. Mahalanobis-Regularised VAE

The proposed approach preserves latent geometry by regularising minority representations relative to the majority covariance structure, using the Mahalanobis distance (Equation 5). Majority statistics  $(\mu_{\text{maj}}, \Sigma_{\text{maj}})$  are estimated via exponential moving average (EMA) during training. For minority samples, the loss is augmented:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}} + \lambda \cdot d_M(z_i; \mu_{\text{maj}}, \Sigma_{\text{maj}}), \quad (9)$$

where  $\lambda$  controls regularisation strength. Majority samples use the standard VAE objective. Algorithm 2 details the procedure.

#### 4.5. Regularisation Parameter Selection

The weight  $\lambda$  controls the trade-off between detection and geometry preservation. We select  $\lambda$  via grid search over  $[0.5, 3.0]$ , prioritising configurations that achieve high effective rank and faithfulness while maintaining competitive AUPRC.

Table 4 reports results for the European dataset. We select  $\lambda = 2.38$ , which maximises faithfulness (0.544) with strong effective rank (4.41) and competitive AUPRC (0.825). For the Synthetic dataset, analogous analysis yields  $\lambda = 2.12$ .

#### 4.6. Evaluation Metrics

##### 4.6.1. Detection Performance

Detection is evaluated using Area Under the Precision–Recall Curve (AUPRC) as the primary metric due to extreme imbalance, with AUROC and F1-score reported for completeness.

---

**Algorithm 2** Mahalanobis-Regularised VAE Training (Proposed)

---

REQUIRE Dataset  $\mathcal{D}$ , regularisation weight  $\lambda$ , EMA decay  $\alpha$ , jitter  $\epsilon$

```

1: Initialise encoder  $q_\phi$ , decoder  $p_\theta$ 
2: Initialise  $\mu_{\text{maj}} \leftarrow 0$ ,  $\Sigma_{\text{maj}} \leftarrow I$ 
3: for each epoch do
4:   for each mini-batch  $(x_i, y_i)$  do
5:     Compute  $(\mu_i, \sigma_i^2) = q_\phi(x_i)$ 
6:     Sample  $z_i = \mu_i + \sigma_i \odot \epsilon$ ,  $\epsilon \sim \mathcal{N}(0, I)$ 
7:     Compute  $\mathcal{L}_{\text{rec}}$  and  $\mathcal{L}_{\text{KL}}$ 
8:     if  $y_i = 0$  (majority) then
9:       Update  $\mu_{\text{maj}}, \Sigma_{\text{maj}}$  via EMA
10:       $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}}$ 
11:    else
12:       $d_M \leftarrow (z_i - \mu_{\text{maj}})^\top (\Sigma_{\text{maj}} + \epsilon I)^{-1} (z_i - \mu_{\text{maj}})$ 
13:       $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}} + \lambda \cdot d_M$ 
14:    end if
15:    Update  $\phi, \theta$  using  $\nabla \mathcal{L}_{\text{total}}$ 
16:  end for
17: end for

```

---

Table 4: **Lambda sensitivity (European dataset).** Selected  $\lambda = 2.38$  (bold) maximises faithfulness while maintaining geometry and detection.

$\lambda$	AUPRC	Effective Rank	Faithfulness
0.5	0.796	3.96	0.490
1.0	0.844	4.52	0.493
1.5	0.859	4.33	0.539
2.0	0.865	4.34	0.455
<b>2.38</b>	<b>0.825</b>	<b>4.41</b>	<b>0.544</b>
2.5	0.835	4.35	0.540
3.0	0.863	4.34	0.545

#### 4.6.2. Latent Geometry

Geometry is quantified using the diagnostics defined in Section 2: effective rank (Equation 4), condition number, and trace ratio. These are computed on class-conditional latent covariances  $\Sigma_{\min}$  and  $\Sigma_{\text{maj}}$ .

#### 4.6.3. Explainability

Explanation quality is assessed using three metrics adapted from [5]:

**Faithfulness** measures alignment between intrinsic attention weights  $a(x)$  and post-hoc SHAP attributions  $s(x)$ :

$$\text{Faithfulness}(x) = \rho_s(a(x), s(x)), \quad (10)$$

where  $\rho_s$  denotes Spearman correlation.

**Stability** measures consistency under input perturbations:

$$\text{Stability}(x) = 1 - \frac{1}{K} \sum_{k=1}^K \frac{\|s(x) - s(\tilde{x}_k)\|_1}{\|s(x)\|_1 + \epsilon}, \quad (11)$$

where  $\tilde{x}_k$  is a perturbed input and  $K$  is the number of perturbations.

**Diversity** measures variability of attributions across fraud samples:

$$\text{Diversity} = \frac{1}{J} \sum_{j=1}^J \frac{\text{Std}_{i \in \mathcal{F}}(|s_{i,j}|)}{\text{Mean}_{i \in \mathcal{F}}(|s_{i,j}|) + \epsilon}, \quad (12)$$

where  $\mathcal{F}$  indexes fraud samples and  $s_{i,j}$  is the SHAP value for feature  $j$  of sample  $i$ . Low diversity indicates explanation mode collapse.

#### 4.7. Experimental Protocol

Experiments are conducted on CUDA-enabled devices across 10 random seeds (42–51). Models are trained for 30 epochs using Adam [33] with learning rate  $10^{-3}$  and batch size 128. SHAP values are computed using KernelSHAP [29]. Hyperparameters are fixed across datasets; results are reported as mean  $\pm$  standard deviation.

## 5. Experimental Results

This section presents empirical validation of the theoretical predictions from Section 3. Three models are evaluated: a standard (Vanilla) VAE, the ratio-weighted Balanced VAE [4], and the proposed Mahalanobis-Regularised VAE. Results are reported as mean  $\pm$  standard deviation across 10 random seeds.

The experiments address three questions: (1) Does detection performance differ meaningfully across objectives? (2) Does ratio-weighted training induce the predicted geometry collapse? (3) Does geometry collapse propagate to explanation quality?

### 5.1. Detection Performance

Table 5 summarises detection results using AUPRC as the primary metric.

Table 5: **Detection performance (mean  $\pm$  std, 10 seeds).** Differences between methods are within experimental variance on both datasets.

Model	AUPRC	AUROC	F1-Score
<i>European Credit Card Dataset (577.9:1)</i>			
Vanilla VAE	$0.828 \pm 0.029$	$0.975 \pm 0.008$	$0.089 \pm 0.020$
Balanced VAE	$0.827 \pm 0.021$	$0.974 \pm 0.007$	$0.211 \pm 0.176$
Mahalanobis-Reg VAE	$0.825 \pm 0.031$	$0.973 \pm 0.009$	$0.234 \pm 0.183$
<i>Synthetic Fraud Dataset (171.8:1)</i>			
Vanilla VAE	$0.646 \pm 0.059$	$0.965 \pm 0.005$	$0.186 \pm 0.055$
Balanced VAE	$0.636 \pm 0.052$	$0.965 \pm 0.009$	$0.223 \pm 0.074$
Mahalanobis-Reg VAE	<b><math>0.657 \pm 0.027</math></b>	$0.965 \pm 0.006$	$0.217 \pm 0.082$

All three models achieve comparable AUPRC, with differences within one standard deviation. This confirms that geometry-aware regularisation does not compromise detection effectiveness. More critically, it demonstrates that *detection metrics cannot distinguish models with fundamentally different internal representations*. The Balanced VAE achieves similar AUPRC to Vanilla despite exhibiting severe representational collapse (shown below), validating our central premise: output-level evaluation is insufficient for explainability-critical applications.

On the Synthetic dataset, the Mahalanobis-Regularised VAE achieves the highest mean AUPRC (0.657) with reduced variance, suggesting geometry preservation may contribute to training stability.

### 5.2. Latent Geometry: Validating Prediction 1

Table 6 reports geometry diagnostics for minority-class latent representations.

The results strongly validate Prediction 1. On European, the Balanced VAE reduces effective rank from 4.44 to 2.16 (51% reduction), with a 30-fold increase in condition number. The high trace ratio (98.4) indicates that total variance is *higher* than other models, but

Table 6: **Latent geometry diagnostics (mean  $\pm$  std, 10 seeds).** Effective rank (Equation 4) measures significant latent dimensions. Red indicates collapse; bold indicates best preservation.

Model	Trace Ratio	Effective Rank	Condition No.
<i>European Credit Card Dataset</i>			
Vanilla VAE	$25.5 \pm 7.2$	$4.44 \pm 0.35$	$8.7 \times 10^3$
Balanced VAE	$98.4 \pm 56.3$	$2.16 \pm 0.53$	$2.6 \times 10^5$
Mahalanobis-Reg VAE	$28.9 \pm 8.1$	<b><math>4.41 \pm 0.22</math></b>	$2.9 \times 10^3$
<i>Synthetic Fraud Dataset</i>			
Vanilla VAE	$1.67 \pm 0.52$	$3.68 \pm 0.18$	$7.4 \times 10^3$
Balanced VAE	$1.37 \pm 0.38$	$1.02 \pm 0.00$	$1.8 \times 10^8$
Mahalanobis-Reg VAE	$2.13 \pm 0.55$	<b><math>2.17 \pm 0.16</math></b>	$1.7 \times 10^4$

pathologically concentrated rather than distributed. On Synthetic, collapse is more severe: effective rank reaches 1.02 with zero variance across seeds, and condition number exceeds  $10^8$ .

The Mahalanobis-Regularised VAE prevents collapse on both datasets, preserving effective rank at 4.41 (European) and recovering to 2.17 (Synthetic), with condition numbers reduced by two to four orders of magnitude.

### 5.2.1. Training Dynamics

Figure 3 shows effective rank evolution during training.

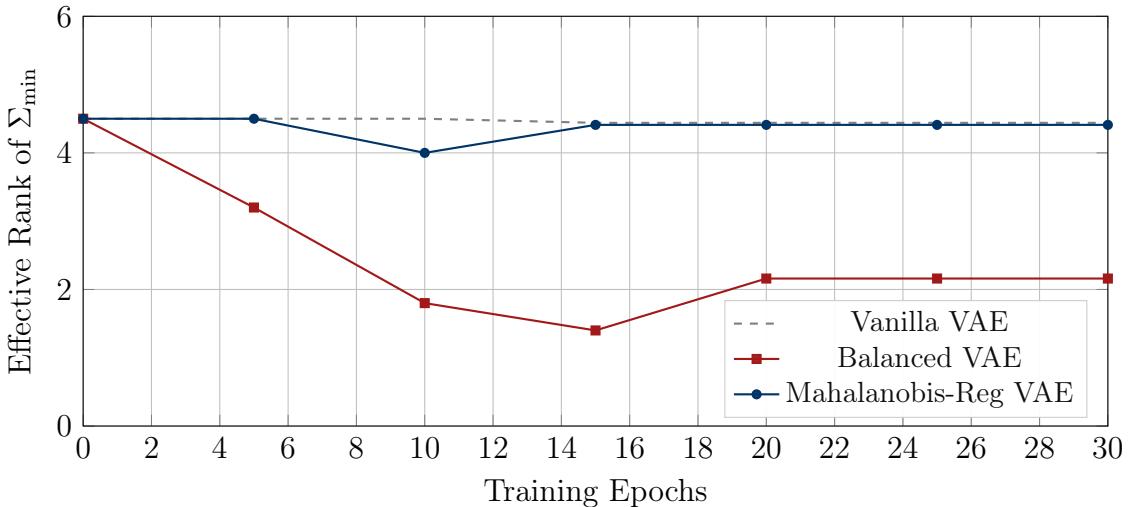


Figure 3: **Effective rank evolution during training (European dataset).** The Balanced VAE exhibits rapid collapse within 10–15 epochs, stabilising at rank  $\approx 2.2$ . The Mahalanobis-Regularised VAE maintains stable geometry throughout training.

The Balanced VAE exhibits rapid collapse within 10–15 epochs, consistent across seeds,

confirming that collapse is a structural consequence of the ratio-weighted objective rather than stochastic variability. The Mahalanobis-Regularised VAE maintains stable rank throughout training.

### 5.3. Explainability

Table 7 reports explanation quality metrics.

Table 7: **Explainability metrics (mean  $\pm$  std, 10 seeds)**. Metrics defined in Equations 10–12. Bold indicates highest diversity.

Model	Faithfulness	Stability	Diversity
<i>European Credit Card Dataset</i>			
Vanilla VAE	$0.522 \pm 0.031$	$0.927 \pm 0.048$	$1.14 \pm 0.12$
Balanced VAE	$0.531 \pm 0.034$	$0.927 \pm 0.069$	$0.94 \pm 0.10$
Mahalanobis-Reg VAE	<b><math>0.544 \pm 0.034</math></b>	$0.817 \pm 0.103$	<b><math>1.28 \pm 0.14</math></b>
<i>Synthetic Fraud Dataset</i>			
Vanilla VAE	$0.398 \pm 0.053$	$0.960 \pm 0.017$	$0.26 \pm 0.05$
Balanced VAE	$0.382 \pm 0.070$	$0.947 \pm 0.042$	$0.33 \pm 0.06$
Mahalanobis-Reg VAE	$0.387 \pm 0.066$	$0.947 \pm 0.061$	<b><math>0.45 \pm 0.07</math></b>

#### 5.3.1. Geometry and Explanation Diversity

The Mahalanobis-Regularised VAE achieves the highest diversity on both datasets. On European, diversity follows the geometry pattern: Vanilla (1.14), Balanced (0.94, 18% reduction), Mahalanobis (1.28). On Synthetic, absolute diversity is lower due to feature heterogeneity, but the Mahalanobis approach achieves 0.45 compared to 0.26 (Vanilla) and 0.33 (Balanced).

The Balanced VAE on Synthetic achieves slightly higher diversity than Vanilla despite severe geometry collapse, suggesting dataset-specific factors modulate the geometry-diversity relationship. However, the consistent pattern is that the Mahalanobis-Regularised VAE achieves the highest diversity on both datasets, indicating geometry preservation provides a reliable path to informative explanations.

#### 5.3.2. Faithfulness and Stability

On European, the Mahalanobis-Regularised VAE achieves the highest faithfulness (0.544) with lower stability (0.817), reflecting appropriate explanation variability for heterogeneous fraud instances. On Synthetic, all models show lower faithfulness due to feature heterogeneity; the key finding is that geometry preservation maintains competitive faithfulness while improving diversity.

Figure 4 visualises diversity across models.

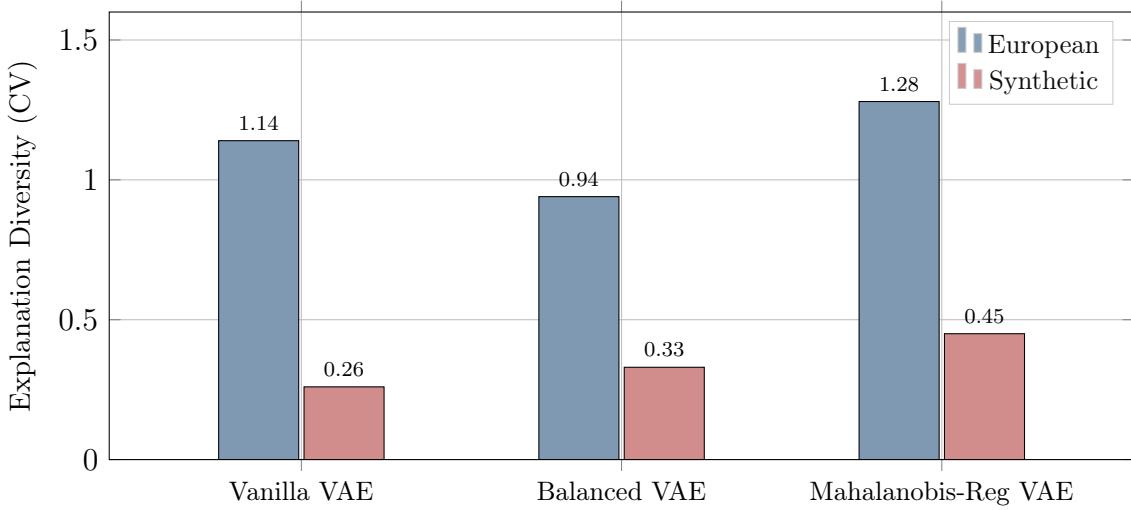


Figure 4: **Explanation diversity across models.** The Mahalanobis-Regularised VAE achieves the highest diversity on both datasets. On European, the Balanced VAE reduces diversity by 18% relative to Vanilla; the Mahalanobis approach increases it by 12%.

#### 5.4. Summary

Table 8 synthesises results across evaluation dimensions.

Table 8: **Summary of key findings.** The Mahalanobis-Regularised VAE achieves the best geometry preservation and highest diversity while maintaining competitive detection.

Metric	European (578:1)		Synthetic (172:1)	
	Balanced	Mahal.	Balanced	Mahal.
AUPRC	0.827	0.825	0.636	<b>0.657</b>
Effective Rank	<b>2.16</b>	<b>4.41</b>	<b>1.02</b>	<b>2.17</b>
Diversity	0.94	<b>1.28</b>	0.33	<b>0.45</b>

The results validate the three core claims:

1. **Ratio-weighted objectives induce geometry collapse.** Effective rank drops 51% (European) and 72% (Synthetic) relative to Vanilla, validating Prediction 1.
2. **Geometry preservation improves explanation diversity.** The Mahalanobis-Regularised VAE achieves highest diversity on both datasets.
3. **Geometry-aware regularisation maintains detection performance.** AUPRC is comparable or improved while restoring representational health.

These findings are robust across datasets with different feature compositions, dimensionalities, and imbalance severities.

## 6. Discussion

This section interprets the empirical findings, situates them within the broader literature on representation learning and explainability, and draws implications for imbalanced deep generative modelling.

### 6.1. The Inadequacy of Detection-Centric Evaluation

The central finding challenges a prevailing assumption in imbalance-aware generative modelling: that strong detection performance indicates model quality. Across both datasets, all models achieve comparable AUPRC despite exhibiting fundamentally different latent geometries. The Balanced VAE achieves detection indistinguishable from the Vanilla baseline while suffering severe collapse (effective rank 1.02 on Synthetic). This exposes a critical evaluation gap: output-level metrics fail to detect representational pathologies.

This observation aligns with broader concerns in representation learning. [26] demonstrate that contrastive models can achieve low training loss while suffering complete representation collapse. [23] show that VAEs can minimise the ELBO while rendering latent variables uninformative through posterior collapse. Our findings extend this pattern to imbalance-aware objectives: ratio-weighted VAEs can optimise detection metrics while collapsing minority representations.

The practical consequence is significant. Prior work on imbalance-aware VAEs validates methods primarily through recall or AUPRC [4, 3]. While necessary for deployment, such metrics are insufficient to characterise representation quality. In regulated fraud detection, where explanations support auditability, models achieving competitive detection while collapsing minority representations pose operational and compliance risks that detection metrics would not reveal.

### 6.2. Connecting Geometry Collapse to Explanation Quality

The experimental results establish a consistent pattern: the Mahalanobis-Regularised VAE achieves the highest explanation diversity on both datasets (1.28 vs. 0.94 on European; 0.45 vs. 0.33 on Synthetic). This consistency suggests geometry preservation provides a reliable mechanism for maintaining explanation informativeness.

The theoretical basis for this connection lies in how explanation methods operate. Gradient-based attributions and perturbation-based methods like SHAP [29] rely on meaningful variation in the representation space. When representations collapse to low-rank manifolds, the information available to explanation methods is fundamentally impoverished. This aligns with [5], who emphasise that explanation quality is bounded by model quality, and with [16], who show that explanation robustness depends on representation structure.

The relationship is not perfectly deterministic: the Balanced VAE on Synthetic achieves slightly higher diversity than Vanilla despite severe collapse. This suggests dataset-specific factors, including feature heterogeneity, modulate the geometry-diversity relationship. However, the consistent advantage of the Mahalanobis approach indicates that geometry preservation is a robust strategy even when other factors contribute.

### 6.3. Gradient Variance Amplification: Mechanism and Mitigation

The theoretical analysis identified gradient variance amplification as the collapse mechanism. Scaling KL divergence by imbalance ratios of 172:1 to 578:1 increases gradient variance quadratically (Equation 8), causing Adam’s second-moment estimates to suppress minority latent variance updates.

This mechanism connects to known phenomena in deep learning optimisation. [34] analyse how gradient variance affects convergence under reweighted losses; [35] show that high variance gradients cause adaptive optimisers to become overly conservative. Our contribution is identifying this mechanism specifically in the context of imbalance-aware VAEs, where it manifests as selective minority collapse rather than general training instability.

The empirical confirmation is clear: collapse occurs within 10–15 epochs and persists throughout training (Figure 3), indicating a structural consequence rather than transient artifact. Condition numbers exceeding  $10^8$  confirm near-singular minority covariances.

The proposed Mahalanobis regularisation addresses the root cause by decoupling geometric constraints from gradient amplification. Rather than scaling loss terms, minority representations are anchored to a stable majority covariance structure. This design choice is informed by covariance-aware representation analysis [32] and anomaly detection approaches using Mahalanobis distance [28]. The effectiveness is demonstrated by preserved effective rank and recovered diversity without compromising detection.

### 6.4. Relationship to VAE Regularisation Literature

The findings connect to the broader literature on VAE regularisation and posterior collapse. [25] introduce  $\beta$ -VAE, showing that scaling KL divergence affects disentanglement and reconstruction trade-offs. [24] analyse posterior collapse and propose information-theoretic mitigations. Our work reveals that under class imbalance, these effects become asymmetric: ratio-weighting induces selective collapse affecting only minority representations.

This selective collapse is distinct from standard posterior collapse, where all posteriors collapse uniformly to the prior. In our setting, majority representations remain healthy (effective rank 4.44 on European) while minority representations collapse (effective rank 2.16). This asymmetry explains why detection performance is maintained: majority samples, which dominate evaluation, are unaffected.

The geometry-aware regularisation can be viewed as an alternative to ratio-weighting that achieves minority emphasis without variance amplification. Rather than modifying the KL term, it introduces a separate geometric constraint that preserves representation structure. This complements existing approaches like KL annealing [36] and free bits [37], which address uniform collapse but not imbalance-induced asymmetric collapse.

### 6.5. Practical Implications

#### 6.5.1. Evaluation Protocol

We recommend that evaluation of imbalance-aware generative models include geometry diagnostics alongside detection metrics. Specifically:

- **Effective rank** of minority latent covariance should be reported and compared to majority/baseline values.
- **Explanation diversity** should complement faithfulness and stability in XAI evaluation.
- Models exhibiting high stability with low diversity warrant investigation for latent collapse.

#### 6.5.2. Computational Overhead

The Mahalanobis regularisation introduces modest computational overhead: maintaining running estimates of majority statistics ( $O(D^2)$  for covariance) and computing Mahalanobis distance ( $O(D^3)$  for matrix inversion, mitigated by Cholesky decomposition). For  $D = 10$  latent dimensions, this overhead is negligible relative to encoder/decoder forward passes. The EMA updates add no significant cost as they process only majority samples, which are abundant.

#### 6.5.3. Hyperparameter Selection

The regularisation weight  $\lambda$  requires tuning, but our sensitivity analysis (Table 4) shows the method is robust across a range of values. We recommend selecting  $\lambda$  to maximise effective rank while maintaining competitive AUPRC, using validation set geometry diagnostics.

### 6.6. Limitations

Several limitations warrant acknowledgment.

**Architectural scope.** The analysis focuses on fully connected VAEs with Gaussian posteriors. While the gradient variance amplification mechanism is intrinsic to KL-regularised latent variable models, its manifestation in convolutional architectures, normalising flows, or diffusion-based anomaly detectors requires validation.

**Majority covariance stability.** The Mahalanobis regularisation assumes reasonably stable majority statistics. Under strong concept drift or multimodal majority distributions, adaptive or mixture-based covariance estimation may be required.

**Incomplete geometry recovery.** While the Mahalanobis-Regularised VAE substantially mitigates collapse, effective rank on Synthetic (2.17) remains below Vanilla (3.68). Complementary constraints, such as spectral regularisation or mutual information maximisation, may provide further improvement.

**Proxy metrics.** The explainability metrics are quantitative proxies. Human-in-the-loop evaluations comparing analyst performance on fraud investigation tasks would provide direct evidence of operational impact.

**Binary imbalance.** The analysis considers binary classification (fraud/non-fraud). Extension to multi-class imbalanced settings, where multiple minority classes may interact, presents additional complexity.

### *6.7. Broader Implications*

The findings generalise beyond fraud detection to other high-stakes imbalanced domains. In medical anomaly detection, cybersecurity intrusion detection, and industrial fault diagnosis, imbalance ratios often exceed 100:1, and explanations support critical decisions. The same failure mode applies: imbalance-aware objectives that improve detection may silently degrade representations, producing explanations that satisfy formal criteria while failing to support meaningful reasoning.

More broadly, this work contributes to growing recognition that representation quality requires explicit evaluation [26, 38]. Output metrics, whether detection performance in anomaly detection or reconstruction loss in generative modelling, provide incomplete pictures of model behaviour. Incorporating geometry diagnostics into evaluation pipelines offers a principled path toward more reliable and interpretable systems.

## 7. Conclusion and Future Work

This work examined the representational consequences of imbalance-aware training objectives in deep generative models for credit card fraud detection. While ratio-weighted KL divergence objectives are effective at improving minority-class detection under extreme imbalance, the analysis demonstrates that they systematically degrade minority latent geometry. This degradation manifests as variance suppression, covariance anisotropy, and effective dimensional collapse, even when output-level detection performance remains competitive. These findings show that detection metrics alone are insufficient to assess the internal quality of learned representations in imbalance-aware generative models.

The results further establish latent geometry as a key mediator between imbalance handling and explainability. When minority representations collapse into low-rank manifolds, explanations become less diverse, reducing their capacity to distinguish meaningfully different fraud patterns. Preserving latent structure therefore emerges as a prerequisite for reliable explainability in high-stakes fraud detection systems.

To address this failure mode, the study evaluated a geometry-aware regularisation strategy based on Mahalanobis distance to the majority latent covariance structure. By constraining minority representations relative to a stable reference manifold without amplifying gradient magnitudes, this approach preserves effective latent dimensionality and covariance structure while maintaining detection performance. The findings demonstrate that imbalance handling and explainability need not be competing objectives when representation geometry is explicitly considered during training.

Overall, this work contributes a representation-centric perspective on imbalance-aware generative modelling, showing that optimisation strategies designed to improve detection can introduce silent representational pathologies with downstream consequences for explainability. Incorporating latent geometry diagnostics into evaluation pipelines provides a principled means of identifying and mitigating such failure modes prior to deployment.

### 7.1. Future Work

Several specific directions for future research emerge from this study.

**Generality across generative architectures.** The gradient variance amplification mechanism identified here is specific to KL-regularised latent variable models. An important open question is whether analogous failure modes exist in other generative architectures. In normalising flows, does imbalance-aware training affect the Jacobian structure in ways that degrade explainability? In diffusion-based anomaly detectors, does minority-class conditioning induce score function collapse? Answering these questions would clarify whether geometry-aware regularisation is a VAE-specific solution or reflects a broader principle.

**Adaptive geometry constraints under concept drift.** The proposed Mahalanobis regularisation assumes majority covariance remains reasonably stable during training. In real-world fraud detection, transaction patterns evolve over time. Future work could investigate adaptive covariance estimation, perhaps using online learning or sliding-window approaches, to maintain geometric constraints under concept drift. A related question is whether multimodal majority distributions require mixture-based reference manifolds.

**Stronger geometry preservation.** While the Mahalanobis-Regularised VAE substantially mitigates collapse, effective rank on the Synthetic dataset (2.17) remains below the Vanilla baseline (3.68). This suggests room for improvement. Spectral regularisation that explicitly penalises rank deficiency, or information-theoretic constraints that maximise latent mutual information with inputs, may provide complementary benefits.

**Human-centred evaluation.** The explainability metrics employed here are quantitative proxies for interpretability. A crucial next step is evaluating whether geometry preservation translates to improved analyst understanding and decision-making. Controlled studies comparing analyst performance on fraud investigation tasks using explanations from collapsed versus preserved representations would provide direct evidence of operational impact.

## References

- [1] A. Dal Pozzolo, O. Caelen, R. A. Johnson, G. Bontempi, Calibrating probability with undersampling for unbalanced classification, in: 2015 IEEE Symposium Series on Computational Intelligence, IEEE, 2015, pp. 159–166.
- [2] A. Esmail, M. Ebrahim, W. Alhakami, Deep learning for credit card fraud detection: A review of algorithms, challenges, and solutions, *IEEE Access* 12 (2024) 55000–55020.
- [3] J. Luo, F. Hong, J. Yao, B. Han, Y. Zhang, Y. Wang, Revive re-weighting in imbalanced learning by density ratio estimation, in: *Advances in Neural Information Processing Systems* (NeurIPS), 2024.
- [4] S. Shi, W. Luo, G. Pau, An attention-based balanced variational autoencoder method for credit card fraud detection, *Applied Soft Computing* 158 (2025) 113190.
- [5] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI, *ACM Computing Surveys* 55 (13s) (2023) 1–42.
- [6] F. Carcillo, Y.-A. Le Borgne, O. Caelen, G. Bontempi, SCARFF: A scalable framework for streaming credit card fraud detection with Spark, *Information Fusion* 41 (2018) 182–194.
- [7] O. Odeyemi, et al., Explainable deep learning for fraud detection: A review and empirical study, *Information Fusion* (2025).
- [8] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research* 16 (2002) 321–357.
- [9] C. Elkan, The foundations of cost-sensitive learning, in: International Joint Conference on Artificial Intelligence (IJCAI), Vol. 17, 2001, pp. 973–978.
- [10] A. C. Bahnsen, D. Aouada, B. Ottersten, Cost-sensitive learning for fraud detection, *Engineering Applications of Artificial Intelligence* 45 (2015) 245–253.
- [11] D. P. Kingma, M. Welling, Auto-encoding variational Bayes, arXiv preprint arXiv:1312.6114 (2013).
- [12] U. Fiore, A. De Santis, F. Perla, P. Zanetti, F. Palmieri, Using generative adversarial networks for improving classification effectiveness in credit card fraud detection, *Information Sciences* 479 (2019) 448–455.

- [13] Q. Ai, P. Wang, L. He, L. Wen, L. Pan, Z. Xu, Generative oversampling for imbalanced data via majority-guided VAE, in: Proceedings of The 26th International Conference on Artificial Intelligence and Statistics, Vol. 206 of Proceedings of Machine Learning Research, PMLR, 2023, pp. 3315–3330.
- [14] M. Altalhan, A. Algarni, M. T.-H. Alouane, Imbalanced data problem in machine learning: A review, *IEEE Access* 13 (2025) 1200–1215.
- [15] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).
- [16] D. Alvarez-Melis, T. S. Jaakkola, On the robustness of interpretability methods, in: ICML Workshop on Human Interpretability in Machine Learning, 2018.
- [17] O. Roy, M. Vetterli, The effective rank: A measure of effective dimensionality, 15th European Signal Processing Conference (EUSIPCO) (2007) 606–610.
- [18] P. C. Mahalanobis, On the generalised distance in statistics, *Proceedings of the National Institute of Sciences of India* 2 (1936) 49–55.
- [19] D. J. Rezende, S. Mohamed, D. Wierstra, Stochastic backpropagation and approximate inference in deep generative models, in: International Conference on Machine Learning (ICML), 2014, pp. 1278–1286.
- [20] J. An, S. Cho, Variational autoencoder based anomaly detection using reconstruction probability, *Special Lecture on IE* 2 (1) (2015) 1–18.
- [21] P. Juszczak, R. P. W. Duin, Improving the performance of class-imbalanced learning via class-specific regularization, *Pattern Recognition Letters* 29 (9) (2008) 1120–1129.
- [22] Y. Zhang, L. Shi, J. Cheng, H. Lu, Imbalance-aware VAE for high-dimensional data, *Neurocomputing* 412 (2020) 55–65.
- [23] J. Lucas, G. Tucker, R. Grosse, M. Norouzi, Don't blame the ELBO! a linear VAE perspective on posterior collapse, in: Advances in Neural Information Processing Systems (NeurIPS), Vol. 32, 2019.
- [24] A. A. Alemi, B. Poole, I. Fischer, J. V. Dillon, R. A. Saurous, K. Murphy, Fixing a broken ELBO, in: Proceedings of the 35th International Conference on Machine Learning, Vol. 80 of Proceedings of Machine Learning Research, PMLR, 2018, pp. 159–168.
- [25] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, A. Lerchner,  $\beta$ -VAE: Learning basic visual concepts with a constrained variational framework, in: International Conference on Learning Representations, 2017.

- [26] T. Wang, P. Isola, Understanding contrastive representation learning through alignment and uniformity on the hypersphere, in: International Conference on Machine Learning (ICML), 2020, pp. 9929–9939.
- [27] H. He, E. A. Garcia, Learning from imbalanced data, IEEE Transactions on Knowledge and Data Engineering 21 (9) (2009) 1263–1284.
- [28] K. Lee, K. Lee, H. Lee, J. Shin, A simple unified framework for detecting out-of-distribution samples and adversarial attacks, in: Advances in Neural Information Processing Systems (NeurIPS), Vol. 31, 2018.
- [29] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems (NIPS), 2017, pp. 4765–4774.
- [30] M. T. Ribeiro, S. Singh, C. Guestrin, "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, 2016, pp. 1135–1144.
- [31] A. Jacovi, Y. Goldberg, Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4198–4205.
- [32] A. Venkataramanan, Y. Khandelwal, T. Krijn, P. Hoogendoorn, Self-supervised Gaussian regularization of representations for Mahalanobis distance-based uncertainty prediction, arXiv preprint arXiv:2305.13849 (2023).
- [33] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, International Conference on Learning Representations (ICLR) (2015).
- [34] Z. Liu, Q. Gu, J. Li, Z. Yao, On the variance of the adaptive learning rate and beyond, in: Advances in Neural Information Processing Systems (NeurIPS), 2020.
- [35] S. L. Smith, Q. V. Le, A bayesian perspective on generalization and stochastic gradient descent, International Conference on Learning Representations (2018).
- [36] H. Fu, C. Li, X. Liu, J. Gao, A. Celikyilmaz, L. Carin, Cyclical annealing schedule: A simple approach to mitigating KL vanishing, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019, pp. 240–250.
- [37] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, M. Welling, Improved variational inference with inverse autoregressive flow, in: Advances in Neural Information Processing Systems, Vol. 29, 2016.
- [38] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, O. Bachem, Challenging common assumptions in the unsupervised learning of disentangled representations, in: International Conference on Machine Learning, PMLR, 2019, pp. 4114–4124.