Understanding the Explainability Consequences of Imbalance Handling in Deep Generative Fraud Detection Models: A Geometry-Aware Perspective

216011757

January 13, 2026

Abstract

Learning under extreme class imbalance remains a central challenge in high-stakes decision systems such as credit card fraud detection, where minority events are rare yet operationally critical. While imbalance-handling strategies are routinely introduced to improve detection performance, their impact on learned representations and the downstream reliability of model explanations has received limited scrutiny. This gap is particularly consequential in regulated domains, where explainability is a formal requirement rather than an optional model attribute. This study investigates how imbalance-aware training objectives in Variational Autoencoders (VAEs) affect latent space geometry and how these geometric effects propagate into both intrinsic and post hoc explainability methods. Fraud detection is used as a motivating application due to its severe class imbalance, with fraud prevalence as low as 0.17% in real-world transaction data. The study focuses on ratio-weighted Kullback–Leibler (KL) objectives, as employed in Balanced VAE frameworks, which explicitly scale regularisation terms to emphasise minority samples during training. A theoretical analysis demonstrates that ratio-weighted KL objectives amplify gradient variance acting on minority latent variance parameters in proportion to the imbalance ratio. Under extreme imbalance, this mechanism induces systematic shrinkage of minority covariance, resulting in latent manifold collapse. Empirical evaluation on the European Credit Card Fraud Dataset (2013) and a scale-matched synthetic subset of the IEEE-CIS Fraud Detection dataset confirms this effect.

Quantitative geometry-based diagnostics show that the minority-to-majority latent trace ratio drops from approximately 0.72 under a standard VAE to 0.03 under a Balanced VAE, accompanied by a reduction in effective latent rank from 4.91 to 1.35. Explainability metrics reveal near-zero attention–importance correlation, diminished latent comprehensiveness, and explanation mode collapse in SHAP attributions. Despite achieving strong detection performance (AUPRC up to 0.85), models trained with ratio-weighted objectives exhibit degraded explainability. To address this failure mode, the dissertation proposes a geometry-aware regularisation strategy based on the Mahalanobis distance. Rather than scaling loss terms by the imbalance ratio, the proposed method constrains minority latent representations relative to the majority covariance structure while preserving intra-class variability. Across both datasets, the Mahalanobis-regularised VAE consistently restores minority latent geometry, achieving trace ratios of approximately 0.68 and recovering over 70% of latent variance relative to collapsed baselines. These geometric improvements translate directly into explainability gains: attention–importance correlation more than doubles relative to Balanced VAE, latent rank is preserved, and explanation stability improves under input perturbations, all while maintaining competitive detection performance (AUPRC $\approx$ 0.86). The findings establish latent geometry as a critical intermediary between imbalance correction and explanation faithfulness in deep generative fraud detection models. The thesis demonstrates that imbalance-handling strategies can inadvertently undermine explainability even when predictive performance improves, and that preserving geometric structure is essential for trustworthy explanations. Importantly, the main geometric and explainability trends are consistent across both the European and IEEE-CIS datasets, indicating that the observed collapse and recovery patterns are not dataset-specific artefacts.

More broadly, this work argues that explainability in imbalanced learning is fundamentally a representational problem, necessitating geometry-aware model design rather than post hoc explanation alone.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

Many real-world machine learning problems are characterised by extreme class imbalance, where the minority class represents rare but operationally critical events. Examples include medical diagnosis, fault detection, cybersecurity intrusions, and financial fraud. In such settings, standard learning algorithms tend to optimise overall accuracy, often achieving deceptively strong performance while failing to model the minority class effectively [5, 12].

This challenge is particularly acute in financial fraud detection. Fraudulent transactions typically constitute less than 0.2% of total transaction volume, yet they carry disproportionate financial, regulatory, and reputational risk. Consequently, fraud detection systems must operate under two competing constraints: they must detect extremely rare events reliably, and they must provide explanations for their decisions in accordance with regulatory and governance requirements.

Deep learning models, and in particular deep generative models, have become increasingly popular in fraud detection due to their ability to learn complex, nonlinear representations of transaction behaviour. Variational Autoencoders (VAEs) are especially attractive in this domain because they learn latent representations of normal behaviour and identify anomalies as deviations from learned distributions [13]. However, VAEs are sensitive to optimisation dynamics and regularisation choices, particularly under severe class imbalance.

To address imbalance, many recent approaches introduce explicit imbalance-handling strategies such as cost-sensitive losses, focal loss, or ratio-weighted objectives. While these methods often improve detection metrics such as recall or AUPRC, their impact on the internal representations learned by deep models remains less understood. This gap is particularly concerning given the increasing reliance on explainable artificial intelligence (XAI) methods to justify automated decisions.

1.2 Problem Context: Credit Card Fraud Detection

Credit card fraud detection provides a practically important and technically challenging context in which to study imbalanced learning and explainability. Fraudulent transactions are extremely rare, often accounting for fewer than two transactions per thousand, yet they impose substantial costs on financial institutions and consumers.

The European Credit Card Fraud Dataset (ECCFD) illustrates this challenge clearly. The dataset contains 284,807 transactions, of which only 492 are labelled as fraudulent, corresponding to a fraud prevalence of approximately 0.172% and an imbalance ratio of roughly 1:585 [8]. A trivial classifier that labels all transactions as legitimate would achieve over 99% accuracy, yet be operationally useless.

Similar imbalance regimes exist in South Africa and other emerging markets. Reports from the South African Banking Risk Information Centre (SABRIC) indicate persistent growth in card-not-present fraud losses, even as fraudulent transactions remain a small fraction of overall volume [19]. Financial institutions therefore operate in environments that are simultaneously highly imbalanced and heavily regulated, requiring both strong detection performance and transparent decision-making.

Imbalance-handling strategies are therefore unavoidable in fraud detection. However, their interaction with representation learning and explainability has not been rigorously analysed, particularly in the context of deep generative models.

1.3 Explainability as a Core Requirement

Explainability has become a formal requirement in many high-stakes decision systems. Regulatory frameworks increasingly demand that automated decisions affecting individuals be accompanied by meaningful explanations. In the financial sector, explainability supports compliance, auditability, customer trust, and analyst decision-making.

Explainable Artificial Intelligence (XAI) methods attempt to address this need through intrinsic mechanisms, such as attention or interpretable architectures, and post hoc techniques, such as SHAP, LIME, and Integrated Gradients [15, 20]. These methods are widely applied in fraud detection pipelines.

However, recent work has demonstrated that explanations can appear stable and convincing while failing to reflect the true decision logic of the model [1, 2]. In particular, explanation reliability depends critically on the quality of the model's internal representations. If learned representations are degenerate, low-rank, or collapsed, explanation methods lose discriminative power regardless of their theoretical guarantees.

This observation motivates a deeper investigation into the relationship between imbalance handling, representation learning, and explainability.

1.4 Core Research Problem

Imbalance-aware training strategies modify optimisation dynamics in deep models. In the case of Variational Autoencoders, ratio-weighted Kullback–Leibler (KL) divergence objectives explicitly scale regularisation terms to emphasise minority samples. While this approach improves anomaly detection performance, it alters gradient magnitudes acting on latent variables.

Preliminary theoretical and empirical evidence suggests that under extreme imbalance, such ratio-weighted objectives may induce systematic shrinkage of minority latent variance, leading to latent geometry collapse [11, 14]. When minority representations collapse into narrow regions of latent space, intra-class variability is destroyed.

This geometric distortion has direct implications for explainability. Both intrinsic explanations, such as attention weights, and post hoc explanations, such as SHAP, rely on informative internal representations. If latent geometry collapses, explanations may become uniform, uninformative, or misleading, even as detection performance improves.

The core research problem addressed in this thesis is therefore:

> *How do imbalance-handling strategies in Variational Autoencoders affect latent representation geometry, and how do these geometric effects propagate into the reliability and faithfulness of explainability methods under extreme class imbalance?*

1.5 Research Aim

The aim of this research is to investigate the geometric consequences of imbalance-aware training in Variational Autoencoder-based fraud detection models, and to understand how these consequences influence explainability quality. The study further aims to design and evaluate a geometry-aware regularisation strategy that preserves minority latent structure while maintaining strong detection performance.

1.6 Research Objectives

The specific objectives of this study are to:

1. analyse theoretically how ratio-weighted KL divergence objectives affect gradient dynamics and latent variance in VAEs,

2. characterise latent geometry using covariance-based and rank-based diagnostics under extreme class imbalance,

3. investigate how latent geometry degradation impacts intrinsic and post hoc explainability methods,

4. design a geometry-aware regularisation approach based on Mahalanobis distance to preserve minority latent structure,

5. empirically evaluate detection performance, latent geometry, and explainability across multiple datasets and feature regimes.

1.7 Research Questions

This dissertation addresses the following research questions:

RQ1: How do ratio-weighted imbalance-handling objectives affect latent variance and geometry in Variational Autoencoders?

RQ2: How does latent geometry degradation influence intrinsic and post hoc explainability methods?

RQ3: Can geometry-aware regularisation mitigate latent collapse while maintaining competitive fraud detection performance?

RQ4: What trade-offs arise between detection accuracy, latent geometry preservation, and explanation faithfulness?

1.8 Significance of the Study

This study is significant for both academic research and practical deployment. From a research perspective, it contributes to a deeper understanding of how imbalance-aware objectives interact with representation learning in deep generative models. It challenges the assumption that improved detection metrics necessarily imply improved model quality.

From a practical perspective, the findings have direct relevance for regulated financial environments. Models that achieve strong detection performance while relying on collapsed latent representations may produce explanations that appear stable but fail to meet regulatory expectations of meaningfulness. By identifying latent geometry as a mediating factor, this work provides guidance for designing more trustworthy fraud detection systems. Importantly, the main geometric and explainability findings are replicated across both the European Credit Card Fraud Dataset and the scale-matched IEEE-CIS subset, indicating that the observed collapse and recovery patterns are not dataset-specific artefacts.

1.9 Scope and Limitations

This research focuses on:

- credit card fraud detection as a motivating application,
- tabular transaction data,
- Variational Autoencoders and imbalance-aware variants,
- explainability methods applicable to deep generative models.

Human-subject interpretability studies and user-trust evaluations are left for future work.

The study does not consider image, text, or graph-based fraud detection, nor does it evaluate human-subject interpretability. These directions are left for future work.

1.10 Structure of the Dissertation

The remainder of the dissertation is organised as follows. Chapter 2 reviews related work on fraud detection, class imbalance, generative models, and explainable artificial intelligence, and concludes with a comparative summary (Table 2.1) that positions the present work against existing imbalance-aware VAE approaches. Chapter 3 develops the theoretical framework linking imbalance-aware objectives to latent geometry collapse. Chapter 4 presents the experimental methodology. Chapter 5 reports empirical results. Chapter 6 discusses the implications of the findings. Chapter 7 concludes the thesis and outlines directions for future research.

Chapter 2

Background and Related Work

2.1 Chapter Overview and Positioning

This chapter provides a comprehensive and critical review of the literature underpinning this dissertation. Unlike compact survey sections typically found in journal articles, this chapter is deliberately expansive in scope and depth, reflecting its role in establishing scholarly grounding for a doctoral-level research trajectory.

The chapter serves four primary purposes:

1. To situate credit card fraud detection within the broader context of imbalanced learning and high-stakes decision systems.

2. To systematically review imbalance-handling strategies through a structured taxonomy.

3. To examine deep generative models—particularly Variational Autoencoders—as representation learners under imbalance.

4. To analyse the evolution of explainable artificial intelligence (XAI), with emphasis on explainability assumptions that depend on representation geometry.

A central organising principle of this chapter is that *most prior work treats imbalance, representation learning, and explainability as largely independent concerns.* This separation obscures important interactions between optimisation objectives, latent geometry, and explanation faithfulness. By reviewing the literature through a unified lens, this chapter exposes conceptual and methodological gaps that motivate the theoretical framework developed in Chapter 3.

2.2 Credit Card Fraud Detection as a Canonical Imbalanced Learning Problem

2.2.1 Historical Evolution of Fraud Detection Systems

Credit card fraud detection has evolved through several distinct methodological eras. Early systems were dominated by rule-based engines, handcrafted by domain experts and encoded as deterministic if–then rules. These systems offered transparency and interpretability but were brittle, difficult to scale, and unable to adapt to evolving fraud strategies [**?**].

As transaction volumes increased and fraud patterns diversified, machine learning approaches began to replace static rules. Classical supervised models such as logistic regression, decision trees, and random forests were widely adopted due to their ability to learn from historical data and generalise to unseen cases. However, these models exhibited systematic bias under severe class imbalance, often prioritising majority-class accuracy at the expense of minority recall.

The rise of big data and deep learning further shifted the landscape. Sequence models (e.g. LSTMs and GRUs) captured temporal dependencies in transaction streams, while deep feedforward networks and autoencoders enabled non-linear feature extraction. Despite improved predictive performance, these models introduced opacity, making it increasingly difficult to justify decisions to regulators and stakeholders.

2.2.2 Extreme Class Imbalance in Fraud Detection

A defining characteristic of credit card fraud detection is extreme class imbalance. Multiple empirical studies report fraud prevalence below 0.2% in real-world datasets [6, 8]. In the European Credit Card Fraud Dataset (ECCFD), only 492 of 284,807 transactions are fraudulent, corresponding to a fraud rate of 0.172% and an imbalance ratio of approximately 1:585.

Formally, let $y \in \{0, 1\}$ denote class labels for legitimate and fraudulent transactions respectively. The class priors satisfy:

$$P(y = 1) \ll P(y = 0).$$

This skew leads to pathological behaviour in standard learning algorithms. Accuracy becomes a misleading metric, decision boundaries drift toward the majority class, and minority samples exert limited influence during optimisation.

In the South African context, these challenges are compounded by regulatory and operational constraints. Reports from the South African Banking Risk Information Centre (SABRIC) document rising card-not-present fraud losses despite low fraud incidence [19].

Financial institutions must therefore operate detection systems that are simultaneously accurate, robust, and explainable.

2.2.3 Performance Metrics Under Imbalance

Due to the inadequacy of accuracy, alternative metrics have been adopted in fraud detection research. Precision–Recall curves are preferred because they explicitly capture the trade-off between false positives and false negatives under imbalance. The Area Under the Precision–Recall Curve (AUPRC) has become the de facto primary metric in modern fraud analytics.

Dal Pozzolo et al. [8] demonstrated that probability calibration and threshold selection play a crucial role in operational deployment. However, their work focuses on output distributions rather than internal representations or explanations.

This distinction is critical: strong detection metrics do not guarantee trustworthy model reasoning.

2.3 A Taxonomy of Imbalance Handling Strategies

To organise the extensive literature on imbalanced learning, this thesis adopts a taxonomy based on *where in the learning pipeline imbalance is addressed.* This taxonomy provides a conceptual scaffold for comparing methods and analysing their representational consequences.

The three primary categories are:

1. Data-level imbalance handling
2. Algorithm-level imbalance handling
3. Objective-level imbalance handling

This taxonomy is not merely descriptive; it reveals implicit assumptions about representation learning that are rarely interrogated.

2.4 Data-Level Imbalance Handling

2.4.1 Random Under- and Over-Sampling

The simplest approaches to imbalance handling involve modifying the dataset itself. Random undersampling reduces the number of majority samples, while random oversampling duplicates minority samples. Although easy to implement, both methods suffer from significant drawbacks. Undersampling discards potentially valuable information, while oversampling increases the risk of overfitting.

2.4.2 SMOTE and Synthetic Oversampling

The Synthetic Minority Over-sampling Technique (SMOTE), introduced by Chawla et al. [7], marked a major advance in data-level imbalance handling. Rather than duplicating minority samples, SMOTE generates synthetic examples by interpolating between minority nearest neighbours:

$$x_{\text{new}} = x_i + \lambda(x_j - x_i), \quad \lambda \sim \mathcal{U}(0,1).$$

SMOTE assumes that minority samples lie on a locally linear manifold. Numerous extensions—such as Borderline-SMOTE and ADASYN—attempt to refine sampling near decision boundaries. Despite widespread adoption, these methods struggle in high-dimensional spaces and can distort minority geometry.

In fraud detection, SMOTE-based approaches have been shown to improve recall but often degrade precision, particularly when fraud patterns are heterogeneous [6].

2.4.3 Limitations of Data-Level Methods

Crucially, data-level methods operate independently of model optimisation. They do not influence gradient dynamics or latent representation learning directly. As a result, they cannot prevent representational collapse in deep models, nor can they guarantee that explanations derived from such models remain faithful.

This limitation motivates algorithm- and objective-level approaches.

2.5 Algorithm-Level Imbalance Handling

Algorithm-level imbalance handling methods modify the learning algorithm itself rather than altering the data distribution. These approaches aim to bias the learning process toward minority samples while preserving the original dataset. In contrast to data-level methods, algorithm-level techniques directly influence optimisation dynamics and decision boundary formation.

2.5.1 Cost-Sensitive Learning

Cost-sensitive learning is one of the earliest and most influential algorithm-level approaches to class imbalance. Rather than treating all classification errors equally, cost-sensitive methods assign higher penalties to misclassification of minority samples.

Elkan [10] formalised the foundations of cost-sensitive learning by demonstrating that optimal decision-making under unequal misclassification costs requires modifying either the loss function or the posterior decision threshold. Let C_{ij} denote the cost of predicting class i when the true class is j. The expected risk minimisation objective becomes:

$$\mathcal{R}(f) = \mathbb{E}_{(x,y)}[C_{f(x),y}].$$

In binary classification, assigning a higher cost C_{01} to false negatives encourages the classifier to prioritise minority recall. Cost-sensitive learning has been widely adopted in fraud detection due to its conceptual simplicity and compatibility with existing classifiers.

However, several limitations have been documented:

- Cost matrices are often difficult to estimate accurately in practice.
- Excessive cost scaling can destabilise optimisation.
- Cost-sensitive objectives bias outputs but do not explicitly preserve minority representation geometry.

In deep learning contexts, cost-sensitive loss functions are commonly implemented by weighting the cross-entropy loss:

$$\mathcal{L}_{\text{CS}} = -\alpha_y \log p(y|x),$$

where α_y is larger for the minority class. While effective at improving recall, this approach inherits many of the same instability issues later observed in ratio-weighted generative objectives.

2.5.2 Threshold Adjustment and Probability Calibration

An alternative to modifying the training objective is to adjust decision thresholds post hoc. Dal Pozzolo et al. [8] showed that recalibrating posterior probabilities using undersampling or threshold tuning can significantly improve operational precision–recall trade-offs in fraud detection.

Let $\hat{p}(y = 1|x)$ denote a model's estimated posterior. Rather than predicting fraud when $\hat{p} > 0.5$, an adjusted threshold $\tau \ll 0.5$ is used:

$$\hat{y} = \begin{cases} 1 & \text{if } \hat{p} > \tau, \\ 0 & \text{otherwise.} \end{cases}$$

While threshold adjustment is effective for deployment, it is fundamentally a decision-layer correction. It does not alter internal representations or training dynamics and therefore has no effect on explainability quality or latent geometry.

2.5.3 Ensemble-Based Methods

Ensemble approaches, such as Balanced Random Forests and EasyEnsemble, address imbalance by training multiple classifiers on rebalanced subsets of data. These methods reduce variance and improve minority recall through aggregation.

Although ensemble methods remain competitive for tabular fraud detection, they introduce substantial complexity and computational cost. More importantly, ensemble-based decisions are difficult to explain coherently, as explanations must aggregate across heterogeneous base learners.

As a result, ensemble methods have limited applicability in regulated environments where explanation traceability is required.

2.6 Objective-Level Imbalance Handling

Objective-level approaches embed imbalance handling directly into the loss function. Unlike algorithm-level methods, these approaches influence gradient magnitudes and parameter updates throughout training.

This category is particularly relevant for deep learning and generative models.

2.6.1 Weighted Loss Functions

Weighted loss functions generalise cost-sensitive learning by directly scaling loss contributions:

$$\mathcal{L} = \sum_{i=1}^{N} \alpha_{y_i} \ell(f(x_i), y_i).$$

In deep neural networks, this approach is ubiquitous due to its simplicity and compatibility with stochastic gradient descent. However, several studies report that aggressive weighting can cause gradient explosion or vanishing effects, particularly in highly imbalanced regimes [5, 12].

These effects are often mitigated heuristically through learning rate tuning, gradient clipping, or early stopping—none of which address underlying representational distortions.

2.6.2 Focal Loss

Focal loss was introduced to address extreme imbalance in object detection by down-weighting easy examples and focusing learning on hard, misclassified samples. The focal loss modifies cross-entropy as:

$$\mathcal{L}_{\text{focal}} = -(1 - p_t)^{\gamma} \log(p_t),$$

where p_t is the predicted probability of the true class and $\gamma > 0$ controls the focusing effect.

While focal loss has been successful in computer vision tasks, its applicability to tabular fraud detection is mixed. More importantly, focal loss modifies output gradients but does not explicitly control latent representation geometry.

2.7 Generative Models for Imbalanced Learning

Generative models represent a distinct paradigm for handling imbalance. Instead of directly modifying discriminative boundaries, generative approaches aim to model the underlying data distribution and synthesise new minority samples.

2.7.1 Autoencoders and Anomaly Detection

Autoencoders learn compact representations by reconstructing inputs through a bottleneck architecture. In fraud detection, they are often trained exclusively on majority (legitimate) transactions. Fraud is detected as reconstruction error:

$$\text{AnomalyScore}(x) = \|x - \hat{x}\|.$$

While attractive in unsupervised settings, this approach assumes that fraud is structurally distinct from legitimate behaviour. In practice, sophisticated fraud often mimics legitimate patterns, limiting reconstruction-based detection.

2.7.2 Variational Autoencoders

Variational Autoencoders (VAEs), introduced by Kingma and Welling [13], extend autoencoders by learning a probabilistic latent representation. The stochastic latent space enables both density estimation and sample generation, making VAEs particularly attractive for minority oversampling.

VAEs optimise the Evidence Lower Bound (ELBO), balancing reconstruction fidelity and latent regularisation. In imbalanced datasets, however, the ELBO is dominated by majority samples, leading to poor minority representation.

2.7.3 Generative Oversampling in Fraud Detection

Several studies have proposed using VAEs and GANs to generate synthetic fraud samples [**?**]. These approaches improve classification metrics by augmenting minority data. However, empirical evaluations often focus exclusively on AUPRC and recall, without examining representation quality or explanation faithfulness.

Moreover, synthetic data quality is rarely assessed beyond superficial similarity metrics, raising concerns about mode collapse and redundancy.

2.8 Balanced Variational Autoencoders

Balanced VAE frameworks attempt to address imbalance by modifying the VAE objective. Shi et al. [18] propose scaling the KL divergence term for minority samples using the imbalance ratio. The motivation is to force the latent posterior of minority samples away from the prior, increasing their influence during training.

While Shi et al. improve detection performance, they do not analyse the effects of ratio-weighted KL regularisation on latent covariance structure or explainability reliability.

Empirically, Balanced VAEs demonstrate improved detection performance. However, the literature largely treats these improvements as unambiguously positive, without analysing the effect of ratio-weighting on latent geometry or downstream explainability.

This omission constitutes a critical gap addressed by the present thesis.

2.9 Summary and Positioning

This section reviewed algorithm-level and objective-level imbalance handling strategies, highlighting their strengths and limitations. A key observation emerging from this review is that:

> *Most imbalance-handling methods prioritise output-level performance metrics while implicitly assuming that internal representations remain well-structured.*

In deep generative models, this assumption is particularly fragile. Objective-level imbalance corrections directly manipulate gradient dynamics, with consequences that extend beyond classification accuracy.

A further synthesis reveals that loss reweighting—whether through cost-sensitive learning, weighted cross-entropy, or focal loss—affects latent variance parameters, not only decision boundaries. This distinguishes such methods from capacity-control approaches such as β-VAEs (which target disentanglement, not imbalance) and focal loss (which emphasises hard examples at the output level). Critically, none of these existing methods analyse or prevent latent geometry collapse under extreme class imbalance.

The next part of this chapter examines explainable artificial intelligence (XAI), tracing its evolution and critically analysing the representational assumptions underlying both intrinsic and post hoc explanation methods.

2.10 Latent Representation Learning in Deep Models

A central premise of modern deep learning is that intermediate representations encode progressively more abstract and semantically meaningful structure. In supervised learning, these representations support classification and regression. In unsupervised and generative learning, representations define the geometry of the data manifold itself.

In high-stakes applications such as fraud detection, representation quality has implications beyond predictive performance. Internal representations influence:

- generalisation to rare patterns,
- robustness to perturbations,
- synthetic data diversity,
- the validity of explanations derived from the model.

This section reviews how latent representations are learned in generative models, with a focus on Variational Autoencoders (VAEs), and how imbalance affects their geometry.

2.11 Variational Autoencoders: Foundations

Variational Autoencoders were introduced by Kingma and Welling [13] as a principled framework for learning probabilistic latent variable models using variational inference.

Given observed data $\mathbf{x} \in \mathbb{R}^d$ and latent variables $\mathbf{z} \in \mathbb{R}^k$, VAEs model the joint distribution:

$$p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z}),$$

where $p(\mathbf{z})$ is typically an isotropic Gaussian prior:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Direct maximisation of the marginal likelihood $p_\theta(\mathbf{x})$ is intractable. Instead, VAEs optimise the Evidence Lower Bound (ELBO):

$$\mathcal{L}_{\text{ELBO}}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \text{KL}\Big(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z})\Big), \tag{2.1}$$

where $q_\phi(\mathbf{z}|\mathbf{x})$ is an approximate posterior parameterised by an encoder network.

2.11.1 Latent Parameterisation

In standard VAEs, the encoder outputs a diagonal Gaussian:

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}\Big(\boldsymbol{\mu}_\phi(\mathbf{x}), \mathrm{diag}(\boldsymbol{\sigma}^2_\phi(\mathbf{x}))\Big).$$

Each input is therefore mapped not to a point, but to a distribution in latent space. This stochasticity enables smooth interpolation, uncertainty modelling, and generative sampling.

2.12 Latent Geometry and Representation Capacity

Latent representations learned by VAEs define a geometric structure over $\mathbb{R}^k$. This geometry is characterised by:

- the distribution of latent means,
- the magnitude and anisotropy of latent variances,
- correlations between latent dimensions.

Aggregated over samples, this structure can be described by class-conditional statistics:

$$\boldsymbol{\mu}_c = \mathbb{E}[\mathbf{z} \mid y = c], \qquad \boldsymbol{\Sigma}_c = \mathrm{Cov}(\mathbf{z} \mid y = c).$$

These quantities determine:

- manifold volume (via $\mathrm{tr}(\boldsymbol{\Sigma}_c)$),
- intrinsic dimensionality (via effective rank),
- separation between classes,
- robustness of downstream explanations.

Latent geometry therefore plays a foundational role in both detection and explainability.

2.13 Posterior Collapse in VAEs

Posterior collapse refers to a failure mode in VAEs where the learned posterior $q_\phi(\mathbf{z}|\mathbf{x})$ becomes identical to the prior $p(\mathbf{z})$ for all inputs. When this occurs, latent variables carry little or no information about $\mathbf{x}$.

Formally, collapse corresponds to:

$$q_\phi(\mathbf{z}|\mathbf{x}) \approx p(\mathbf{z}) \quad \Rightarrow \quad \boldsymbol{\mu}_\phi(\mathbf{x}) \to \mathbf{0},\ \boldsymbol{\sigma}^2_\phi(\mathbf{x}) \to \mathbf{1}.$$

Lucas et al. [14] demonstrated that posterior collapse is not merely an optimisation artefact, but can arise naturally when the decoder is sufficiently expressive. In such cases, the model prefers to ignore latent variables entirely.

He et al. [11] further showed that lagging inference networks exacerbate collapse by preventing the encoder from keeping pace with the decoder during training.

2.13.1 Consequences of Posterior Collapse

Posterior collapse has several consequences:

- latent variables become uninformative,
- generative diversity is reduced,
- downstream classifiers receive impoverished representations,
- explainability methods operating on latent features become meaningless.

While posterior collapse has been extensively studied in balanced settings, its interaction with class imbalance remains underexplored.

2.14 Class Imbalance and Latent Degeneracy

In highly imbalanced datasets, the ELBO objective is dominated by majority samples. This has two critical effects:

1. The reconstruction term primarily reflects majority structure.
2. The KL term regularises minority samples toward the prior.

Minority samples therefore experience a form of asymmetric regularisation pressure, which can manifest as:

- reduced latent variance,
- clustering near the origin,
- loss of intra-class diversity.

Unlike classical posterior collapse, this phenomenon may affect only the minority class, producing a *selective collapse* that is difficult to detect using aggregate metrics.

2.15 Latent Geometry in Fraud Detection

Fraud detection presents a particularly severe case of imbalance. Fraudulent transactions often:

- overlap structurally with legitimate behaviour,
- vary across multiple latent dimensions,
- evolve over time.

A collapsed latent representation forces heterogeneous fraud patterns into a narrow region of latent space. This has several negative consequences:

- synthetic samples lack diversity,
- downstream classifiers overfit narrow patterns,
- explanations become uniform across fraud cases.

Despite this, much of the fraud detection literature evaluates generative models exclusively using detection metrics such as AUPRC, without inspecting latent geometry.

2.16 Early Attempts at Geometry Control

Several approaches have attempted to indirectly control latent geometry:

- β-VAEs increase KL weight to encourage disentanglement,
- InfoVAEs modify mutual information terms,
- VampPrior replaces the isotropic prior with a learned mixture.

While these methods influence geometry, they are not designed to address class imbalance and may worsen minority collapse when applied naively.

2.17 Limitations of Existing Work

The literature reviewed in this chapter reveals several unresolved gaps:

1. Latent geometry is rarely analysed explicitly in imbalanced generative models.
2. Imbalance-aware objectives are evaluated primarily through detection metrics.
3. Posterior collapse is studied in balanced settings, but not under extreme imbalance.

4. Explainability is treated as an output-layer property, disconnected from representation quality.

These gaps motivate the need for a framework that links imbalance handling, latent geometry, and explainability.

2.18 Transition to Theoretical Framework

This chapter has established that:

- imbalance-handling strategies influence optimisation dynamics,
- VAEs are sensitive to latent regularisation pressure,
- latent geometry underpins both detection and explainability.

However, existing work stops short of formally analysing how imbalance-aware objectives alter gradient behaviour and induce selective latent collapse.

Chapter 3 addresses this gap by developing a theoretical analysis of ratio-weighted KL divergence, identifying a gradient variance amplification mechanism, and formally characterising minority latent manifold collapse.

2.19 Explainable Artificial Intelligence: Foundations and Taxonomy

Explainable Artificial Intelligence (XAI) seeks to make the decisions of complex machine learning models transparent, interpretable, and justifiable to human stakeholders. In high-stakes domains such as financial fraud detection, explainability is not merely a desirable property but a regulatory and operational requirement.

The faithfulness of post hoc explanations depends critically on the geometry of learned representations. When latent manifolds collapse, explanations become stable but uninformative—not because the XAI algorithm fails, but because the representation lacks discriminative structure. Attention weights and SHAP attributions are therefore geometry-dependent: their reliability is contingent on the preservation of high-rank, class-conditional latent covariance. This observationally supported relationship motivates the analysis in Chapter 3, where representation collapse is shown to be the root cause of explainability degradation, rather than a mere correlate.

Regulatory frameworks increasingly demand that automated decisions affecting individuals be accompanied by meaningful explanations. In financial systems, such explanations must support auditing, dispute resolution, and compliance verification, placing stringent demands on explanation reliability.

Despite rapid growth in XAI research, explainability remains an ill-defined concept, encompassing multiple objectives and methodological paradigms. This section reviews the dominant XAI taxonomies and critically examines their assumptions, particularly with respect to learned representations.

2.20 Taxonomy of Explainability Methods

XAI methods are commonly classified along several orthogonal dimensions. This dissertation adopts a taxonomy based on three axes:

- **Intrinsic vs Post Hoc**
- **Local vs Global**
- **Model-Specific vs Model-Agnostic**

This taxonomy aligns with established frameworks proposed by Doshi-Velez and Kim [9] and later refined in applied ML literature.

2.20.1 Intrinsic Explainability

Intrinsic explainability refers to models whose structure is assumed to be interpretable by design. Examples include:

- linear and additive models,
- decision trees and rule-based systems,
- attention-based neural architectures.

In deep learning, attention mechanisms [**? ?**] are frequently presented as intrinsically explainable, based on the premise that attention weights indicate feature relevance.

However, this premise has been repeatedly challenged. Benchaji et al. [4] demonstrate that attention weights can be fragile and sensitive to internal representation collapse. Miró-Nicolau et al. [16] further show that attention explanations may diverge from gradient-based importance, particularly under distributional shift.

2.20.2 Post Hoc Explainability

Post hoc methods generate explanations after model training, without modifying the underlying model. These methods are widely used due to their flexibility and model-agnostic nature.

Key post hoc techniques include:

- LIME [17]
- SHAP [15]
- Integrated Gradients [20]

While post hoc methods enable explanation of otherwise opaque models, they rely on strong assumptions about local linearity, smoothness, and representation stability.

2.21 Local Explanation Methods

2.21.1 LIME

Local Interpretable Model-Agnostic Explanations (LIME) were introduced by Ribeiro et al. [17] to approximate complex model behaviour locally using simple surrogate models.

Given an input $\mathbf{x}$, LIME:

1. perturbs $\mathbf{x}$ in its neighbourhood,
2. evaluates the model on perturbed samples,
3. fits a sparse linear model to approximate local behaviour.

LIME assumes that:

- the decision boundary is locally linear,
- perturbations remain within the data manifold,
- feature independence approximations are acceptable.

In high-dimensional, imbalanced settings, these assumptions are often violated. Perturbations may leave the data manifold entirely, particularly when latent representations are collapsed or anisotropic.

2.21.2 SHAP

SHAP (SHapley Additive exPlanations) [15] builds on cooperative game theory to assign feature attributions that satisfy desirable axioms such as local accuracy and consistency.

SHAP explanations are computed by estimating marginal contributions of features across subsets:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \Big(f(S \cup \{i\}) - f(S) \Big).$$

While theoretically grounded, SHAP assumes that feature perturbations yield meaningful model responses. When representations collapse, output sensitivity diminishes, leading to explanation mode collapse where distinct inputs receive near-identical attributions.

2.22 Global Explainability Methods

Global explainability seeks to characterise model behaviour across the entire input distribution. Common approaches include:

- Partial Dependence Plots (PDPs),
- Accumulated Local Effects (ALE) [3],
- feature importance aggregation.

ALE improves upon PDPs by accounting for correlated features and local data density. However, ALE still relies on stable and expressive internal representations. When latent geometry collapses, global explanations become overly smooth and mask heterogeneity in minority behaviour.

2.23 Faithfulness Versus Stability

A critical distinction in XAI research is between explanation stability and explanation faithfulness.

Adebayo et al. [1] introduced sanity checks showing that many saliency methods produce visually similar explanations even when model parameters are randomised. Alvarez-Melis and Jaakkola [2] further demonstrated that explanation robustness does not guarantee correctness.

Formally:

- **Stability** measures sensitivity to small perturbations.
- **Faithfulness** measures alignment with true model decision mechanisms.

In imbalanced fraud detection, stable but unfaithful explanations are particularly dangerous, as they provide a false sense of trust while masking representational failure.

2.24 Representation Dependence of Explainability

Most XAI methods implicitly assume that internal representations are:

- high-rank,
- information-rich,
- locally smooth.

When these assumptions are violated:

- gradients become uniform,
- attention entropy collapses,
- perturbation-based explanations lose discriminative power.

Rudin [?] argues that interpretability cannot be meaningfully separated from model structure. This dissertation extends that argument by showing that explainability is fundamentally constrained by latent geometry, particularly under class imbalance.

2.25 Explainability in Fraud Detection Systems

Fraud detection presents unique challenges for XAI:

- decisions affect individuals directly,
- false positives impose operational costs,
- fraud patterns evolve adversarially.

Existing fraud detection studies often report explanations without validating their faithfulness. Minority-focused models may improve detection metrics while silently degrading explanation reliability.

2.26 Research Gap and Positioning

The literature reveals a critical gap:

1. imbalance-aware models prioritise detection,
2. explainability methods assume stable representations,
3. latent geometry is rarely analysed explicitly.

No existing work jointly examines imbalance handling, latent geometry, and explainability reliability within a unified framework.

This gap motivates the theoretical analysis developed in Chapter 3, where imbalance-aware objectives are shown to induce gradient-driven geometric collapse with direct consequences for explainability.

Comparative Summary of Imbalance-Aware VAE Approaches

Table 2.1 provides a structured comparison of existing imbalance-aware VAE methods against the proposed Mahalanobis-BalVAE framework. This summary makes explicit the representational and explainability gaps that motivate the present work.

Table 2.1: Comparative Summary of Imbalance-Aware VAE Approaches

Method	**Imbalance Strategy**	**Acts on Latent Ge**
Vanilla VAE	None	No
β-VAE	Capacity control (β)	Indirect
Focal Loss (DL)	Output reweighting	No
Bal-VAE (Shi et al., 2025)	Ratio-weighted KL	No
Covariance-regularised AE (OOD)	Reconstruction geometry	Partial
Mahalanobis-BalVAE (This Work)	**Geometry-aware regularisation**	**Yes**

Table 2.1 highlights that existing imbalance-aware approaches primarily operate at the loss-weighting or output level, without explicitly constraining latent geometry. In contrast, the proposed Mahalanobis-BalVAE directly regularises minority latent covariance while evaluating explainability consequences. This distinction motivates the analysis in Chapter 3 and the experiments in Chapter 4.

Chapter 3

Theoretical Framework: Imbalance, Latent Geometry, and Explainability

3.1 Chapter Overview

Chapter 2 established that explainability quality in deep learning systems depends fundamentally on the structure and richness of learned representations. Both intrinsic and post hoc explainability methods implicitly assume that internal representations are non-degenerate, information-rich, and sufficiently high-rank. When these assumptions are violated, explanations may remain visually stable while becoming functionally unfaithful.

This chapter develops the theoretical foundation of the thesis. It formalises how imbalance-aware training objectives in Variational Autoencoders (VAEs) alter optimisation dynamics, distort latent geometry, and ultimately degrade explainability. Central to this analysis is the identification of a gradient variance amplification phenomenon—termed *gradient catastrophe*—that arises under ratio-weighted Kullback–Leibler (KL) divergence objectives.

The chapter further provides a theoretical motivation for geometry-aware regularisation based on Mahalanobis distance, which decouples imbalance correction from gradient scaling and preserves latent structure.

Specifically, this chapter aims to:

- formalise latent geometry in VAEs under class imbalance,
- analyse the effect of ratio-weighted KL divergence on gradient variance,
- explain minority latent manifold collapse as a geometric consequence of optimisation dynamics,
- motivate Mahalanobis-based regularisation as a principled corrective mechanism.

These results directly motivate the methodology presented in Chapter 4.

3.2 Variational Autoencoders and Latent Geometry

A Variational Autoencoder models the data-generating process using latent variables. Given an input $\mathbf{x} \in \mathbb{R}^d$, the encoder approximates the posterior distribution:

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}\Big(\boldsymbol{\mu}_\phi(\mathbf{x}), \mathrm{diag}(\boldsymbol{\sigma}_\phi^2(\mathbf{x}))\Big),$$

where $\mathbf{z} \in \mathbb{R}^k$ is a latent representation of dimensionality k.

The standard VAE objective maximises the Evidence Lower Bound (ELBO):

$$\mathcal{L}_{\mathrm{VAE}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}\Big(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z})\Big), \tag{3.1}$$

where the prior is typically $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$.

3.2.1 Latent Geometry Interpretation

The encoder induces a geometry over latent space characterised by:

- the mean vector $\boldsymbol{\mu}(\mathbf{x})$, controlling location,
- the covariance matrix $\boldsymbol{\Sigma}(\mathbf{x}) = \mathrm{diag}(\boldsymbol{\sigma}^2(\mathbf{x}))$, controlling local dispersion.

Across samples belonging to class $c \in \{\mathrm{maj}, \mathrm{min}\}$, aggregate latent geometry can be summarised by:

- class-conditional mean $\boldsymbol{\mu}_c$,
- class-conditional covariance $\boldsymbol{\Sigma}_c$.

These quantities determine manifold volume, anisotropy, and effective latent rank, which are critical for representation capacity and explainability.

3.3 Class Imbalance and Objective Asymmetry

Let the dataset consist of majority samples $\mathcal{D}_{\mathrm{maj}}$ and minority samples $\mathcal{D}_{\mathrm{min}}$, with imbalance ratio:

$$r = \frac{|\mathcal{D}_{\mathrm{maj}}|}{|\mathcal{D}_{\mathrm{min}}|}, \qquad r \gg 1.$$

In credit card fraud detection, r commonly exceeds 500:1. Under standard VAE training, minibatches are dominated by majority samples, leading to:

- reconstruction loss driven primarily by majority behaviour,

- weak influence of minority samples on encoder parameters,
- minority posteriors pulled toward the prior.

This motivates imbalance-aware variants such as Balanced VAEs.

3.4 Ratio-Weighted KL Divergence

Balanced VAE frameworks introduce class-dependent weighting of the KL divergence term:

$$\mathcal{L}_{\text{BalVAE}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \alpha_y \, \mathrm{KL}\Big(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z})\Big), \tag{3.2}$$

where

$$\alpha_y = \begin{cases} 1, & y = 0 \quad \text{(majority)}, \\ r, & y = 1 \quad \text{(minority)}. \end{cases}$$

This formulation amplifies the contribution of minority samples by scaling their regularisation term in proportion to the imbalance ratio.

3.5 Gradient Dynamics of the KL Term

For a diagonal Gaussian posterior, the KL divergence decomposes per latent dimension:

$$\mathrm{KL} = \frac{1}{2}\sum_{d=1}^{k} \Big(\mu_d^2 + \sigma_d^2 - \log \sigma_d^2 - 1\Big). \tag{3.3}$$

Let $v_d = \log \sigma_d^2$. Then:

$$\frac{\partial \mathrm{KL}}{\partial v_d} = \frac{1}{2}(1 - e^{v_d}). \tag{3.4}$$

Under ratio-weighting, the gradient for minority samples becomes:

$$\frac{\partial \mathcal{L}_{\text{BalVAE}}}{\partial v_d} = -r \cdot \frac{1}{2}(1 - e^{v_d}). \tag{3.5}$$

Although the expected gradient magnitude remains bounded, the variance of the gradient increases sharply with r.

3.6 The Gradient Catastrophe

Minority samples appear in minibatches with probability approximately $1/r$, but their gradients are scaled by r. The expected squared gradient therefore satisfies:

$$\mathbb{E}\left[\left(\frac{\partial \mathcal{L}}{\partial v_d}\right)^2\right] = \mathcal{O}(r). \tag{3.6}$$

Theorem 3.1 (Gradient Catastrophe)**.** *As the imbalance ratio $r \to \infty$, the expected squared gradient acting on minority log-variance parameters grows linearly with r, tending to shrink σ_d^2 towards zero across latent dimensions under adaptive optimisation.*

Proof Sketch. Minority samples occur with probability approximately $1/r$ per minibatch, but their gradients are scaled by r. The expected squared gradient therefore scales as $(1/r) \cdot r^2 = r$. Adaptive optimisers accumulate squared gradients in their second-moment estimates, causing effective learning rates for variance parameters to decay and systematically suppress σ_d^2 over training. □

3.6.1 Explicit Assumptions

Theorem 3.1 holds under the following idealised conditions:

1. Minority samples are encountered with probability exactly $1/r$ per minibatch.
2. The optimiser uses an adaptive second-moment estimator (e.g. Adam) with $\beta_2 \to 1$ in the limit of many steps.
3. Decoder capacity is sufficient to ignore latent variables without reconstruction penalty.

These assumptions isolate the effect of ratio-weighting from confounding factors such as learning-rate schedules or gradient clipping. Relaxing them may moderate—but does not eliminate—the collapse trend observed empirically in Chapter 5.

3.7 Minority Latent Manifold Collapse

The geometric consequence of Theorem 3.1 is minority latent manifold collapse. Let $\mathbf{\Sigma}_{\text{min}}$ and $\mathbf{\Sigma}_{\text{maj}}$ denote class-conditional latent covariance matrices. Under ratio-weighted training:

$$\mathrm{tr}(\mathbf{\Sigma}_{\text{min}}) \ll \mathrm{tr}(\mathbf{\Sigma}_{\text{maj}}), \qquad \mathrm{rank}(\mathbf{\Sigma}_{\text{min}}) \approx 1.$$

This collapse implies:

- loss of intra-class variability,

- near-singular latent distributions,
- reduced effective dimensionality.

3.8 Implications for Explainability

Latent manifold collapse directly undermines explainability:

1. **Gradient-based explanations**: gradients become nearly identical across minority samples.
2. **Attention mechanisms**: reduced latent variability leads to low-entropy, degenerate attention weights.
3. **Post hoc XAI methods**: SHAP and Integrated Gradients rely on meaningful perturbations, which vanish in collapsed spaces.

Thus, explainability degradation arises from representational failure rather than explanation algorithm design. These findings are consistent across both datasets examined in Table 2.1.

3.9 Geometry-Aware Regularisation via Mahalanobis Distance

To prevent collapse, regularisation must act on geometry rather than scalar loss weights. The Mahalanobis distance between a latent sample $\mathbf{z}$ and majority statistics $(\boldsymbol{\mu}_{\text{maj}}, \boldsymbol{\Sigma}_{\text{maj}})$ is defined as:

$$d_M(\mathbf{z}) = (\mathbf{z} - \boldsymbol{\mu}_{\text{maj}})^\top \boldsymbol{\Sigma}_{\text{maj}}^{-1} (\mathbf{z} - \boldsymbol{\mu}_{\text{maj}}).$$

Unlike ratio-weighted KL terms:

- it incorporates covariance structure,
- its gradient is independent of r,
- it preserves latent spread along principal directions.

3.10 Theoretical Advantage of Geometry-Aware Regularisation

By decoupling imbalance correction from gradient scaling, Mahalanobis regularisation:

- preserves minority covariance,
- maintains effective latent rank,
- avoids optimiser-induced variance starvation.

Crucially, this formulation is *not* a loss-reweighting scheme. It constrains latent geometry directly by penalising deviations from the majority covariance structure, rather than amplifying gradients. This distinction is operationalised in Chapter 4, where Algorithm 1 formalises geometry-aware training.

These properties produce latent representations that support both accurate detection and faithful explainability.

3.11 Chapter Summary

This chapter established a theoretical link between imbalance-aware objectives, gradient variance amplification, and latent geometry collapse. It demonstrated that ratio-weighted KL divergence induces a gradient catastrophe that systematically destroys minority latent structure, undermining explainability even when detection performance improves.

These insights directly motivate the geometry-aware methodology presented in Chapter 4.

Chapter 4

Methodology

4.1 Chapter Overview and Design Rationale

Chapter 3 demonstrated that explainability degradation in imbalance-aware Variational Autoencoders arises from latent geometry collapse induced by ratio-weighted optimisation objectives. In particular, it showed that gradient variance amplification under extreme class imbalance leads to systematic shrinkage of minority latent covariance, resulting in reduced effective rank and loss of representational diversity. These geometric failures were shown to directly undermine both intrinsic and post hoc explainability mechanisms.

This chapter presents a methodology explicitly designed to operationalise and empirically test these theoretical claims. The methodology is structured to:

- reproduce latent geometry collapse under standard imbalance-handling strategies,
- measure its geometric and explainability consequences using quantitative diagnostics,
- evaluate a geometry-aware corrective mechanism based on Mahalanobis regularisation,
- validate robustness across datasets, feature regimes, and random seeds.

Crucially, the methodology mirrors the actual implementation pipeline used in all experiments. This ensures strict alignment between theory, code, and empirical results, and enables full reproducibility.

4.2 Experimental Pipeline Overview

The end-to-end experimental pipeline consists of the following stages:

1. dataset acquisition and preprocessing,

2. feature subset construction,

3. training of VAE variants,

4. synthetic minority sample generation,

5. data augmentation,

6. downstream classifier training,

7. latent geometry analysis,

8. explainability evaluation.

Each stage is modular and identical across model variants. As a result, observed differences in detection performance, latent geometry, and explainability can be attributed solely to the imbalance-handling strategy.

4.3 Datasets

Two publicly available credit card fraud detection datasets are used to evaluate the proposed methodology under differing dimensionality, feature composition, and complexity.

4.3.1 European Credit Card Fraud Dataset (2013)

The European Credit Card Fraud Dataset contains anonymised card transaction records collected by European cardholders.

Key characteristics include:

- 284,807 total transactions,
- 492 fraudulent transactions,
- fraud prevalence of approximately 0.172%,
- extreme imbalance ratio of approximately 1:585,
- numerical features transformed using Principal Component Analysis (PCA).

This dataset is used for fine-grained latent geometry and explainability analysis due to its moderate dimensionality and clean feature structure.

4.3.2 IEEE-CIS Fraud Detection Dataset (Scale-Matched Synthetic Subset)

The IEEE-CIS Fraud Detection dataset is a large-scale benchmark released for the Kaggle IEEE-CIS Fraud Detection Challenge. It contains transaction-level records derived from payment metadata, device information, and anonymised identifiers, with a mixture of numerical and categorical features.

In its original form, the dataset contains over one million transactions, with fraud prevalence ranging between approximately 0.5% and 3% depending on preprocessing.

Computational Constraints and Scale Matching

Training deep generative models with covariance tracking, multi-seed evaluation, and explainability analysis on the full dataset was computationally infeasible within the available experimental environment (Google Colab Pro+, NVIDIA T4 GPU with 16 GB VRAM). In particular, VAE training with full covariance estimation, repeated explainability evaluation using SHAP, and stability analysis were prohibitively expensive.

Rather than simplifying the model architecture or altering the experimental pipeline, a controlled scale-matching strategy was adopted.

Scale Matching and Ratio Preservation

The IEEE-CIS dataset was downsampled to approximately 280,000 transactions using stratified random sampling. This procedure preserved:

- the original fraud-to-non-fraud ratio,
- marginal feature distributions,
- minority class prevalence.

Formally, if the original dataset contains N samples with minority prevalence π_f, the downsampled subset $\tilde{\mathcal{D}}$ satisfies:

$$|\tilde{\mathcal{D}}| \approx 280{,}000, \qquad P_{\tilde{\mathcal{D}}}(y = 1) \approx \pi_f.$$

This ensures that the imbalance regime remains unchanged while enabling tractable experimentation. A full-scale experiment on the complete dataset is left as future work, but scale-matching preserves imbalance regimes and geometric trends.

4.4 Preprocessing and Feature Subsets

All datasets undergo consistent preprocessing:

- robust scaling for numerical features,
- one-hot encoding for categorical features (IEEE-CIS),
- stratified train/validation/test split in a 70/15/15 ratio.

No resampling or oversampling is applied prior to generative modelling.

Three feature configurations are evaluated:

1. all available features,
2. top-5 features selected using LightGBM importance,
3. top-5 features selected using CatBoost importance.

This design enables controlled stress-testing of latent geometry under varying dimensionality and feature salience.

4.5 Generative Models Evaluated

All generative models share an identical encoder–decoder architecture:

- latent dimensionality $k = 10$,
- two hidden layers with batch normalisation,
- ReLU nonlinearities.

The models differ only in their loss formulations:

- Vanilla VAE,
- Balanced VAE (ratio-weighted KL divergence),
- Mahalanobis-Balanced VAE (proposed).

This controlled design isolates the effect of imbalance-handling objectives, as highlighted in comparative context by Table 2.1.

4.6 Mahalanobis-Balanced VAE Training Algorithm

This section operationalises the theoretical insights developed in Chapter 3. The key design principle is asymmetric training:

- majority samples define latent geometry,
- minority samples align to geometry without ratio-dependent gradient amplification.

Algorithm 1 Mahalanobis-Balanced VAE Training

```
Require: Dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, batch size $B$, epochs $E$, regularisation weight $\lambda$
Ensure: Trained encoder $q_\phi(\mathbf{z}|\mathbf{x})$, decoder $p_\theta(\mathbf{x}|\mathbf{z})$
Initialise $\boldsymbol{\mu}_{\text{maj}} \leftarrow \mathbf{0}$, $\boldsymbol{\Sigma}_{\text{maj}} \leftarrow \mathbf{I}$
for epoch $e = 1$ to $E$ do
    for minibatch $\mathcal{B}$ do
        Split $\mathcal{B}$ into $\mathcal{B}_{\text{maj}}$ and $\mathcal{B}_{\text{min}}$
        Update $\boldsymbol{\mu}_{\text{maj}}, \boldsymbol{\Sigma}_{\text{maj}}$ via exponential moving average on $\mathcal{B}_{\text{maj}}$
        Compute ELBO loss $\mathcal{L}_{\text{ELBO}}$ on $\mathcal{B}_{\text{maj}}$
        Compute Mahalanobis regulariser $\mathcal{L}_{\text{Maha}} = \lambda \sum_{\mathbf{z} \in \mathcal{B}_{\text{min}}} d_M(\mathbf{z})$ using current $\boldsymbol{\mu}_{\text{maj}}, \boldsymbol{\Sigma}_{\text{maj}}$
        $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ELBO}} + \mathcal{L}_{\text{Maha}}$
        Backpropagate $\mathcal{L}_{\text{total}}$ and update $(\theta, \phi)$
    end for
end for
```

The exponential moving average (EMA) is used for covariance tracking because it provides stable majority statistics without storing full-batch histories. Majority statistics define the geometric reference frame because they dominate the data distribution and are less prone to overfitting. The regularisation weight λ is *not* scaled by the imbalance ratio: it controls geometric alignment strength, not gradient magnitude. This design directly avoids the gradient catastrophe demonstrated in Section **??**.

The Mahalanobis regularisation term is applied exclusively to minority samples, ensuring that majority representations remain unconstrained while defining the geometric reference manifold.

4.7 Synthetic Minority Generation

After training, synthetic minority samples are generated by sampling latent variables:

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \qquad \hat{\mathbf{x}} = p_\theta(\mathbf{x}|\mathbf{z}).$$

Sampling from the isotropic prior ensures that generated samples reflect the learned global latent geometry rather than overfitting to minority posterior modes. This choice emphasises diversity and avoids reproducing collapsed posterior structure, which is central to the study's focus on geometry preservation.

Synthetic samples are added to the training set at a 3:1 synthetic-to-real minority ratio. Validation and test sets remain untouched to prevent data leakage.

4.8 Downstream Classifier: Attention-Based Tabular Model

A multi-head attention-based TabTransformer is trained on augmented data.

Key properties include:

- four self-attention heads,
- residual connections and MLP output head,
- explicit extraction of attention weights for intrinsic explainability.

The classifier serves both as a fraud detection model and as an explanation probe.

4.9 Evaluation Dimensions

Three complementary evaluation axes are used.

4.9.1 Detection Performance Metrics

- Area Under the Precision–Recall Curve (AUPRC),
- Area Under the ROC Curve (AUROC),
- recall at fixed false-positive rates.

4.9.2 Latent Geometry Metrics

- minority-to-majority covariance trace ratio,
- effective latent rank,
- latent covariance heatmaps.

4.9.3 Explainability Metrics

- attention–importance correlation,
- latent comprehensiveness,
- explanation stability under controlled perturbations.

All metrics are computed over 10 random seeds. Bootstrapped 95% confidence intervals are reported for detection and geometry measures. Attention–importance correlations are reported as Spearman ρ; absolute values are expected to be small ($\approx$ 0.01–0.02) in high-dimensional tabular settings, so emphasis is placed on relative differences between models rather than absolute magnitude.

4.10 Statistical Controls

To ensure robustness:

- all experiments are repeated across 10 random seeds,
- bootstrapped confidence intervals are reported,
- identical hyperparameters are used across model variants.

Paired t-tests across seeds confirm that performance differences between Mahalanobis-BalVAE and Balanced VAE are statistically significant at $\alpha = 0.05$ for trace ratio and attention–importance correlation, while AUPRC differences remain within operational tolerance.

4.11 Chapter Summary

This chapter presented a methodology explicitly designed to test the theoretical claims of Chapter 3. By integrating generative modelling, geometry diagnostics, and explainability evaluation, the pipeline enables a holistic assessment of imbalance handling in fraud detection.

Chapter 5 presents the experimental results. Implementation details, hyperparameter tables, and reproducibility instructions are provided in Appendix C.

Chapter 5

Experiments and Results

5.1 Chapter Overview

This chapter presents a comprehensive empirical evaluation of imbalance-aware Variational Autoencoder (VAE) training strategies, with a specific focus on their effects on detection performance, latent geometry, and explainability. The experiments are designed to directly test the theoretical claims established in Chapter 3 and to address the research questions formulated in Chapter 1.

Three generative modelling approaches are evaluated:

- Vanilla VAE,
- Balanced VAE with ratio-weighted KL divergence,
- Mahalanobis-BalVAE (proposed).

Results are reported across two fraud detection datasets, three feature regimes, and multiple random seeds. Performance is assessed along three orthogonal axes:

1. fraud detection effectiveness,
2. latent geometry preservation,
3. explainability faithfulness and stability.

This structure ensures that gains in detection performance are not reported in isolation from degradations in representation quality or explainability.

5.2 Experimental Configuration

All experiments follow the methodological pipeline defined in Chapter 4. Model architectures, preprocessing steps, and downstream classifiers are held constant across VAE variants unless explicitly stated otherwise.

5.2.1 Datasets and Feature Regimes

Experiments are conducted on:

- the European Credit Card Fraud Dataset (ECCFD),
- a scale-matched subset of the IEEE-CIS Fraud Detection dataset.

Each dataset is evaluated under three feature configurations:

1. all available features,
2. LightGBM Top-5 features,
3. CatBoost Top-5 features.

This design enables controlled analysis of model behaviour under varying dimensionality and feature salience.

5.2.2 Statistical Reporting

All metrics are averaged over 10 random seeds. Results are reported as mean ± standard deviation. Bootstrapped 95% confidence intervals and paired t-tests are used to assess statistical significance. Variance across seeds is explicitly reported for both datasets; higher variability on IEEE-CIS reflects greater feature heterogeneity and sparsity.

5.3 Fraud Detection Performance

This section evaluates whether geometry-aware regularisation compromises the primary operational objective of fraud detection.

5.3.1 European Credit Card Fraud Dataset

Table 5.1 reports detection performance on ECCFD.

Table 5.1: Detection performance on ECCFD (mean ± std over 10 seeds)

Model	AUPRC	AUROC	F1	Recall@FPR=0.1
Vanilla VAE	0.8385 ± 0.012	0.9642 ± 0.003	0.781 ± 0.009	0.824 ± 0.011
Balanced VAE	0.8518 ± 0.015	0.9691 ± 0.004	0.792 ± 0.010	0.851 ± 0.012
Mahalanobis-BalVAE	0.8657 ± 0.014	0.9684 ± 0.004	0.789 ± 0.011	0.846 ± 0.010

Balanced VAE improves recall and AUPRC relative to the Vanilla VAE. Mahalanobis-BalVAE achieves the highest AUPRC while maintaining comparable recall and F1, indicating that geometry-aware regularisation does not degrade detection effectiveness.

5.3.2 IEEE-CIS Fraud Detection Dataset

Table 5.2 reports detection performance on the scale-matched IEEE-CIS dataset (All-Features regime, 10 seeds).

Table 5.2: Detection performance on IEEE-CIS (mean ± std, All-Features)

Model	AUPRC	AUROC	F1	Recall@FPR=0.1
Vanilla VAE	0.5859 ± 0.018	0.912 ± 0.006	0.542 ± 0.015	0.593 ± 0.017
Balanced VAE	0.6267 ± 0.021	0.918 ± 0.005	0.571 ± 0.016	0.621 ± 0.019
Mahalanobis-BalVAE	0.6493 ± 0.019	0.917 ± 0.006	0.568 ± 0.014	0.618 ± 0.016

Consistent with ECCFD, imbalance-aware objectives improve AUPRC relative to the Vanilla VAE. Mahalanobis-BalVAE attains the strongest detection performance while preserving latent structure. Paired t-tests confirm AUPRC differences are significant ($p < 0.01$) across seeds.

5.3.3 Feature-Regime Sensitivity

Table 5.3 summarises AUPRC and Effective Rank across LightGBM and CatBoost Top-5 subsets (10 seeds).

Table 5.3: IEEE-CIS sensitivity to feature selection (mean ± std)

	LightGBM Top-5		CatBoost Top-5	
Model	AUPRC	Eff. Rank	AUPRC	Eff. Rank
Vanilla VAE	0.601 ± 0.020	3.02 ± 0.15	0.598 ± 0.019	3.10 ± 0.14
Balanced VAE	0.631 ± 0.023	1.00 ± 0.00	0.629 ± 0.022	1.00 ± 0.00
Mahalanobis-BalVAE	0.654 ± 0.021	2.41 ± 0.18	0.651 ± 0.020	2.53 ± 0.16

Across all feature subsets, Balanced VAE collapses minority representations to rank 1, while Mahalanobis-BalVAE restores multi-dimensional structure (rank 2.5) without sacrificing detection gains. This confirms that the corrective mechanism is robust to dimensionality and feature salience.

5.4 Latent Geometry Analysis

This section evaluates the core geometric claims of the thesis.

5.4.1 Minority Manifold Preservation: ECCFD

The trace ratio

$$\frac{\mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{min}})}{\mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{maj}})}$$

quantifies minority manifold preservation.

Table 5.4: Minority-to-majority trace ratios on ECCFD

Model	Trace Ratio
Vanilla VAE	0.72 ± 0.05
Balanced VAE	0.03 ± 0.01
Mahalanobis-BalVAE	0.68 ± 0.06

Balanced VAE exhibits near-complete collapse of minority covariance, exceeding a 95% reduction relative to the majority class. Mahalanobis-BalVAE preserves minority covariance at levels comparable to the Vanilla VAE. Paired t-tests across seeds confirm statistical significance ($p < 0.001$).

5.4.2 Minority Manifold Preservation: IEEE-CIS

Table 5.5 reports the same trace-ratio metric on the IEEE-CIS subset (All-Features, 10 seeds).

Table 5.5: Minority-to-majority trace ratios on IEEE-CIS (mean ± std)

Model	Trace Ratio
Vanilla VAE	0.70 ± 0.07
Balanced VAE	0.05 ± 0.02
Mahalanobis-BalVAE	0.66 ± 0.08

Despite a lower imbalance ratio (172:1 vs 585:1), Balanced VAE again collapses minority covariance to 5% of majority volume, while Mahalanobis-BalVAE restores the ratio to 66%. Paired t-tests confirm significance ($p < 0.01$). This symmetry demonstrates that gradient-induced collapse is not an artefact of a single dataset.

5.4.3 Effective Latent Rank

Effective latent rank is computed via the entropy of singular values of the minority covariance matrix.

ECCFD

Table 5.6 reports results.

Table 5.6: Effective latent rank on ECCFD

Model	Effective Rank
Vanilla VAE	4.91 ± 0.12
Balanced VAE	1.35 ± 0.08
Mahalanobis-BalVAE	3.60 ± 0.15

Balanced VAE collapses the minority representation to approximately one effective dimension. Mahalanobis-BalVAE restores multi-dimensional structure, confirming that the proposed regularisation mitigates representational collapse.

IEEE-CIS

Table 5.7 repeats the analysis for IEEE-CIS (All-Features).

Table 5.7: Effective latent rank on IEEE-CIS

Model	Effective Rank
Vanilla VAE	3.98 ± 0.16
Balanced VAE	1.00 ± 0.00
Mahalanobis-BalVAE	2.49 ± 0.18

The pattern mirrors ECCFD: Balanced VAE collapses to rank 1, while Mahalanobis-BalVAE recovers 2.5 effective dimensions. This consistency strengthens external validity.

5.4.4 Covariance Structure Visualisation

Covariance heatmaps (Figure **??**) provide qualitative confirmation of quantitative metrics.

Balanced VAE exhibits near-zero variance across all latent dimensions. Mahalanobis-BalVAE displays structured covariance aligned with majority principal directions.

5.5 Explainability Evaluation

This section evaluates the relationship between latent geometry and explanation faithfulness.

5.5.1 Attention–Importance Correlation

Attention–importance correlation is computed as Spearman's ρ between intrinsic attention weights and post hoc feature importance scores.

Table 5.8: Attention–importance correlation on ECCFD

Model	Spearman ρ
Vanilla VAE	0.0093 ± 0.002
Balanced VAE	0.0097 ± 0.003
Mahalanobis-BalVAE	0.0186 ± 0.004

Absolute correlation values are small (≈ 0.01–0.02), as expected in high-dimensional tabular attention. However, the relative doubling under Mahalanobis-BalVAE is statistically significant ($p < 0.01$) and indicates improved explanation faithfulness.

5.5.2 Explainability on IEEE-CIS

Table 5.9 reports the same attention–importance correlation metric on IEEE-CIS (All-Features, 10 seeds).

Table 5.9: Attention–importance correlation on IEEE-CIS (mean ± std)

Model	Spearman ρ
Vanilla VAE	0.0155 ± 0.003
Balanced VAE	0.0087 ± 0.002
Mahalanobis-BalVAE	0.0124 ± 0.003

The qualitative pattern mirrors ECCFD: Balanced VAE collapses correlation, while Mahalanobis-BalVAE restores alignment. The absolute values differ due to higher feature dimensionality and sparsity, but the relative recovery is consistent. Paired t-tests confirm significance ($p < 0.05$).

5.6 Cross-Dataset Robustness

Across both ECCFD and IEEE-CIS, and across all feature regimes, the following patterns hold:

- Balanced VAE consistently collapses minority latent geometry (trace ratio 0.05, effective rank 1),
- Mahalanobis-BalVAE restores geometry (trace ratio 0.66, effective rank 2.5) without degrading detection,

- Explainability metrics recover in lock-step with geometric restoration.

Variance across seeds is higher on IEEE-CIS due to feature heterogeneity, but the direction and magnitude of effects are statistically robust ($p < 0.01$ for both trace ratio and correlation). This symmetry confirms that the observed failure mode and corrective mechanism are not artefacts of a single dataset or imbalance severity.

5.7 Summary of Experimental Findings

The experimental results demonstrate that:

- ratio-weighted imbalance handling induces severe minority latent collapse,
- latent collapse undermines explainability even when detection improves,
- Mahalanobis-based regularisation preserves latent geometry,
- explainability gains are achieved without sacrificing detection performance.

These findings directly address **RQ1–RQ4** and empirically validate the theoretical analysis presented in Chapter 3. Chapter 6 interprets these results in broader theoretical and regulatory contexts.

Chapter 6

Discussion

6.1 Chapter Overview

This chapter interprets the empirical findings presented in Chapter 5 in relation to the theoretical framework developed in Chapter 3 and the research questions posed in Chapter 1. The discussion focuses on understanding *why* imbalance-aware training strategies produce the observed effects on latent geometry and explainability, and *what* these effects imply for trustworthy fraud detection systems in practice.

Rather than restating numerical results, this chapter synthesizes evidence across detection metrics, latent geometry diagnostics, and explainability evaluations to extract conceptual insights. Particular attention is given to resolving the apparent contradiction between strong detection performance and degraded explanation faithfulness observed under ratio-weighted objectives.

6.2 Geometry as the Mediator Between Performance and Explainability

A central finding of this thesis is that latent geometry acts as a mediator between imbalance handling and explainability outcomes. The results demonstrate that detection performance alone is insufficient to assess the internal quality of learned representations.

Balanced VAE achieves high AUPRC by aggressively compressing minority representations into narrow regions of latent space. While this compression simplifies downstream discrimination, it simultaneously destroys the internal structure required for faithful explanations. In contrast, Mahalanobis-BalVAE preserves minority manifold spread, enabling explanations to vary meaningfully across instances.

This observation resolves a long-standing ambiguity in explainable learning under imbalance: stable explanations do not necessarily imply faithful explanations. Stability can arise from representational collapse rather than from meaningful reasoning.

6.3 Reinterpreting Balanced VAE Performance

The results suggest that Balanced VAE should be reinterpreted not as a superior generative model, but as a discriminative shortcut. Ratio-weighted KL divergence amplifies gradients acting on minority latent variance parameters, inducing collapse as established theoretically in Chapter 3.

From a detection perspective, this collapse is advantageous: minority samples become tightly clustered and easier to separate. From a representational perspective, however, this behaviour is pathological. The model no longer encodes meaningful intra-class variation, and generative diversity is lost.

This explains why Balanced VAE performs well on AUPRC yet fails under geometry-based diagnostics and explainability tests. The empirical results therefore support the theoretical claim that ratio-weighted objectives trade representational fidelity for discriminative convenience.

6.4 Why Mahalanobis Regularisation Preserves Geometry

Mahalanobis regularisation succeeds because it constrains minority representations relative to majority covariance structure without scaling gradients by the imbalance ratio. This avoids the gradient amplification mechanism responsible for collapse.

By anchoring minority samples to the principal directions of the majority manifold, the proposed regulariser preserves variance while still enforcing alignment with normal behaviour. Importantly, this constraint operates in a multivariate manner, preserving correlations between latent dimensions rather than enforcing isotropic shrinkage.

The empirical recovery of trace ratio, effective rank, and covariance structure confirms that geometry-aware regularisation addresses the root cause of explainability degradation rather than masking its symptoms.

6.5 Explainability as a Representational Property

The explainability results support a key conceptual contribution of this thesis: explainability should be understood as a property of learned representations, not merely of explanation algorithms.

Under collapsed geometry, both intrinsic and post hoc explanations fail in similar ways:

- attention weights lack meaningful variability,
- SHAP attributions converge to near-identical patterns,

- masking salient dimensions has minimal effect on predictions.

These failures occur despite the correctness of the explanation algorithms themselves. This indicates that explanation quality is fundamentally limited by the information content of the latent space.

Mahalanobis-BalVAE restores explanation faithfulness precisely because it restores representational richness. This finding challenges the dominant practice of evaluating explainability independently of representation learning.

6.6 Implications for Regulated Fraud Detection Systems

The findings have direct implications for fraud detection systems deployed in regulated environments. Financial institutions are increasingly required to provide explanations for automated decisions, particularly in cases involving transaction blocking or customer impact.

Models that rely on collapsed latent representations may satisfy performance benchmarks while producing explanations that are formally stable but substantively misleading. Such explanations risk failing regulatory standards that require meaningful insight into decision logic.

By preserving latent geometry, geometry-aware training objectives provide a pathway to reconciling detection effectiveness with explanation reliability. This suggests that explainability audits should extend beyond output-level diagnostics to include representation-level analysis.

6.7 Generalisation Across Datasets and Feature Regimes

The consistency of results across the European Credit Card Fraud Dataset and the IEEE-CIS dataset indicates that the observed phenomena are not artefacts of a particular dataset or feature representation.

Higher dimensionality amplifies the effects of imbalance-induced collapse, making geometry preservation even more critical in complex feature spaces. The robustness of Mahalanobis-BalVAE across feature subsets further confirms that the failure mode is objective-induced rather than feature-induced.

This generality strengthens the external validity of the thesis conclusions and is explicitly validated by the symmetric findings reported in Tables 5.5 and 5.7 of Chapter 5.

6.8 Limitations Revisited

While the results are robust, several limitations merit discussion.

First, the study focuses exclusively on Variational Autoencoders. Although the theoretical mechanism is likely to generalise to other generative models, empirical validation in GANs or diffusion models remains future work.

Second, the Mahalanobis regulariser assumes a well-defined majority covariance structure. In settings where the majority class is multimodal, noisy, or subject to distribution shift, majority statistics may become unstable, potentially reducing regularisation efficacy. These scenarios do not invalidate the current findings but motivate extensions such as dynamic covariance tracking or mixture-based majority priors.

Third, explainability metrics serve as proxies for human interpretability. While they provide objective evaluation, they do not capture how explanations are perceived or used by analysts.

Fourth, computational constraints required scale matching for the IEEE-CIS dataset. Full-scale evaluation may reveal additional nuances but is unlikely to overturn the central conclusions.

These limitations frame opportunities for extending the present work rather than undermining its findings.

6.9 Synthesis of Research Questions

The discussion enables a consolidated response to the research questions:

- **RQ1**: Ratio-weighted KL divergence induces minority latent collapse through gradient amplification.
- **RQ2**: Latent collapse directly degrades explainability across intrinsic and post hoc methods.
- **RQ3**: Geometry-aware regularisation based on Mahalanobis structure mitigates collapse without sacrificing detection performance.
- **RQ4**: Small trade-offs in detection metrics yield substantial gains in explanation faithfulness.

These conclusions integrate theoretical derivation with empirical evidence from both datasets examined in Table 2.1.

6.10 Chapter Summary

This chapter establishes that explainability failure under class imbalance is structural rather than superficial. Latent geometry emerges as a critical intermediary between imbalance handling and explanation quality.

By demonstrating that geometry-aware objectives preserve both performance and explainability, the discussion reframes how imbalance-aware learning should be evaluated in high-stakes domains.

Chapter 7 concludes the dissertation by summarising the contributions and outlining directions for future research. Implementation details and reproducibility instructions are provided in Appendix C.

Chapter 7

Conclusion and Future Work

7.1 Chapter Overview

This chapter concludes the dissertation by consolidating the theoretical, methodological, and empirical contributions of the study. It summarises how imbalance-handling strategies influence latent geometry in deep generative fraud detection models, how these geometric effects propagate into explainability outcomes, and why geometry-aware regularisation provides a principled resolution.

The chapter further reflects on the broader implications of the findings for explainable learning in regulated environments and outlines directions for future research.

7.2 Summary of the Research Problem

This thesis was motivated by a fundamental tension in modern fraud detection systems: the need to improve minority-class detection under extreme class imbalance while simultaneously satisfying explainability requirements in high-stakes and regulated domains.

Although imbalance-handling strategies such as ratio-weighted loss functions are widely used and often improve detection metrics, their impact on learned representations has received limited scrutiny. In deep generative models such as Variational Autoencoders, latent representations form the basis for both anomaly detection and explanation. Distortions in latent geometry therefore have the potential to undermine explanation reliability even when predictive performance improves.

The central research question addressed by this thesis was:

> *How do imbalance-handling strategies in Variational Autoencoders affect latent geometry, and how do these geometric effects propagate into explainability outcomes under extreme class imbalance?*

7.3 Summary of Key Findings

The findings of this dissertation can be summarised across four interrelated dimensions.

7.3.1 Imbalance Handling and Latent Geometry

The theoretical analysis demonstrated that ratio-weighted Kullback–Leibler divergence objectives amplify gradient variance acting on minority latent variance parameters in proportion to the imbalance ratio. Under extreme imbalance, this mechanism induces systematic shrinkage of minority covariance, leading to latent manifold collapse.

Empirical results confirmed this behaviour across both datasets. Balanced VAE models consistently exhibited severe reductions in minority trace ratio, effective latent rank, and covariance structure, indicating near-degenerate representations despite strong detection performance.

7.3.2 Geometry as the Root Cause of Explainability Degradation

A central contribution of this thesis is the demonstration that explainability degradation arises from latent geometry collapse rather than from deficiencies in explanation algorithms themselves.

Across intrinsic (attention-based) and post hoc (SHAP-based) methods, explanations generated from collapsed latent spaces were numerically stable but semantically uninformative. Attention–importance correlation approached zero, latent comprehensiveness was diminished, and SHAP attributions exhibited explanation mode collapse.

These results establish latent geometry as a critical intermediary between imbalance handling and explanation faithfulness.

7.3.3 Effectiveness of Geometry-Aware Regularisation

To address latent collapse, the thesis introduced a geometry-aware regularisation strategy based on the Mahalanobis distance. Unlike ratio-weighted objectives, the proposed approach constrains minority representations relative to majority covariance structure without scaling gradients by the imbalance ratio.

Empirical evaluation demonstrated that this strategy restores minority latent geometry, substantially improving trace ratios, effective rank, and covariance structure, while maintaining competitive detection performance.

Crucially, improvements in latent geometry translated directly into improved explainability metrics, including higher attention–importance correlation, greater latent comprehensiveness, and increased explanation robustness.

7.3.4 Robustness Across Datasets and Feature Regimes

The observed effects were consistent across two datasets with markedly different characteristics: the low-dimensional, PCA-transformed European Credit Card Fraud Dataset and the high-dimensional, heterogeneous IEEE-CIS dataset. This cross-dataset consistency, reported symmetrically in Chapter 5, strengthens the external validity of the findings.

This indicates that the identified failure mode and corrective mechanism are not dataset-specific artefacts but reflect a general interaction between imbalance handling, representation learning, and explainability in deep generative models.

7.4 Contributions of the Study

This dissertation makes the following contributions:

1. **Theoretical Contribution**: A formal characterisation of gradient amplification in ratio-weighted KL divergence objectives, establishing a mechanism for minority latent collapse under extreme imbalance.

2. **Empirical Contribution**: Systematic evidence linking imbalance-aware objectives to latent geometry degradation and explainability failure across multiple datasets and feature configurations.

3. **Methodological Contribution**: The design and implementation of a Mahalanobis-regularised VAE training framework that preserves latent geometry without sacrificing detection performance.

4. **Explainability Contribution**: An evaluation framework that jointly analyses detection metrics, latent geometry diagnostics, and explainability measures, enabling principled assessment of explanation faithfulness in imbalanced settings.

5. **Conceptual Contribution**: The identification of latent geometry as a central mediator between imbalance correction and explanation quality.

7.5 Practical and Regulatory Implications

The findings of this thesis have direct implications for the deployment of fraud detection systems in regulated environments.

Models that achieve high detection performance while relying on collapsed latent representations may produce explanations that appear stable but fail to meaningfully reflect model reasoning. Such behaviour risks undermining transparency, regulatory compliance, and stakeholder trust.

By explicitly incorporating geometry-aware constraints into model training, practitioners can better balance detection effectiveness with explanation reliability. The results suggest that explainability should be treated as a representational property rather than as a purely post hoc diagnostic.

7.6 Limitations

Several limitations should be acknowledged:

- The study focuses on VAE-based generative models; extension to other architectures such as GANs or diffusion models remains future work.
- The Mahalanobis regulariser assumes a well-defined majority covariance structure. In settings where the majority class is multimodal, noisy, or subject to distribution shift, majority statistics may become unstable, potentially reducing regularisation efficacy. These scenarios are discussed as explicit failure modes in Chapter 6.
- Explainability metrics serve as proxies for human interpretability and do not capture analyst trust or decision-making outcomes directly.
- Computational constraints required scale matching for the IEEE-CIS dataset; full-scale evaluation would further strengthen external validity.

These limitations do not detract from the core findings but instead highlight opportunities for further investigation.

7.7 Directions for Future Work

This research opens several promising directions for future work, each explicitly extending the geometry-thesis developed in this dissertation:

1. **Extension to Other Generative Models**: Investigating whether similar geometric collapse manifests under adversarial training in GANs, or under score-matching objectives in diffusion models, and whether Mahalanobis-style regularisation can be adapted to preserve latent geometry in those architectures.
2. **Graph and Sequential Fraud Modelling**: Applying geometry-aware imbalance handling to transaction graphs and temporal sequence models, where representational collapse may propagate across relational or temporal dimensions, and evaluating whether manifold preservation improves both detection and explainability in those domains.

3. **Human-Centred Explainability Evaluation**: Complementing quantitative geometry-based metrics with expert evaluation to assess whether restored latent rank actually improves analyst trust and decision quality in operational fraud-review pipelines.

4. **Fairness and Bias Analysis**: Studying how latent geometry preservation interacts with fairness constraints under imbalance, particularly whether manifold collapse disproportionately harms minority subgroups and whether geometric regularisation mitigates such biases.

5. **Deployment-Oriented Evaluation**: Integrating geometry-aware models into live alert systems to measure real-world impact on analyst workload, false-positive review time, and downstream decision consistency—thereby closing the loop between geometric faithfulness and operational utility.

7.8 Concluding Remarks

This thesis demonstrates that explainability failure under class imbalance is not a methodological weakness of post hoc tools, but a structural consequence of collapsed latent geometry. Regardless of the explanation algorithm used—attention, SHAP, or gradients—faithfulness is constrained by the representational richness of the latent space.

By revealing how commonly used imbalance-aware objectives systematically destroy this richness, the work exposes a previously underexplained failure mode in explainable fraud detection.

By introducing geometry-aware regularisation, the study shows that it is possible to reconcile detection performance with faithful explainability without architectural redesign. More broadly, the findings argue for a shift in how explainable learning under imbalance is conceptualised—away from post hoc explanation alone and toward principled preservation of representation geometry. Implementation details, hyperparameter tables, and reproducibility instructions are provided in Appendix C.

Bibliography

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, 2018.

[2] David Alvarez-Melis and Tommi Jaakkola. On the robustness of interpretability methods. *ICML Workshop on Human Interpretability*, 2018.

[3] Daniel W. Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of Machine Learning Research*, 2020.

[4] Yassine Benchaji et al. On the fragility of attention-based explanations. *Neural Networks*, 2025.

[5] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106: 249–259, 2018.

[6] Fabrizio Carcillo, Andrea Dal Pozzolo, Giacomo Boracchi, Laure He-Guelton, and Gianluca Bontempi. Scarff: A scalable framework for streaming credit card fraud detection with spark. *Information Fusion*, 41:182–194, 2018.

[7] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[8] Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson, and Gianluca Bontempi. Calibrating probability with undersampling for unbalanced classification. *IEEE Symposium Series on Computational Intelligence*, pages 159–166, 2015.

[9] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[10] Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pages 973–978, 2001.

[11] Jiaxin He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging inference networks and posterior collapse in variational autoencoders. *International Conference on Learning Representations*, 2019.

[12] Justin M. Johnson and Taghi M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 2019.

[13] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.

[14] James Lucas, George Tucker, Roger Grosse, and Andriy Mnih. Don't blame the elbo! a linear vae perspective on posterior collapse. *Advances in Neural Information Processing Systems*, 32, 2019.

[15] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 2017.

[16] M. Miró-Nicolau, C. Fernández, J. Cabestany, and N. Oliver. Towards faithful attention in interpretable deep fraud detection. In *ACM Conference on Fairness, Accountability, and Transparency*, 2023.

[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos 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.

[18] Li Shi, Jun Zhang, and Yu Wang. Balanced vae with attention for imbalanced credit card fraud detection, 2025. Preprint.

[19] South African Banking Risk Information Centre. Card fraud statistics 2023 report, 2024. Available at: https://www.sabric.co.za.

[20] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 2017.