AI/ML Driven Methods for Fraud Anomaly Detection

by Eva Okhankhuele

B.S. in Computer Engineering Technology, December 2004, University of Houston, Houston, Texas
M.Ed. in Instructional Design and Technology, December 2022, West Texas A&M University, Canyon, Texas

A Praxis submitted to

The Faculty of
The School of Engineering and Applied Science
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Engineering

January 9, 2026

Praxis directed by
Domingo Rivera Professorial Lecturer in Engineering and Applied Science

The School of Engineering and Applied Science of The George Washington University certifies that Okhankhuele E Eva has passed the Final Examination for the degree of Doctor of Engineering as of November 21, 2025. This is the final and approved form of the Praxis.

AI/ML-Driven Methods for Fraud Anomaly Detection

Okhankhuele E Eva

Praxis Research Committee:

Blackford Joseph, Adjunct Professor of Engineering and Applied Science, Committee Chair

Domingo Rivera, Professorial Lecturer in Engineering and Applied Science, Praxis Advisor

Morton John, Professorial Lecturer in Engineering and Applied Science, Committee Member

© Copyright 2025 by Eva Okhankhuele
All rights reserved.

Dedication

This research is dedicated to my family, whose patience and support enabled me throughout this journey, and to every mentor and colleague who encouraged me to persevere. May this research contribute, in its own small way, to building systems that protect people as much as they yield profits.

Acknowledgements

This process of Praxis required the counsel, kindness, and support of many people. I would like to offer a special thanks to my advisor, Dr. Domingo Rivera, of SEAS, to the entire faculty, and to colleagues whose mentorship has been a source of strength in this process, as well as their valuable insights, resources, and keen interest in Cybersecurity and machine learning techniques.

Also, thanks to my family and friends, whose support has been a source of strength to enable me to succeed in this journey. Your encouragement turned late nights of research into purposeful moments.

Abstract of Praxis

AI/ML-Driven Methods for Fraud Anomaly Detection

A key challenge arising from the ever-increasing instances of financial fraud is the need to develop a fraud detection model that is cost-effective, understandable, as well as private. This Praxis research develops a cost-aware and privacy-suited method for detecting fraud in imbalanced financial transaction data.

The methods combine four interdependent components:

1. Cost-sensitive evaluation, optimizing decision thresholds through a loss function $\alpha \cdot FP + \beta \cdot FN$ to reduce operational loss rather than maximize metrics.
2. Synthetic data augmentation using Conditional Tabular Generative Adversarial Networks (CTGAN) to estimate minority-class representation under auditable pre-merge gates that govern real-data precision at t^* and Cost at t^*.
3. Anomaly-feature fusion, combining unsupervised Isolation Forest and Autoencoder signals into supervised models to advance cost efficiency while maintaining recall.
4. Explainable validation, employing SHAP-based interpretability on real and synthetic validation data to support regulator-ready transparency while protecting confidentiality.

Experiments on a publicly available credit-card transaction dataset used a fixed, leakage-controlled 60/20/20 split. Results show that cost-aware thresholding substantially reduced total expected loss compared with metric-based operating points. Privacy-safe SHAP parity confirmed that synthetic validation can reproduce real-data feature-

importance structure ($\rho \approx 0.96$); and anomaly-feature fusion improved cost efficiency while preserving recall. CTGAN augmentation did not improve performance on this dataset/split, a documented negative finding that informs stricter, policy-based acceptance tests for synthetic data.

Overall, the method is transparent, auditable, and aligned with how fraud teams measure success in cost and workload, not just in scores. Future work extends the framework toward federated learning to enable multi-institutional collaboration without raw-data sharing, thus preserving cost efficiency, transparency, and privacy at scale.

Keywords: Fraud detection, cost-aware learning, CTGAN, SHAP, explainable AI, synthetic data governance, privacy-preserving analytics, financial risk management.

Table of Contents

Dedication iii
Acknowledgements iv
Abstract of Praxis v
List of Figures xiii
List of Tables xiv
List of Symbols xv
Chapter 1: Introduction 1
1.1 Background and Context 1
1.2 Problem Statement 2
1.3 Purpose of Study 3
1.4 Research Questions 4
1.5 Research Hypotheses 4
1.6 Significance of the Study 5
1.7 Scope and Delimitations 5
1.8 Expected Outcomes 6
1.9 Organization of Praxis 7
Chapter 2: Literature Review 9
2.1 Purpose and Roadmap 9
2.2 Fraud Detection as a Rare-Event and Cost Problem 10
2.2.1 Moving from Metric Scores to Operating Decisions 10
2.2.2 Practical Ingredients of Cost-Aware Policies 11
2.2.3 Implications for Model Comparison 12
2.3 Supervised Baselines and Ensemble Learners 13
2.3.1 Logistic Regression as a Reference Point 13
2.3.2 Addressing Imbalance During Training 14
2.3.3 Calibration and Threshold Transfer 15
2.3.4 Stacking and Model Fusion 15
2.4 Unsupervised Anomaly Detection 16
2.4.1 Strengths and Limits 17

2.4.2 Hybrid Pipelines 18
2.4.3 Evaluation Under a Cost Policy 18
2.5 Deep Learning for Tabular and Sequential Transactions 19
2.5.1 Tabular Deep Nets Versus Boosted Trees 20
2.5.2 Sequence Modeling 20
2.5.3 Representation Learning with Autoencoders 21
2.6 Synthetic Data for Imbalance: SMOTE Family and CTGAN 21
2.6.1 Why Tabular GANs 23
2.6.2 Guardrails for Synthetic Generation 23
2.6.3 How CTGAN Interacts with Cost 24
2.6.4 Federated and LLM-enhanced Synthesis 25
2.7 Hybrid and Stacked Designs 26
2.7.1 Generative - Sequential Hybrids 26
2.7.2 Mixture-of-Experts and Graph Fusion 27
2.8 Explainability for Regulated Use 27
2.8.1 Local and Global Explanations 28
2.8.2 Synthetic-First Explainability 29
2.8.3 Stability and Governance 29
2.9 Privacy-Preserving Collaboration 30
2.9.1 Practical Considerations 30
2.9.2 Governance in a Federated Setting 31
2.10 Evaluation Protocols, Leakage, and Calibration 31
2.10.1 Confidence Intervals and Significance 32
2.11 Cost-Aware Thresholding and Policy Curves 33
2.11.1 Operating the Policy 33
2.11.2 Ethical and Customer Experience Considerations 34
2.12 CTGAN Audits and Validity Checks 35
2.13 Role of Representation Learning and Embeddings 35
2.14 Governance, Fairness, and Audit Trails 36

2.14.1 Fairness Checks and Redress 36
2.14.2 Adversarial robustness 37
2.15 Summary and Gap Statement 38
2.16 Concept Drift and Adaptive Learning 39
2.16.1 Scheduled Retraining and Sliding Windows 39
2.16.2 Online Updates and Threshold Adaptation 39
2.16.3 Reinforcement Learning for Policy Control 40
2.17 Diffusion Models and Generative Extensions 40
2.17.1 Potential advantages for fraud 41
2.17.2 Evaluation discipline still applies 42
2.18 Causal Explainability and Counterfactual Analysis 42
2.18.1 Methods and Safeguards 43
2.18.2 Linking to Fairness 43
2.19 Fairness, Bias Mitigation, and Ethical AI 43
2.19.1 Measuring Fairness 44
2.19.2 Mitigation Strategies 44
2.20 Cross-Family Synthesis and Theoretical Implications 45
2.21 Emerging Research Directions and Future Trends 46
2.22 Summary 48
Chapter 3: Methodology 51
3.1 Purpose and Problem–Solution Framing 51
3.2 Data and splits 52
3.2.1 Exploratory Data Analysis 52
3.2.2 Sources and Data Characteristics 56
3.2.3 Partitioning and Leakage Control 57
3.2.4 Rationale and Link to Research Questions 58
3.3 Preprocessing and Feature Handling 58
3.3.1 Numeric Scaling and Outlier Stability 58
3.3.2 Categorical Encoding and Optional Embeddings 59

3.3.3 Temporal Features and Velocity 59
3.3.4 Data Conditioning and Rare Value Treatment 60
3.3.5 Preprocessing Alignment to RQ1- RQ3 60
3.4 CTGAN Synthesis and Audits 61
3.4.1 Generator Objectives and Configuration 61
3.4.2 Three-Part Audit to Prevent Drift and Memorization 62
3.4.3 Privacy Posture 63
3.4.4 Expected Outcome and Link to Hypotheses 63
3.5 Supervised Learners and Calibration 63
3.5.1 Baseline Learners 63
3.5.2 Class Weighting and Imbalance Parameters 64
3.5.3 Probability Calibration 64
3.5.4 Hyperparameters and Compact Search 65
3.6 Unsupervised Anomaly Signals 65
3.6.1 Isolation Forest and Autoencoder Training 65
3.6.2 Score Fusion with Supervised Learners 66
3.6.3 Expected Outcome and Link to Hypotheses 66
3.7 Cost-Aware Decision Policy 66
3.7.1 Loss Function and Threshold Selection 66
3.7.2 Primary and Secondary Metrics at the Policy Point 67
3.7.3 Sensitivity to Policy Coefficients 67
3.8 Explainability and Governance 68
3.8.1 SHAP Workflow on Real and Synthetic Validation 68
3.8.2 Governance Artifacts and Drift Hooks 68
3.9 Training, Tuning, and Evaluation Loop 69
3.9.1 End-to-end Loop 69
3.9.2 Implementation Details 70
3.10 Exact Ablation Plan 70
3.11 Threats to Validity and Limitations 72

3.12 Ethical Considerations and Fairness 73
3.13 Reproducibility and Experiment Tracking 73
3.14 Computational Resources and Efficiency 74
3.15 Software Stack and Environment Controls 74
3.16 Equations and Definitions 74
3.17 Robustness, Drift Monitoring, and Recalibration 75
3.18 Security and Adversarial Resilience 76
3.19 Data Management and Privacy Posture 76
3.20 Supplementary Analysis 76
3.21 Alignment to Research Questions and Hypotheses 77
3.22 Implementation Blueprint 78
3.23 Statistical Analysis and Uncertainty Quantification 79
3.24 Documentation for Audit and Review 80
3.25 Deployment Notes and Operations 80
3.26 Summary 81
Chapter 4 - Results 82
4.1 Introduction 82
4.2 Experimental Context 83
4.2.1 Cost-Sensitive Threshold Selection 84
4.2.2 Data Quality and Leakage Checks 88
4.2.3 Probability Calibration and Score Reliability 88
4.3 Baseline Ensemble Performance 89
4.4 CTGAN Augmentation for Minority Coverage 90
4.4.1 Distributional Sanity Checks for CTGAN 94
4.5 Anomaly Feature Fusion 96
4.5.1 Error Analysis and Slice Performance 100
4.6 Explainability and Synthetic Validation Parity 101
4.7 Comparative Evaluation of Cost Terms 102
4.7.1 Sensitivity to Cost Weights and Alert Budget 104

4.8 Confidence Intervals and Robustness 104
4.8.1 Stability Across Random Seeds 106
4.9 Operational Summary 106
4.9.1 Practical Deployment Notes 107
4.10 Discussion of Findings 108
4.10.1 Why the Winner Works 109
4.11 Strengths and Limitations 110
4.11.1 Reproducibility Checklist 110
4.12 Summary and Integration of Findings 111
Chapter 5: Discussion and Future Work 112
5.1 Overview and Purpose of Discussion 112
5.2 Interpretation of Results 113
5.3 Integration and Contribution 115
5.4 Theoretical and Practical Implications 117
5.5 Ethical, Regulatory, and Interpretability Considerations 117
5.6 Limitations and Scope Boundaries 118
5.7 Future Work and Federated Learning Outlook 119
5.8 Conclusion 120
References 122

List of Figures

Figure 1. Class Distribution of the creditcard.csv Dataset.....53

Figure 2. KDE Plot of Transaction Amount by Fraud Status 54

Figure 3. Correlation Matrix for PCA Components.....55

Figure 4. Number of Fraudulent Transactions Over Time.....56

Figure 5. XGBoost Precision–Recall Curve (test set) 85

Figure 6. Cost versus Threshold for XGBoost with validation selected $t*$ 86

Figure 7. Confusion Matrix for XGBoost at t* on test 87

Figure 8. XGBoost Precision–Recall Curve with CTGAN (train only) 91

Figure 9. Cost versus Threshold for XGBoost with CTGAN (train only).....92

Figure 10. Confusion Matrix for XGBoost with CTGAN at $t*$on test.....93

Figure 11. Real vs CTGAN Distribution Overlays for Top PCA Components 95

Figure 12. XGBoost Precision–Recall Curve with anomaly feature fusion 97

Figure 13. Cost versus Threshold for XGBoost with anomaly feature fusion.....98

Figure 14. Confusion Matrix for XGBoost with anomaly feature fusion at $t*$on test.....99

Figure 15. SHAP parity overlap at top ten between real and synthetic validation 102

Figure 16. Models cost leaderboard.....103

Figure 17. Cost at $t*$ with ninety five percent confidence interval.....105

List of Tables

Table 1. Baseline results at t* 89

Table 2. CTGAN augmented results at t* 93

Table 3. Anomaly feature fusion results at t* 99

Table 4. SHAP parity summary 102

Table 5. Consolidated results at t* 103

Table 6. Bootstrap confidence intervals 105

List of Symbols

α	Cost weight assigned to false positives
β	Cost weight assigned to false negatives
t	Decision threshold
t^*	Optimal threshold minimizing total expected cost

List of Acronyms

AI	Artificial Intelligence
AUC	Area Under the Curve
CTGAN	Conditional Tabular Generative Adversarial Network
FL	Federated Learning
FP	False Positive
FN	False Negative
TP	True positives
TN	True negatives
ML	Machine Learning
SHAP	Shapley Additive Explanations
SMOTE	Synthetic Minority Over-sampling Technique
XAI	Explainable Artificial Intelligence

Chapter 1: Introduction

1.1 Background and Context

It is no doubt that the scale, complexity, sophistication, as well as economic effects of financial fraud, continue to develop at a rapid pace. As stated by the U.S. Federal Trade Commission (2024), financial fraud losses in 2023 reached a record-breaking total of over $10 billion, thanks in a large part to online payments, synthetic identities, as well as AI-driven fraudulent activity (Lin, 2025). As such, those detection systems that apply rule-based filters as well as human-based review processes no longer possess the capabilities to effectively deal with today's complicated fraud activity in so far as they lack effectiveness in contemporary payment systems (Odufisan et al., 2025).

Rule-based systems, however, also exist in the realm of compliance, but they fall short in identifying fraudulent techniques hidden in honest transaction activity. However, with the proliferation of digital networks, financial institutions are increasingly considering AI & ML solutions that help in detecting multidimensional deviations that represent fraud (Faisal et al., 2024; Hashemi et al., 2022).

However, the problem is larger than predictive accuracy. Fraudulent activities make up less than 0.2% of transactions, which is a highly imbalanced dataset. Models optimized for accuracy or AUC might seem effective but fail to catch most fraud cases or

trigger too many false positives. Recent studies identify this imbalance as one of the main challenges in real-world implementation (Alothman et al., 2022; Kennedy et al., 2023; Tayebi & Kafhali, 2025). On the contrary, finding the optimal balance between recall and precision while quantifying the financial impact of errors becomes much more important than chasing the maximum general metrics. Cost-sensitive models that minimize loss functions such as $\alpha \cdot FP + \beta \cdot FN$ better fit with the business goals in question (Cheng et al., 2024).

1.2 Problem Statement

Despite progress in both ensemble and deep learning, effective balancing between cost and interpretability in fraud detection under class imbalance remains a great challenge. Techniques such as SMOTE and ADASYN improve recall at the price of distorting feature distributions, leading to overfitting and generalization reductions commonly observed by Ahammad et al. (2020).

Recent approaches to deep learning-based advances include GAN-based models such as CTGAN and transformer architectures such as Finsformer, equipped with the ability to learn complex temporal and tabular relationships (Kong et al., 2024; Mienye & Swart, 2024; An et al., 2024). These types of approaches are often based on cross-validation in the same dataset but are not intended to simulate real-world deployment, where the model would be tested on fixed, unseen test partitions.

Of these, privacy and transparency have a significant place. Financial institutions have strict regulations on the sharing of data. FL allows decentralized model training without sharing raw data. Shakeri et al. (2024) proposed a privacy-preserving, auditable

fraud detection technique using federated learning combined with explainable AI tools such as SHAP and LIME. However, very few integrated studies consider synthetic data generation, cost-aware learning, and explainability under one reproducible framework. This paper aims to bridge that gap.

1.3 Purpose of Study

This paper presents a clear, auditable, and reliable AI/ML solution to the fraud detection that tackles difficulties in four methodological areas: cost-sensitive decision optimization, data augmentation using synthetic data, fusion of anomaly features, as well as privacy-preserving explainability.

Firstly, cost-sensitive modeling incorporates parameters α and β in order to estimate the effects of FP & TN in a way that allows the minimization of the expected total loss in accordance with business risk policies (Lin, 2025; Cheng et al., 2024).

Secondly, data augmentation via the Conditional Tabular Generative Adversarial Networks (CTGAN) is employed in this process to address the issue of minority class data limitation. distributions. This assists in well-leaned data under high imbalances (Awosika et al.,2024). Thirdly, the fusion strategy of the anomaly feature with both Isolation Forest, as well as Autoencoder model techniques, combined supervised learning strategies in order to reinforce cost efficiency with little compromise on the recall of results in representing the small fraud marks that conventional variables overlook (Aljunaid et al., 2025).

Lastly, privacy safe explainability uses the concept of SHAP-based interpretability techniques not only on the validation data set but also on synthetic data,

offering a regulator-compliant level of transparency with no concern to customer privacy (Lee et al., 2024).

These four factors complement each other in enhancing recall, accuracy, and cost efficiency, creating a firm foundation of operational trust in high-stakes financial settings. This framework is also flexible in terms of implementing federated learning in the future in maintaining its capabilities of having cost efficiency and protecting user privacy.

1.4 Research Questions

RQ1: Can a cost-aware method be developed to improve fraud detection on imbalanced financial data while reducing operational loss through an $\alpha \cdot FP + \beta \cdot FN$–based threshold?

RQ2: Can a controlled CTGAN-based augmentation method be applied to improve minority-class representation without degrading operational cost at t^*?

RQ3: Can a feature-fusion method incorporate unsupervised anomaly signals (Isolation Forest and Autoencoder scores) into supervised fraud models to lower Cost at t^* without sacrificing recall?

RQ4: Can a synthetic-validation explainability method be implemented to provide regulator-ready insights while preserving privacy?

1.5 Research Hypotheses

H1: Selecting the operating point by minimizing $\alpha \cdot FP + \beta \cdot FN$ on validation could yield a lower Cost at t^* than common metric-based or default thresholds.

H2: Applying CTGAN-based augmentation could result in improved recall at comparable precision, thereby reducing Cost at t^*.

H3: Integrating anomaly scores into supervised models could result in reduced Cost at t^* at comparable or higher recall relative to the same model without fusion.

H4: Using a synthetic validation twin could result in SHAP feature-importance rankings that closely mirror those from real validation (high overlap and rank correlation)

1.6 Significance of the Study

This research brings together modern AI tools and the financial industry's growing demand for clear, cost-effective, and regulation-compliant systems. By directly including cost as part of the model's design, it creates a strong link between model performance and real business value (Devi et al., 2024).

The approach also supports better governance by offering features like explainability using synthetic data and readiness for federated learning. These tools help organizations share insights safely while staying in line with changing privacy laws.

Overall, the framework promotes the development of AI systems in finance that are not only effective but also ethical and trustworthy. In real-world settings, it can help guide the design of fraud detection tools that are clear, financially justified, and reliable in today's fast-changing digital world.

1.7 Scope and Delimitations

This study is centered on supervised binary classification, using the public creditcard.csv dataset. It looks at ensemble and hybrid models, including Logistic

Regression and a combined XGBoost-AdaBoost model, tested under both natural class distribution and one improved using CTGAN-generated synthetic data.

To avoid data leakage, the dataset is split into 60% for training, 20% for validation, and 20% for testing. Only the training data is augmented, which ensures that validation and testing reflect true performance and generalization.

The study limits its scope to structured tabular data and cost-sensitive supervised learning. Although topics like multimodal learning and federated learning are discussed briefly, they are not applied in this project. This narrow focus allows for a deep, detailed analysis of the selected methods, all within a framework that is designed to be transparent and reproducible, as supported by the work of Leevy et al. (2023) and Hayat and Magnier (2025).

1.8 Expected Outcomes

This research introduces a full fraud detection framework that combines synthetic data creation, cost-aware evaluation, and explainable validation. It is designed to be ready for real-world operations and compliance with regulations.

The first part of the system uses CTGAN to generate synthetic tabular data. This is expected to improve recall and make the model more robust by better representing rare fraud patterns without disrupting the overall data structure. The goal is to improve the model's ability to detect fraud while still being accurate with normal transactions (Lin, 2025; Tayebi & Kafhali, 2025).

The second part involves a cost-aware thresholding method, which uses the $\alpha \cdot FP + \beta \cdot FN$ formula to reduce financial losses. Instead of relying only on statistical scores, this approach weighs the real-world costs of false positives and false negatives to help set better decision points (Cheng et al., 2024; Aljunaid et al., 2025).

The third component focuses on explainability. Using SHAP values, it checks whether explanations made with synthetic data match those from real data. If successful, this offers a way to explain model decisions clearly without revealing sensitive financial or personal details (Awosika et al., 2024; Shakeri et al., 2024).

Together, these elements offer a practical and well-rounded method for building fraud detection systems that are accurate, cost-effective, transparent, and privacy-conscious.

Recent industrial research by Khosravi et al. (2024) showed that CTGAN-based data augmentation can significantly improve recall in highly imbalanced manufacturing datasets. Their study achieved recall improvements between 30% and 90% across models like Decision Trees, Random Forests, and Logistic Regression, without lowering precision. This supports the idea of testing CTGAN's impact in the context of financial fraud detection (Khosravi et al., 2024).

1.9 Organization of Praxis

Chapter 1: Introduction - This covers motivation, scope, purpose, research questions, hypotheses, and significance.

Chapter 2: Literature Review - Reviews methods for fraud detection, class imbalance, GANs, cost-aware modeling, explainability, and privacy.

Chapter 3: Methodology - Describes the division of data, cost function, the setting up of CTGAN, SHAP analysis, and plans for ablation studies.

Chapter 4: Results - Comprises of results on baseline models, impact of CTGAN, anomaly fusion, and explainability parity.

Chapter 5: Discussion - The implications, limitations, and future work, including FL-based expansion, are discussed.

Chapter 2: Literature Review

2.1 Purpose and Roadmap

This chapter reviews current research on applying artificial intelligence and machine learning to financial fraud detection. The discussion follows the four challenges outlined in Chapter 1. Prior studies highlight that fraud detection involves rare events, unstable behavior, and constraints imposed by cost, transparency, and privacy (Nicholls et al., 2021; Psychoula et al., 2021; Zhang et al., 2023). However, many of those studies evaluate algorithms in isolation rather than showing how these challenges interact. This chapter takes a different direction by surveying the literature while explicitly linking each theme to the design choices made in this praxis.

Works such as Hashemi et al. (2022) and Pundkar and Zubei (2023) summarize the imbalance challenge but often stop at descriptive statements about rarity. In contrast, this study treats imbalance as the foundation for cost-aware thresholding and for deciding when augmentation or anomaly signals are worth adding. Likewise, while many papers report AUC or F1 as primary indicators, researchers such as Nicholls et al. (2021) and Psychoula et al. (2021) observe that these metrics do not always reflect real-world financial impact. This praxis follows that critique directly by embedding α and β at the threshold-selection stage instead of relying on metric-based operating points.

Transparency and privacy also appear in much of the literature, especially in governance-oriented work by Nicholls et al. (2021) and Psychoula et al. (2021). Yet, most prior studies either discuss these issues conceptually or apply explainability tools without

integrating them into a governance framework. By contrast, the design here draws from the privacy-first orientation in Zhang et al. (2023) and Awosika et al. (2024) and implements a practical mechanism, namely synthetic validation for SHAP, to produce explanations that can be shared without using sensitive real transactions.

2.2 Fraud Detection as a Rare-Event and Cost Problem

Fraud detection is fundamentally a rare-event classification problem, with fraudulent transactions typically representing only a fraction of a percent of the total. Rao et al. (2025) and Mehta (2024) both note that despite this reality, many evaluations still rely on accuracy or AUC alone, which can mask poor performance on the minority class. A model that labels all transactions as legitimate can achieve high accuracy but fail operationally. This gap is acknowledged in recent surveys but is not consistently addressed in experimental practice. The present study differs from these metric-centered approaches by treating rarity not just as a descriptive property but as a constraint that shapes the entire evaluation pipeline. Thresholds, comparisons, cost measurement, and the use of augmentation are all grounded in the rare-event nature of the task. Instead of reporting a mix of thresholds, this work fixes a single cost-aware operating point across baselines, CTGAN-augmented models, and anomaly-fusion models and ensures that comparisons reflect true operational trade-offs.

2.2.1 Moving from Metric Scores to Operating Decisions

Recent papers increasingly argue that the most important fraud-detection decision is not which model scores highest on AUC, but which threshold minimizes expected loss under defined cost weights (Cheng et al., 2024; Psychoula et al., 2021). While many

authors describe this idea, fewer implement it rigorously. Several studies still choose thresholds informally or depend on default cutoffs, leaving the connection between model scores and business impact unclear. This praxis adopts the cost-policy approach directly. The $\alpha \cdot FP + \beta \cdot FN$ formulation used in this work mirrors the decision-centered framing in Cheng et al. (2024) and is applied strictly on the validation set with a recall floor, producing a threshold t* that is then transferred without change to the test set. Compared to prior work that treats cost discussion as a conceptual appendix, this study incorporates cost into the core evaluation loop and uses it to define RQ1 and H1. This distinction shifts the focus from improving leaderboard metrics to improving real-world decision outcomes.

2.2.2 Practical Ingredients of Cost-Aware Policies

The practical ingredients of cost-aware decision-making include loss estimates, investigation capacity, recall targets, and threshold selection. These components are defined explicitly in governance-oriented work such as Nicholls et al. (2021) and Cheng et al. (2024), who describe how institutions should translate predictive outputs into operational policies.

Rao et al. (2025) acknowledge financial cost considerations but still determine operating points using metric-driven criteria such as maximizing F1. Mehta (2024) reports AUC and F1 as the primary evaluation metrics without demonstrating how the selected thresholds satisfy loss constraints or investigation capacity. These specific examples show how α and β are mentioned conceptually but are not incorporated into the threshold-selection process.

This study follows the full procedural guidance in Nicholls et al. (2021) and Cheng et al. (2024). It sets α and β explicitly, imposes a minimum recall requirement, and selects operating points through cost sweeps on the validation split before applying the chosen threshold once to the locked test set. This directly supports RQ1 and motivates H1. The distinction is clear: Rao et al. (2025) and Mehta (2024) reference cost principles without operationalizing them, whereas this praxis implements a repeatable, auditable cost-aware protocol applied consistently across all models and experiments.

2.2.3 Implications for Model Comparison

A recurring theme in the literature is that comparisons are only meaningful when models operate under the same decision policy (Pundkar & Zubei, 2023; SP et al., 2024). Studies that switch thresholds between models or rely on different operating criteria risk drawing misleading conclusions. For example, SP et al. (2024) show that differences in calibration can shift thresholds enough to change relative performance even when raw AUC differs only slightly. This study follows that guidance closely. All models, Logistic Regression, tree ensembles, CTGAN-augmented variants, and anomaly-fusion hybrids are evaluated under the same cost function and recall floor. The comparison is therefore "controlled" in the sense advocated by Pundkar & Zubei (2023). The contrast with other studies lies in what is held constant: many works vary both model and operating criteria; this work varies only the model, allowing observed differences in cost to be attributed to the models themselves rather than to differences in evaluation practices.

2.3 Supervised Baselines and Ensemble Learners

Ensemble learners provide dependable performance on tabular fraud data. Random Forest reduces variance by averaging many decorrelated trees. Gradient boosting techniques such as XGBoost and LightGBM fit shallow trees to residuals and capture non-linear interactions with modest feature engineering. Across recent fraud studies, these families usually deliver strong precision–recall performance and can be calibrated effectively (Rocha et al., 2025; Alothman et al., 2022). Rocha et al. (2025) and Alothman et al. (2022) largely treat these ensembles as the main end point and compare them mainly on AUC or F1. This praxis uses the same model families but assigns them a different role. Here, ensembles serve as the supervised backbone that supports cost-aware thresholding, CTGAN augmentation, and anomaly-feature fusion. The comparison with prior work is therefore not in the choice of models, which remains conventional, but in how they are evaluated and combined. Instead of asking which ensemble wins a metric benchmark, this study asks how these standard learners behave once they are placed inside a strict cost policy and audit framework.

2.3.1 Logistic Regression as a Reference Point

Logistic Regression remains a fair and interpretable baseline against which more complex models can be compared. It is easy to regularize, fast to train, and straightforward to explain to stakeholders. The model often underperforms when non-linear relationships or strong interactions are present, yet it continues to appear in fraud studies because of its simplicity and governance value. Zhang et al. (2021) and Patel et al. (2024), for example, rely on logistic coefficients as an initial view of feature

directionality and as a support for early feature engineering. In many institutions, these coefficients act as a “sanity check” for more complex learners. This study follows that pattern closely. Logistic Regression is used as an anchor model and as a reference point for governance. The contribution is not a new use of Logistic Regression, but a different way of reading its performance. The model is evaluated under the same cost-aware threshold policy as XGBoost and other ensembles. When the ensemble improves cost relative to Logistic Regression at t*, the result parallels the progression reported in Zhang et al. (2021) and Patel et al. (2024) but is interpreted in terms of operational loss rather than generic accuracy. This ties the familiar baseline directly to RQ1 by framing it as the starting policy rather than as a purely statistical benchmark.

2.3.2 Addressing Imbalance During Training

Imbalance-sensitive training improves ensemble performance. Ahammad et al. (2020) and Patel et al. (2024) describe three main strategies: adjusting class weights or scale_pos_weight to discourage majority-class dominance, forming more balanced mini-batches, and applying oversampling such as SMOTE inside each training fold. All three approaches share the principle that train, validation, and test data must remain strictly separated so that reported performance is honest.

The present work adopts this separation principle but makes a different choice about how imbalance is handled. Instead of relying on SMOTE-family oversamplers as in Ahammad et al. (2020) and Patel et al. (2024), this study combines class-weighted ensembles with CTGAN-based synthesis applied only to the training split. This choice is consistent with authors who caution that interpolation oversamplers can blur decision

boundaries, but it still aims to expand minority coverage. Chapter 4 therefore compares a class-weighted, cost-aware baseline directly against a CTGAN-augmented version under identical splits and policies. In that way, the study uses the same concern about imbalance as prior work but tests a different remedy and evaluates it explicitly in terms of Cost at t*.

2.3.3 Calibration and Threshold Transfer

Tree ensembles can be overconfident, which complicates threshold transfer from validation to production. SP et al. (2024) and Rocha et al. (2025) show that validation-time calibration, through Platt scaling or isotonic regression, improves probability quality and stabilizes downstream decisions. They also note that calibrated scores make it easier to create segment-specific thresholds when loss differs across channels. The praxis adopts this calibration advice in a direct way. Rather than reporting raw ensemble scores, as some earlier studies do, this work calibrates probabilities before any cost sweeps are run. The chosen t* is then moved from validation to test without adjustment. This design choice narrows the gap between experimental practice and operational guidance in SP et al. (2024) and Rocha et al. (2025). The difference is that calibration is treated as a required step in the evaluation pipeline rather than as an optional refinement, because reliable threshold transfer is central to RQ1 and to the comparison between baseline and augmented models.

2.3.4 Stacking and Model Fusion

Stacked generalization blends complementary learners. Abdelghafour et al. (2024) and Almalki and Masud (2025) describe stacks that combine Logistic Regression, Random Forest, XGBoost, and LightGBM with a meta-learner that consumes out-of-fold

predictions. When leakage controls and calibration are in place, these stacks can lift recall at roughly constant precision, because each base learner focuses on different structure. This study takes a more selective approach. Ensembles are still used as anchors, which is aligned with Rocha et al. (2025), but stacking is only considered when it offers measurable gains under the cost policy defined in Chapter 3. Institutions described by Abdelghafour et al. (2024) often follow a similar practice, starting with a tuned XGBoost or LightGBM model and adding a stacker only if recall improves without sacrificing calibration or explainability. The anomaly-feature fusion used here is intentionally closer to a modest extension of a strong baseline than to a deep hierarchical stack. The goal is to keep the governance burden lower while still testing whether additional signals can reduce Cost at t*. This positions the praxis between stacking-heavy research that pursues maximum accuracy and governance-sensitive work that favors simpler, auditable combinations.

2.4 Unsupervised Anomaly Detection

Unsupervised methods search for unusual behavior without labels. Isolation Forest isolates points along random splits until they are separated, with shorter paths indicating outliers. Local Outlier Factor compares local density to neighbors and flags points that are much sparser than their surroundings. One-class SVM estimates a boundary around normal points. Autoencoders learn to reconstruct input data and flag cases with high reconstruction error (Psychoula et al., 2021; Oreški, 2023). These tools are attractive in fraud detection because they can surface novel patterns that supervised models have not yet seen. Psychoula et al. (2021) and Oreški (2023) also point out an important limitation. When prevalence is extremely low, unsupervised detectors used as

stand-alone decision makers tend to generate many false positives, which can quickly overwhelm operations. Several anomaly-detection studies still evaluate models primarily with unsupervised scores and do not show how those scores would fit within a realistic alert budget. This study aligns with the more cautious interpretation in Psychoula et al. (2021) by not deploying unsupervised models as final decision rules. Instead, it uses their anomaly scores as additional supervised features and evaluates them under the same cost-aware policy as the baseline.

2.4.1 Strengths and Limits

The main strength of unsupervised techniques is their sensitivity to novelty. Olowu et al. (2024) and Psychoula et al. (2021) show that such methods can highlight emerging fraud patterns when labels are sparse or delayed. At the same time, they document the drawback that these detectors often flag a large number of benign transactions in highly imbalanced domains. Even modest false-positive rates can translate into untenable alert volumes. This praxis adopts the strengths identified in those studies while sidestepping the documented weaknesses. Instead of treating anomaly scores as direct triggers for investigations, they are fed into a supervised cost-aware learner that can learn when a high anomaly score is genuinely informative and when it is noise. The design therefore respects the caution expressed in Olowu et al. (2024) and positions anomaly detection as a supporting signal rather than as an independent detector. This directly informs RQ3, which asks whether such fusion can lower cost at t* without sacrificing recall.

2.4.2 Hybrid Pipelines

Hybrid pipelines that combine anomaly detection and supervised learning are increasingly common. Psychoula et al. (2021) and Kamuangu (2024) describe architectures where one or two anomaly scores are computed on training data and then appended as features for a downstream classifier. Some variants introduce a second stage to classify the type of anomaly or use variational autoencoders to provide both reconstruction error and latent representations. These designs often show improved recall but sometimes lack a clear link to operational cost. The anomaly-feature fusion in this study follows the same general idea of combining anomaly scores with a supervised learner, but with two key differences. First, the backbone is fixed to a single XGBoost model, which reduces architectural complexity compared to multi-stage cascades described by Kamuangu (2024). Second, the value of the anomaly features is judged entirely by their impact on Cost at t* under the fixed policy introduced earlier. The comparison with prior hybrids is therefore not mainly architectural. It is about how the additional signals are evaluated. The question is whether the extra complexity is justified by a measurable reduction in cost on a locked test set.

2.4.3 Evaluation Under a Cost Policy

Rocha et al. (2025) argue that hybrid systems should be evaluated within the same leakage-controlled and cost-aware framework as simpler baselines. Olowu et al. (2024) echo this by noting that it is hard to know whether anomaly-based features really help if they are not compared at the same operating point. Many anomaly-detection papers still report qualitative improvements or use thresholds that differ from those of the supervised

baselines. This praxis follows the more disciplined line of work. Anomaly features are computed on training data only, the supervised learner is calibrated on validation, and the operating threshold is chosen to minimize expected loss subject to a recall floor. The same locked holdout is then used for all model families. Any cost reduction can therefore be attributed to the anomaly features themselves rather than to relaxed thresholds or different evaluation regimes. In this way, the study uses the recommendations in Rocha et al. (2025) and Olowu et al. (2024) as explicit design rules and ties the hybrid evaluation directly to RQ3 and H3.

2.5 Deep Learning for Tabular and Sequential Transactions

Deep learning supports fraud detection when non-linear patterns or temporal structure play a central role. Multilayer perceptrons (MLPs) uncover complex feature interactions in tabular settings, while recurrent networks and transformers capture behavior across time, such as spending sequences or device-session trails. Studies that use dense transaction histories report strong results with sequence-aware architectures because recurrent layers learn regularity and deviation more effectively than shallow models (An et al., 2024; Odufisan et al., 2025). This thesis differs from sequence-heavy research because the public creditcard.csv dataset contains only isolated transactions, not full histories. As a result, the study aligns more closely with work on tabular fraud detection and focuses on ensembles, but still includes a simple MLP to maintain comparability with prior tabular deep learning studies. This creates a direct link back to the literature while keeping the evaluation grounded in the cost policy defined in Chapter 3.

2.5.1 Tabular Deep Nets Versus Boosted Trees

On tabular data, boosted trees often match or surpass MLPs, especially when engineered features already capture meaningful structure. Rocha et al. (2025) and Alothman et al. (2022) note that gradient boosting tends to be more data-efficient, particularly when labels are scarce or feature interactions are strong. Many deep-learning-centered fraud studies emphasize architectural novelty, but they sometimes evaluate models mainly with conventional metrics such as AUC or F1. The design in this study follows the guidance noted by Rocha et al. (2025): include an MLP for parity with prior research, but rely on XGBoost as the main supervised backbone because of its stability on tabular fraud data. This makes the contribution different from deep-learning-first papers. Here, the deep net is not a proposed improvement; instead, it is one of several baselines tested under the same cost-aware thresholding routine. The comparison therefore asks whether an MLP provides any cost benefit at t*, not whether it maximizes accuracy.

2.5.2 Sequence Modeling

Sequence models—GRU, LSTM, and transformers—perform well when full customer histories, device sequences, or session data are available. An et al. (2024) show that such architectures detect behavioral drifts, changes in spending rhythm, and irregular transition patterns across events. These models often outperform tabular methods when sequences are complete and well structured. In contrast, this study does not use sequence modeling because the public dataset contains time as an index, not an ordered series tied to customer sessions. This limitation distinguishes the present work from sequence-

oriented studies. Instead of adapting sequence models to incomplete histories, the praxis focuses on what can be done with tabular methods, synthetic augmentation, anomaly features, and cost-based evaluation. This difference also clarifies why RQ1–RQ3 center on cost, thresholding, and augmentation—not on sequential architectures.

2.5.3 Representation Learning with Autoencoders

Autoencoders compress inputs into latent features that can serve as new representations. In fraud studies, denoising autoencoders often act as regularizers by filtering noise while retaining meaningful signal, and their latent space can be passed into downstream learners (Oreški, 2023). Some studies treat learned embeddings as the core contribution.

This thesis takes a different approach. Autoencoder-generated features are used only inside the anomaly-feature fusion pathway, not as a global replacement for tabular models. This is consistent with the goal of maintaining interpretability and cost clarity. The choice also responds to the findings of An et al. (2024), who show that representation learning can help but does not always justify added complexity when data lack rich structure. In the present work, the added anomaly features are evaluated strictly under the same cost-aware policy, which ties representation learning directly to RQ3 rather than presenting it as a stand-alone improvement.

2.6 Synthetic Data for Imbalance: SMOTE Family and CTGAN

Imbalance is a defining feature of fraud data. Classical oversamplers such as SMOTE, Borderline-SMOTE, and SMOTE-ENN create synthetic points in feature space

by interpolating between minority neighbors or by filtering noisy examples after oversampling. Patel et al. (2024) and Ahammad et al. (2020) report that these methods can improve minority recall, but they also show that interpolation sometimes blurs decision boundaries and can propagate noise in regions where the minority class is sparse. These limitations are important for this study because the credit card dataset provides few confirmed frauds, and methods that interpolate too aggressively may misrepresent rare combinations rather than strengthen them.

SMOTE, ADASYN, and SMOTE-ENN are frequently used to mitigate class imbalance in fraud datasets. However, this thesis does not employ these oversampling methods. Synthetic augmentation is limited to CTGAN and is applied only to the training split to preserve strict leakage control. This choice follows the concerns noted in Patel et al. (2024) and Ahammad et al. (2020) about distortion, and aligns more closely with recent CTGAN-focused work. For example, Zhu et al. (2024) and Tayebi and Kafhali (2025) show that CTGAN can generate minority samples that capture more complex structure than interpolation methods. Their studies, however, evaluate CTGAN largely through changes in recall or AUC without tying improvements to explicit cost thresholds.

The current study addresses that gap by placing CTGAN inside a cost-aware evaluation. CTGAN outputs are judged by whether they reduce expected loss at the chosen threshold rather than by recall alone. This approach differs from work such as Lee (2024), who demonstrates improved minority recall but does not incorporate cost-aware operating points. It also contrasts with studies such as Mienye and Swart (2024), who apply CTGAN without distribution audits. In this thesis, CTGAN outputs are audited with distributional checks, KS overlays, and SHAP-parity validation before any

comparison is made. By applying these stricter evaluation steps, the study tests CTGAN under conditions where synthetic augmentation must justify itself in terms of cost reduction, not only metric lift.

2.6.1 Why Tabular GANs

Tabular GANs learn the joint distribution of the minority class and can generate samples that reflect complex, non-linear relationships. CTGAN is designed for mixed data types and heavy-tailed distributions. It conditions generation on selected features and uses a training objective that resists pitfalls documented in earlier GAN work, such as mode collapse. Studies by Tayebi and Kafhali (2025) and Zhu et al. (2024) show that CTGAN can capture minority structure that interpolation methods fail to reproduce. Their results, however, focus mainly on improvements in minority recall or AUC and do not evaluate CTGAN under an explicit cost-aware threshold. This praxis adopts CTGAN for the same structural reasons described in those studies but tests it under stricter evaluation. Rather than assuming that improved minority coverage is beneficial, the study evaluates CTGAN only through expected cost at the chosen threshold. This approach differs from Zhu et al. (2024), who do not incorporate cost-based operating points, and extends the findings by asking whether CTGAN's structural strengths translate into real cost benefit when compared to a class-weighted, calibrated baseline. In this way, CTGAN is assessed not only as a generator but as a component in a cost-aware fraud workflow.

2.6.2 Guardrails for Synthetic Generation

Responsible use of CTGAN follows three rules. The first is to generate data on the training split only. The second is to audit synthetic distributions against real marginals

and pairwise relationships. The third is to evaluate once on a fixed real holdout. These practices respond directly to concerns raised in Strelcenia and Prakoonwit (2022) and Oreški (2023), who show that synthetic data can leak information or hide overfitting when generation and evaluation are not separated. Strelcenia and Prakoonwit (2022) in particular recommend discriminator checks to identify drift or over-smoothing in synthetic samples. The methodology in this study follows those guardrails exactly. CTGAN is trained only on training data, its outputs are checked with distributional plots and KS overlays, and the final comparison is made on a locked real test set. This contrasts with some CTGAN studies such as Mienye and Swart (2024), who report gains without accompanying distribution audits. The present approach aligns more closely with the governance-focused recommendations in Oreški (2023), by requiring synthetic data to pass explicit quality checks before entering the modeling pipeline.

2.6.3 How CTGAN Interacts with Cost

CTGAN can increase recall by expanding minority coverage, but improvements in recall alone do not guarantee better operational outcomes. Studies such as Cheng et al. (2024) and Pundkar and Zubei (2023) emphasize that minority expansion can raise false positives, which increases investigation burden and cost. Gains in recall must therefore be weighed against changes in alert volume at the chosen threshold. This principle guides the generation budget: synthetic data should not dominate the training set. The experiments in Chapter 4 apply this cost-aware logic directly. Unlike Zhu et al. (2024), who evaluate CTGAN primarily on classification metrics, this study evaluates CTGAN only in terms of expected loss at the threshold selected by the cost policy. The negative result, where CTGAN does not reduce cost on this dataset, stands in contrast to some of

the optimistic findings in Tayebi and Kafhali (2025). That contrast reinforces the importance of evaluating synthetic generation under cost constraints rather than by metric lift alone.

2.6.4 Federated and LLM-enhanced Synthesis

Federated synthesis enables multiple institutions to train generators collaboratively without sharing raw data. This helps address scarcity of minority patterns that may appear in one institution but not in others. Recent work also explores using large language models to guide tabular synthesis, especially where categorical features carry narrative or procedural meaning. Shakeri et al. (2024) and Lin (2025) describe early promising results in producing more realistic combinations under privacy limits.

The literature suggests that these methods may extend CTGAN's benefits to distributed settings. Tayebi and Kafhali (2025) also highlight that CTGAN improves minority representation when trained responsibly on the local data distribution. The present study references these developments but confines experimentation to local CTGAN synthesis on tabular credit card transactions. This narrower scope reflects the dataset constraints and focuses the empirical work on how synthetic augmentation behaves under a fixed cost policy, which directly supports H2. The difference is that the thesis evaluates CTGAN through cost and governance rather than broader claims of generative improvement.

2.7 Hybrid and Stacked Designs

High-performing fraud detection pipelines often mix techniques. Common patterns include combining ensemble classifiers with anomaly detection scores or pairing learned representations with a boosted-tree decision layer. Graph-based components add a relational view that uncovers coordinated behavior across accounts (Motie and Raahemi, 2023; Alarfaj and Shahzadi, 2024). These hybrid configurations can show strong performance, but in many cases they become large multi-stage systems that are difficult for institutions to govern. This thesis takes a more focused approach. The hybrid architecture is intentionally lean. It uses a strong ensemble backbone, optional CTGAN-based augmentation on the training split, and a limited set of anomaly features. All components are evaluated under a single cost-aware policy. Compared with the multi-part hybrid systems described by Motie and Raahemi (2023), the present design trades architectural breadth for operational clarity and transparency.

2.7.1 Generative - Sequential Hybrids

Several studies combine CTGAN with sequence models such as GRU or LSTM. Mienye and Swart (2024) report that when CTGAN supplies additional minority sequences, the recurrent model learns rarer transitions more effectively, which raises recall at constant precision if the synthetic sequences are audited for quality. These results rely on access to full transactional histories with meaningful temporal structure. The current research does not use GRU or LSTM models because the public creditcard.csv dataset lacks sessionized sequences. Instead, it borrows the underlying principle from generative–sequential hybrids: synthetic data should be added only when it plausibly

enhances coverage and should be evaluated under cost-aware thresholds. The comparison is therefore conceptual rather than architectural. The lesson from Mienye and Swart (2024) is applied by emphasizing careful auditing and cost-based acceptance rather than attempting to duplicate a sequence-based setup.

2.7.2 Mixture-of-Experts and Graph Fusion

Mixture-of-experts systems route transactions to specialized submodels, such as a graph expert or a sequence expert, and then combine the outputs. Graph neural networks can expose fraud patterns that appear normal on a single record but suspicious when connected across entities. Studies such as Thilagavathi et al. (2024) and Motie and Raahemi (2023) illustrate how graph embeddings reveal fraud rings and mule networks that would otherwise remain hidden. The approach in this thesis remains simpler. Ensembles serve as the supervised anchor, CTGAN offers minority coverage, and anomaly scores are added cautiously. This narrower design supports RQ2 and RQ3 while keeping the final decision layer transparent and easy to calibrate. Compared with the complex routing and multi-expert systems in the literature, the present work reduces governance overhead while still testing whether modest fusion improves cost at the chosen operating point.

2.8 Explainability for Regulated Use

Financial institutions must justify automated decisions to analysts, auditors, and regulators. Prior work notes that SHAP and LIME can provide explanations, but many studies stop after showing one-off plots without embedding explanation tools into a full governance cycle (Nicholls et al., 2021). Psychoula et al. (2021) also highlight that

explanations must be consistent and traceable across time, not just informative at a single moment. The current study aligns with these concerns by pairing SHAP with synthetic validation twins and parity checks. This placement of explainability inside a governance framework goes beyond uses of SHAP that treat it only as a visualization tool and supports both privacy and stability by ensuring that explanations can be reviewed without exposing sensitive records.

2.8.1 Local and Global Explanations

SHAP provides local reasoning for individual predictions and global rankings across the dataset. LIME approximates local behavior with surrogate models. Both are used in research to identify influential features and diagnose problematic patterns (Psychoula et al., 2021). Psychoula et al. (2021) also note that these tools are often applied in isolation, without clear integration into evaluation or privacy workflows. This praxis uses SHAP in a more structured way. It produces global and local summaries on both synthetic and real validation sets as part of a parity test. That design responds to the governance concerns raised in Nicholls et al. (2021) and moves beyond ad hoc plots by anchoring explanations inside the evaluation routine. Studies such as Awosika et al. (2024) and Tayebi and Kafhali (2025) emphasize feature-importance quality but do not evaluate explanations under confidentiality constraints. In contrast, the present work evaluates explanations under requirements for privacy stability by pairing SHAP summaries with a synthetic-validation twin.

2.8.2 Synthetic-First Explainability

Awosika et al. (2024) describe how synthetic validation data can be used to generate explanations without exposing sensitive features of real customers. Their work suggests that synthetic twins can support disclosure and auditing if they match real validation behavior closely. This study applies that principle directly. SHAP is computed on both real and synthetic validation sets. The comparison of feature rankings becomes a parity check that confirms whether synthetic explanations preserve the structure needed for governance. Rather than treating synthetic data solely as training input, this work uses CTGAN outputs for privacy-aware explanation. The same generator is evaluated for both classification impact and explanatory reliability.

2.8.3 Stability and Governance

Explainability must remain consistent across slices and retraining cycles. Nicholls et al. (2021) and Psychoula et al. (2021) note that instabilities can signal drift or calibration errors. They propose governance measures such as tracking feature stability indices. The approach in this thesis reflects those recommendations. Instead of using SHAP only once, the model's explanations are evaluated with synthetic-real parity checks to ensure consistency. Explainability must remain consistent across slices and retraining cycles. Nicholls et al. (2021) and Psychoula et al. (2021) note that instabilities can signal drift or calibration errors. They propose governance measures such as tracking feature stability indices. The approach in this thesis reflects those recommendations. Instead of using SHAP only once, the model's explanations are evaluated with synthetic-real parity

checks to ensure consistency. This approach treats explanation generation as an ongoing governance task that supports RQ4 and H4, rather than as a one-time demonstration.

2.9 Privacy-Preserving Collaboration

Federated learning enables institutions to collaborate without sharing raw data. Zhang et al. (2023) and Shakeri et al. (2024) show how local training with secure aggregation can improve fraud detection across diverse portfolios while preserving privacy. These approaches are attractive in regulated industries but require coordination and infrastructure across multiple sites. The current work does not implement federated learning but adopts its guiding ideas. Data remain local to each split, and synthetic validation replaces sensitive records for SHAP-based explanations. This narrower setting mirrors some governance principles of federated research but applies them in a single-dataset environment where privacy concerns arise mainly at the explanation stage.

2.9.1 Practical Considerations

Federated systems face challenges such as communication overhead, model inversion risks, and heterogeneity across institutions (Aljunaid et al., 2025; Zhang et al., 2023). These issues motivate secure aggregation, differential privacy, and personalization layers that allow the global model to adapt to local conditions. This study operates without multi-party heterogeneity. Instead, it focuses on privacy within a single dataset by using synthetic validation for SHAP reporting. The contrast illustrates how federated studies aim to secure collaboration across institutions, while this praxis applies similar governance principles to protect validation explanations. The underlying idea from Zhang

et al. (2023), that privacy must be balanced with utility, is preserved but applied in a simpler context.

2.9.2 Governance in a Federated Setting

Federated learning requires local calibration, local threshold selection, and local governance packs with cost policies and explanation logs (Awosika et al., 2024). These artifacts ensure that each institution aligns the shared model with its own risk appetite. Although federated deployment is future work here, the governance tools built in Chapter 4, such as cost-aware thresholds and SHAP parity reports, can transfer naturally to a federated environment. The distinction from Shakeri et al. (2024) is that this research demonstrates the approach within a single institution first. This alignment supports H4 without expanding the empirical scope.

2.10 Evaluation Protocols, Leakage, and Calibration

Evaluation design can inflate performance if not controlled. The literature repeatedly stresses four pitfalls.

First, preprocessing leakage occurs when normalization or target encoding is fit on the full dataset instead of the training split. Patel et al. (2024) note that this produces overly optimistic scores because the model gains indirect access to the test distribution.

Second, temporal leakage is common in fraud datasets with timestamps. Pundkar and Zubei (2023) show that models trained without proper time splits tend to overfit future patterns.

Third, uncalibrated probabilities limit threshold transfer. SP et al. (2024) demonstrate that thresholds chosen on validation do not generalize when score calibration is absent.

Fourth, metric inflation occurs when AUC or F1 is used without considering precision at the operating recall. Studies such as Pundkar and Zubei (2023) explain that high AUC does not guarantee acceptable alert volumes.

This study follows the evaluation discipline described above. All preprocessing is fit on the training split only, CTGAN is trained only on the training minority rows, and the operating threshold is selected on validation under a fixed cost policy before being applied once to the locked test set. This protocol is stricter than setups that rely on cross-validation alone or permit preprocessing across splits, and it follows the recommendations in SP et al. (2024) and Pundkar and Zubei (2023). This consistency ensures that comparisons between baseline, CTGAN-augmented, and anomaly-fused models remain credible.

2.10.1 Confidence Intervals and Significance

Rare-event estimates can be noisy. SP et al. (2024) recommend stratified bootstrap intervals for metrics such as precision, recall, F1, AUC-PR, and expected cost. Rocha et al. (2025) describe paired bootstrap tests for cost differences between competing models to determine whether improvements are statistically meaningful. This study mirrors that advice. Bootstrap intervals are applied to the locked test predictions to quantify uncertainty for each model under the same cost policy. Psychoula et al. (2021) highlight the importance of reproducible evaluation pipelines, and Chapter 3 adopts this practice through fixed splits and notebooks that regenerate all figures. Rather than reporting point

estimates alone, this work adds statistical discipline by showing which cost reductions are likely to reflect real improvements and which fall within sampling noise.

2.11 Cost-Aware Thresholding and Policy Curves

Policy selection benefits from visual tools that ground technical choices in business terms. Cheng et al. (2024) use cost curves to visualize $\alpha \cdot FP + \beta \cdot FN$ across thresholds, while Pundkar and Zubei (2023) annotate precision–recall curves with operating points that meet recall requirements. These tools help risk teams understand alert volume, financial exposure, and staffing implications. Some studies present these curves mainly as illustrations. In contrast, this praxis adopts the operational style described in Cheng et al. (2024) and uses cost curves to select thresholds systematically. The chosen threshold becomes part of the documented evaluation protocol rather than a hidden parameter.

2.11.1 Operating the Policy

Cheng et al. (2024) describe a practical playbook for applying cost-aware policies: set α and β, generate cost curves, identify feasible thresholds, choose the one that minimizes expected loss, and document the decision. Rocha et al. (2025) add that thresholds can be updated as prevalence shifts, provided calibration holds. This study follows that operational sequence directly. Thresholds are selected on the validation split under fixed α and β, and the chosen point is transferred unchanged to the test set. Cheng et al. (2024) describe a practical playbook for applying cost-aware policies: set α and β, generate cost curves, identify feasible thresholds, choose the one that minimizes expected loss, and document the decision. Rocha et al. (2025) add that thresholds can be updated

as prevalence shifts, provided calibration holds. This study follows that operational sequence directly. Thresholds are selected on the validation split under fixed α and β, and the chosen point is transferred unchanged to the test set. This is more transparent than reporting thresholds without explaining how they were derived, and it treats the policy layer as an explicit part of the contribution.

2.11.2 Ethical and Customer Experience Considerations

Nicholls et al. (2021) emphasize that cost is not the only criterion. False positives create friction that affects customers and analysts, while segment-level disparities can raise fairness concerns. Pundkar and Zubei (2023) recommend coupling cost analysis with fairness checks that monitor subgroup precision and false-positive rates. This study adopts that ethical framing by selecting thresholds that meet recall requirements while avoiding excessive alert volume. Nicholls et al. (2021) emphasize that cost is not the only criterion. False positives create friction that affects customers and analysts, while segment-level disparities can raise fairness concerns. Pundkar and Zubei (2023) recommend coupling cost analysis with fairness checks that monitor subgroup precision and false-positive rates. This study adopts that ethical framing by selecting thresholds that meet recall requirements while avoiding excessive alert volume. The design supports RQ1 and H1 by treating the operating point itself as a managed policy choice. Instead of optimizing metrics alone, the work explicitly links thresholding to customer experience and fairness considerations.

2.12 CTGAN Audits and Validity Checks

Synthetic augmentation raises concerns about drift, memorization, and unrealistic combinations. Strelcenia and Prakoonwit (2022) recommend univariate tests such as Kolmogorov–Smirnov to check marginal alignment between real and synthetic data. Abdi (2023) and Chen, Sun, and Ge (2024) analyze pairwise dependence to ensure synthetic features preserve relationships present in the real data. Oreški (2023) suggests counting out-of-support values and using a discriminator to detect synthetic batches that deviate too strongly.

This study applies those same safeguards. CTGAN is trained on the training split, synthetic rows are compared with real marginals and correlations, and all models are evaluated on a locked real test set. Tayebi and Kafhali (2025) describe scenarios where CTGAN improves minority recall, but they do not use cost-aware thresholds or audit synthetic data before merging. In contrast, this praxis combines CTGAN with strict audit checks and tests whether gains survive under the same cost policy. Reporting the negative finding that CTGAN did not improve cost on this dataset addresses the gap noted in prior work, where synthetic augmentation is often assumed to help but rarely evaluated under policy constraints.

2.13 Role of Representation Learning and Embeddings

High-cardinality categorical fields such as merchants or devices present challenges for traditional encodings. Lin (2025) shows that learned embeddings can capture structure that one-hot encoding misses, while Alarfaj and Shahzadi (2024) report improved stability when embeddings are combined with boosted trees. The present study

does not introduce embeddings because the public dataset does not contain high-cardinality identifiers. The contrast with Lin (2025) and Alarfaj and Shahzadi (2024) clarifies the scope: this work uses PCA-based features and a cost-aware framework as a baseline. Representation learning is acknowledged as a future extension rather than part of the central contribution.

2.14 Governance, Fairness, and Audit Trails

Regulators expect traceability. A minimal governance pack for fraud models includes a model card, a cost policy document that states the α and β ranges, an explanation parity report comparing synthetic and real SHAP, and a drift monitor for prevalence and precision at the policy recall. Documentation should also capture the data lineage, feature transformations, calibration steps, and how thresholds were selected. These elements support change management for model updates and threshold adjustments, as described in Nicholls et al. (2021) and Psychoula et al. (2021). In this study, the same governance ideas are applied through fixed train, validation, and test partitions, a documented threshold selection procedure, and SHAP parity checks using synthetic validation data. This mirrors the documentation practices recommended by Nicholls et al. (2021) but adapts them to a single-institution analytic environment rather than a multi-party operational setting.

2.14.1 Fairness Checks and Redress

Fairness checks should be embedded within the governance cycle. Subgroup precision, recall, and false-positive rates can expose disparities that require mitigation. Awosika et al. (2024) show how fairness reporting can accompany cost and drift

monitoring so that decisions remain defensible. When disparities occur, teams may retrain with re-weighting, penalize subgroup gaps in the loss, or adjust thresholds by segment. Redress mechanisms should allow customers to challenge decisions through plain-language explanations. In this research, fairness checks serve as guiding principles rather than implemented metrics, because the public creditcard.csv dataset lacks demographic attributes. Fairness is therefore addressed indirectly through cost-aware thresholds and privacy-preserving explanations, which aligns with the constraints described by Awosika et al. (2024).

2.14.2 Adversarial robustness

Fraud is adversarial. Attackers probe thresholds and try to mimic legitimate behavior. Robustness checks include stress testing with scenario perturbations, reviewing feature importance for dependencies on manipulable fields, and verifying that small changes in the input do not flip predictions unexpectedly. Rao et al. (2025) emphasize that robustness measures should be part of governance rather than optional enhancements. This study uses governance artifacts to connect experimental findings with operational policy. This emphasis is closer to the governance-aware perspective in Nicholls et al. (2021), who argue that models, thresholds, and explanations must be reviewed as a single package. This contrasts with technical studies that evaluate models only through metrics and do not link them to controls required for regulated environments.

2.15 Summary and Gap Statement

The reviewed literature provides a broad set of models and extensions for fraud detection, yet two persistent gaps remain. First, most studies do not use strict leakage-controlled or cost-aware evaluation protocols. For example, several ensemble-based studies benchmark performance using only AUC or F1 scores without documenting how thresholds were selected or transferred to validation and test sets. Second, few investigations present privacy-compatible explainability workflows capable of producing stable, regulator-traceable interpretations, a limitation noted by Rocha et al. (2025) and Shakeri et al. (2024).

This study addresses these gaps by integrating ensemble learning under consistent evaluation settings, CTGAN-based augmentation applied only to the training partition with audit checks, and a cost-aware threshold selection process that minimizes $\alpha \cdot FP + \beta \cdot FN$. SHAP explanations are validated through synthetic-to-real parity checks to support privacy and transparency.

These design choices create a reproducible and privacy-conscious framework aligned with RQ1 through RQ4 and H1 through H4. They also guide the ablations in Chapter 4 and the interpretation in Chapter 5. Compared with studies that emphasize a single technique or propose a new model, this praxis demonstrates how existing components can be wired together under cost, governance, and leakage-control requirements that match operational fraud detection.

2.16 Concept Drift and Adaptive Learning

Fraud behavior shifts with incentives, defenses, and regulatory changes. Static models decay, sometimes abruptly. Concept drift may involve changes in feature distributions, label prevalence, or the relationship between predictors and outcomes. Awosika et al. (2024) describe several adaptive strategies, including scheduled retraining and threshold rebalancing. This study acknowledges these challenges but uses static models on a fixed data partition, consistent with the limitations of the public dataset. Drift-handling techniques are reviewed here to place the study within the broader literature while clarifying that they remain future extensions rather than part of the core experimental setup.

2.16.1 Scheduled Retraining and Sliding Windows

One approach is to retrain on a sliding window that prioritizes recent data. Psychoula et al. (2021) discuss how drift detectors can trigger retraining when calibration or SHAP distributions shift, or when precision at the target recall declines. In contrast, this study uses a single fixed 60/20/20 split to produce a snapshot rather than a rolling window. Even so, the cost curves and SHAP summaries in Chapter 4 are structured so that organizations could later integrate them into a scheduled drift-aware retraining routine.

2.16.2 Online Updates and Threshold Adaptation

A second approach is to update the operating threshold more often than the model. Rocha et al. (2025) show that calibrated probabilities allow thresholds to be rebalanced as prevalence shifts. This is fast to implement and can be combined with a periodic model

refresh. The praxis does not implement online threshold adjustments. Instead, it chooses a single threshold on the validation split and applies it to the test set without change. This choice keeps the experiment simple while demonstrating how cost curves could support threshold adaptation in a real deployment.

2.16.3 Reinforcement Learning for Policy Control

A third strategy treats threshold selection as a sequential decision problem. Reinforcement learning can update thresholds based on realized cost, using a reward function that reflects financial loss and constraints that prevent unsafe exploration. (Cheng et al., 2024) demonstrate how policy-gradient methods can adjust decision thresholds in response to changing base rates and shifting fraud patterns.

This study does not implement reinforcement learning because the public creditcard.csv dataset does not provide the streaming or temporally indexed setting required for sequential policy updates. Instead, the study uses a static threshold chosen on the validation split. The difference is that Cheng et al. (2024) evaluate threshold movement over time, while the current research focuses on validating a single cost-aware operating point under strict leakage control. This keeps the evaluation aligned with the study's goals of assessing CTGAN augmentation, anomaly-feature fusion, and synthetic-explainability parity under a stable and reproducible protocol.

2.17 Diffusion Models and Generative Extensions

While GANs have been the dominant generative approach for tabular fraud data, diffusion models are emerging as a credible alternative. Diffusion models learn to denoise

samples step by step and can capture complex, multimodal distributions using likelihood-style training. Tayebi and Kafhali (2025) show that diffusion models can be conditioned on selected features to respect categorical structure and often train more reliably than adversarial methods.

Although diffusion models offer clear potential, this study relies on CTGAN because it is more mature for tabular financial data and is already validated in fraud-related research such as Zhu et al. (2024). The contrast is intentional. While diffusion may eventually outperform GAN-based tabular synthesis, CTGAN represents a practical and well-documented choice that can be evaluated rigorously under the audit and cost policies used in this study.

2.17.1 Potential advantages for fraud

Diffusion models may reduce mode collapse and can generate more diverse minority-class samples, especially for rare transaction patterns. Zhu et al. (2024) demonstrate that diffusion-based generators can recover complex joint structures that GANs sometimes miss. Hybrid diffusion models that incorporate conditional encoders provide additional control over feature combinations during synthesis.

This praxis does not adopt diffusion models because CTGAN is already well-established for this type of tabular fraud data. Diffusion models are therefore presented as a forward-looking alternative that could extend the current methodology once the CTGAN audit framework and cost evaluation pipeline are fully validated.

2.17.2 Evaluation discipline still applies

Diffusion-based synthesis would still require the same safeguards recommended in Strelcenia and Prakoonwit (2022) and Oreški (2023), including marginal checks, dependency tests, out-of-support filtering, and discriminator-based drift screening. Shakeri et al. (2024) highlight that such audits remain crucial even when advanced generators are used in regulated environments.

This study remains with CTGAN and treats diffusion as a conceptual extension. The comparison with diffusion is therefore limited to framing how future work could build on the same audit discipline and cost-aware evaluation rather than introducing new generators into the present experimental pipeline.

2.18 Causal Explainability and Counterfactual Analysis

Post-hoc feature importance explains correlations, but it does not establish causation. Causal explainability methods evaluate how predictions change under controlled interventions. Psychoula et al. (2021) note that counterfactual analysis can show which minimal, plausible changes would reverse a fraud decision. In operational settings, these tools can help analysts understand decision boundaries and provide actionable guidance to improve investigative triage.

This study does not implement causal methods because the public dataset lacks the richer attributes normally required for structural or interventional modeling. Instead, the study focuses on SHAP-based explanations and synthetic-validation parity to preserve clarity and privacy while keeping the evaluation reproducible.

2.18.1 Methods and Safeguards

Causal methods generally rely on structural causal models or counterfactual generators constrained to realistic changes. Nicholls et al. (2021) stress that counterfactual explanations must exclude protected attributes and should avoid unrealistic or unethical interventions. As a result, causal tools are most effective as internal validation instruments for boundary stability and fairness. This study acknowledges these safeguards conceptually but does not implement them. The focus remains on CTGAN auditing, cost-aware thresholding, and SHAP-based interpretability.

2.18.2 Linking to Fairness

Causal tools can determine whether protected variables or their proxies exert undue influence. Awosika et al. (2024) show that combining SHAP with counterfactual analysis can reveal whether certain features introduce disparate impact. This deepens the fairness narrative by moving from descriptive rankings to intervention-aware explanations. The present work remains within descriptive SHAP applied to real and synthetic validation sets, framing causal methods as a future extension once richer institutional attributes are available.

2.19 Fairness, Bias Mitigation, and Ethical AI

Fairness is central in fraud detection because both false positives and false negatives have human and financial consequences. Nicholls et al. (2021) describe how disproportionate false positives can burden particular customer segments, while disproportionate false negatives expose institutions to uneven risk. Psychoula et al.

(2021) emphasize that fairness obligations extend to both model development and operational monitoring.

Because the creditcard.csv dataset does not include demographic attributes, this study incorporates fairness principles indirectly. Cost-aware thresholds and privacy-preserving explainability help avoid operational bias even when direct subgroup analysis is not possible. This approach stays consistent with fairness expectations while remaining grounded in the limits of the dataset.

2.19.1 Measuring Fairness

Fairness should be evaluated using rare-event-sensitive metrics. Awosika et al. (2024) recommend subgroup precision, subgroup recall, and subgroup false-positive rates rather than overall accuracy, since accuracy obscures disparities in low-prevalence settings. They further emphasize that fairness must be monitored over time to catch drift-induced disparities.

In this study, subgroup fairness cannot be computed directly. Instead, fairness is reflected in stable performance at the chosen cost-aware threshold, combined with explainability checks that discourage reliance on noisy or potentially biased surrogate features.

2.19.2 Mitigation Strategies

Fairness mitigation can occur during data collection, model training, or post-processing. Psychoula et al. (2021) describe strategies such as re-weighting, adjusting

loss functions to penalize disparities, or calibrating thresholds across subgroups. Nicholls et al. (2021) stress that all decisions must be explainable and documented for oversight.

Since demographic attributes are unavailable here, the mitigation strategy centers on calibrated probabilities, fixed cost-aware thresholds, and transparency through SHAP-based summaries. This aligns with Awosika et al. (2024), who argue that responsible deployment requires documentation and interpretability even when demographic fairness cannot be computed directly.

The difference lies in the type of data available. Psychoula et al. (2021) and Awosika et al. (2024) evaluate fairness using explicit demographic slices, while this study cannot perform those analyses because no protected attributes exist in the public dataset. As a result, fairness is reflected indirectly through consistent cost-aware thresholding, calibration, and SHAP-based explanations. This study treats more detailed fairness interventions as future work once richer institutional attributes become available.

2.20 Cross-Family Synthesis and Theoretical Implications

The literature points toward convergence across model families. Ensembles provide stable discrimination on tabular data, while deep models contribute temporal and contextual awareness. Generative models expand minority coverage, and federated learning extends reach without compromising privacy. Rocha et al. (2025) and Pundkar and Zubei (2023) report that the largest gains often arise not from new architectures but from disciplined evaluation, careful calibration, cost policy alignment, and strong governance. This suggests a theoretical stance in which both architecture and process

design hold equal weight. The decision rule becomes a coordinated product of the learner, the calibration layer, and the cost policy, and all three require tuning and audit checks.

This cross-family view supports RQ1 through RQ4 and connects directly to the methodology described in Chapter 3. Pundkar and Zubei (2023) highlight that improvements in fraud detection often stem from choices about thresholding and evaluation rather than from changes in the classifier itself. This study follows that guidance by placing cost-aware thresholding and strict leakage control at the center of its design rather than proposing a new model architecture.

The present work provides a concrete example of this cross-family synthesis. It combines ensembles, CTGAN-based synthetic augmentation, anomaly features, and SHAP explainability within a single evaluation and governance framework. This orientation contrasts with model-centric contributions that introduce a new classifier but provide limited detail on calibration, cost policy, or privacy constraints. The contribution here is to demonstrate how established components can be aligned into a unified, reproducible pipeline that addresses both operational and regulatory needs.

2.21 Emerging Research Directions and Future Trends

Several research directions appear promising.

Self-supervised learning. Contrastive and masked modeling approaches can learn strong representations from unlabeled transactions, reducing dependence on scarce fraud labels and improving downstream classifiers (An et al., 2024).

Multimodal fusion. Combining tabular signals with text, voice, device biometrics, or customer interaction logs can enrich feature quality. This expansion must be paired with explainability and privacy safeguards (Alarfaj & Shahzadi, 2024).

Graph attention and scalable inference. Graph attention networks and efficient neighbor sampling make real-time scoring on large interaction networks feasible, which supports streaming or session-based fraud analysis (Motie & Raahemi, 2023).

Differential privacy and secure computation. Advances in privacy protection continue to improve multi-institution collaboration without exposing sensitive records (Zhang et al., 2023).

Governance automation. Embedding fairness checks, drift monitors, and explanation parity tests into production pipelines can reduce operational overhead and ensure continuous compliance (Awosika et al., 2024).

These forward-looking developments extend the ideas behind H4 and anticipate systems that blend generative augmentation, adaptive learning, and explainable governance (Shakeri et al., 2024). The unifying theme in this literature is that progress is less about a single novel algorithm and more about orchestrating components that work together under real-world constraints.

Relative to these emerging directions, this praxis intentionally focuses on a smaller scope. The aim is to determine what can be achieved today with ensembles, CTGAN, anomaly fusion, and SHAP when evaluated within a disciplined cost and governance framework on a well-known public dataset. This provides a practical baseline. Future work can build on this foundation by adding self-supervised methods,

multimodal data sources, or differential privacy techniques once richer data become available.

2.22 Summary

This chapter reviewed the evolving landscape of AI and machine learning for fraud anomaly detection through the lenses of imbalance, cost, explainability, and privacy. Four practical themes emerged.

First, evaluation must be leakage-resistant and explicitly cost-aware, with thresholds optimized to minimize expected loss at a defined recall floor.

Second, CTGAN and related synthetic generators can extend minority-class coverage when applied only to training data and governed by pre-merge audits that confirm real-data precision and cost stability.

Third, integrating anomaly signals such as Isolation Forest depth and Autoencoder reconstruction error into supervised models can enrich detection features without compromising recall.

Finally, explainability should be privacy-attuned, using synthetic-validation parity checks and structured governance artifacts that support regulatory transparency, as emphasized by (Psychoula et al., 2021) and (Tayebi & Kafhali 2025)
These themes map to the research questions and hypotheses leading this study:

- RQ1 / H1 – Cost-aware, method. Establish a method that translates model outputs into financial terms by minimizing $\alpha \cdot FP + \beta \cdot FN$ on validation, reducing operational loss relative to metric-based thresholds.

- RQ2 / H2 – Synthetic augmentation (CTGAN). Evaluate whether a controlled, audited CTGAN-based augmentation can enhance minority-class representation without degrading precision or increasing total Cost at t^*, thereby testing the practical limits of synthetic data in fraud modeling.
- RQ3 / H3 – Anomaly-feature fusion. Determine whether integrating unsupervised anomaly scores from Isolation Forest and Autoencoder models into supervised ensembles can reduce Cost at t^* while maintaining recall, showing that richer feature fusion strengthens cost efficiency.
- RQ4 / H4 – Privacy-safe explainability. Assess whether a synthetic-validation twin can reproduce real-data feature-importance rankings closely enough for regulator-ready interpretability while safeguarding customer privacy.

Chapter 3 translates these research directions into a reproducible design by outlining the data partitions, CTGAN configuration and audit process, ensemble parameters, anomaly-feature generation, and the cost-aware thresholding routine. It also specifies bootstrap confidence intervals and SHAP-parity validation so that Chapter 4 can report results with statistical consistency and governance-grade credibility (Rocha et al., 2025; Abdelghafour et al., 2024).

Compared with the literature base summarized here, the praxis positions itself as a disciplined synthesis of known components rather than a proposal of a single new algorithm. Its contribution lies in showing how cost-aware policies, CTGAN-based synthesis, anomaly fusion, and privacy-preserving explainability can be combined and audited within a unified evaluation pipeline. These choices speak directly to the needs of fraud teams, risk committees, and regulators. Together, they position the study as a

structured response to specific gaps identified in recent work. Instead of introducing a new model architecture, the contribution rests on coordinating cost-aware evaluation, synthetic auditing, and explainability into one reproducible and governance-aligned framework.

Chapter 3: Methodology

3.1 Purpose and Problem–Solution Framing

This chapter translates the research questions and hypotheses from Chapter 1 into an executable protocol that can be audited, reproduced, and adapted to institutional settings. The study aims to build and evaluate an auditable fraud anomaly detection pipeline that addresses four recurring problems at once. The first problem is severe class imbalance and the resulting brittleness of models that see too few positive examples during training. The second problem is the gap between common leaderboard metrics and the economic reality of false positives and false negatives. The third problem is the requirement for explanations that are stable, repeatable, and acceptable to regulated environments. The fourth problem is data privacy, which often prevents open sharing of transaction-level records across institutions.

The evaluation proceeds through three structured phases that share a fixed data partition and a consistent cost-aware policy. **Phase A** constructs supervised baselines to set a clear yardstick for accuracy and calibration on tabular transactions, addressing **H1** and **RQ1**. **Phase B** introduces **CTGAN** augmentation on the training split to test whether synthetic minority coverage improves detection while preserving calibration and interpretability, addressing **H2** and **RQ2**. **Phase C** adds an **anomaly-feature fusion** layer that appends unsupervised signals to the strongest supervised learner, thereby increasing novelty sensitivity without discarding the dependable ensemble backbone, addressing **H3** and **RQ3**. In parallel, the study evaluates **explainability parity** between real and synthetic validation using **SHAP** feature importance and local reason codes, addressing **H4** and **RQ4**.

The pipeline is intentionally quantitative and relies on public or synthetic data to preserve reproducibility and avoid proprietary dependencies. This decision aligns with the direction in the literature that combines generative modeling with cost-aware operating policies to reduce expected loss and stabilize alert volumes under drift (Backiyalakshmi & Umadevi, 2025; Devi et al., 2024; Geng & Zhang, 2023; Lin, 2025; Mienye & Swart, 2024; Zhu et al., 2024). The chapter lays out the dataset, the locked split, the feature handling, the CTGAN generator and audits, the supervised and unsupervised learners, the cost policy, the metric suite, the explainability workflow, the ablation plan, and the statistical methods for uncertainty. It closes with threats to validity, fairness and ethics considerations, deployment notes, and a reproducibility blueprint that matches the artifact list in Chapter 4.

3.2 Data and splits

3.2.1 Exploratory Data Analysis

A preliminary exploratory review was conducted to understand the structure and behavior of the dataset before modeling. Because most of the features (V1 through V28) were generated through a dimensionality-reduction process prior to release, the dataset did not require extensive cleaning. However, descriptive statistics and distribution checks helped confirm several characteristics that shaped the preprocessing decisions used in the study.

First, the data exhibited a severe class imbalance (see Figure 1). Fraudulent transactions represented well under one percent of all observations, consistent with findings reported across financial fraud datasets in the literature. Researchers have

emphasized that such rarity makes traditional accuracy metrics unreliable and encourages the use of evaluation approaches that prioritize minority detection (Hashemi et al., 2022; Olowu et al., 2024; Pundkar & Zubei, 2023). This confirmed that precision, recall, AUC-PR, and cost metrics would be more informative than accuracy alone.

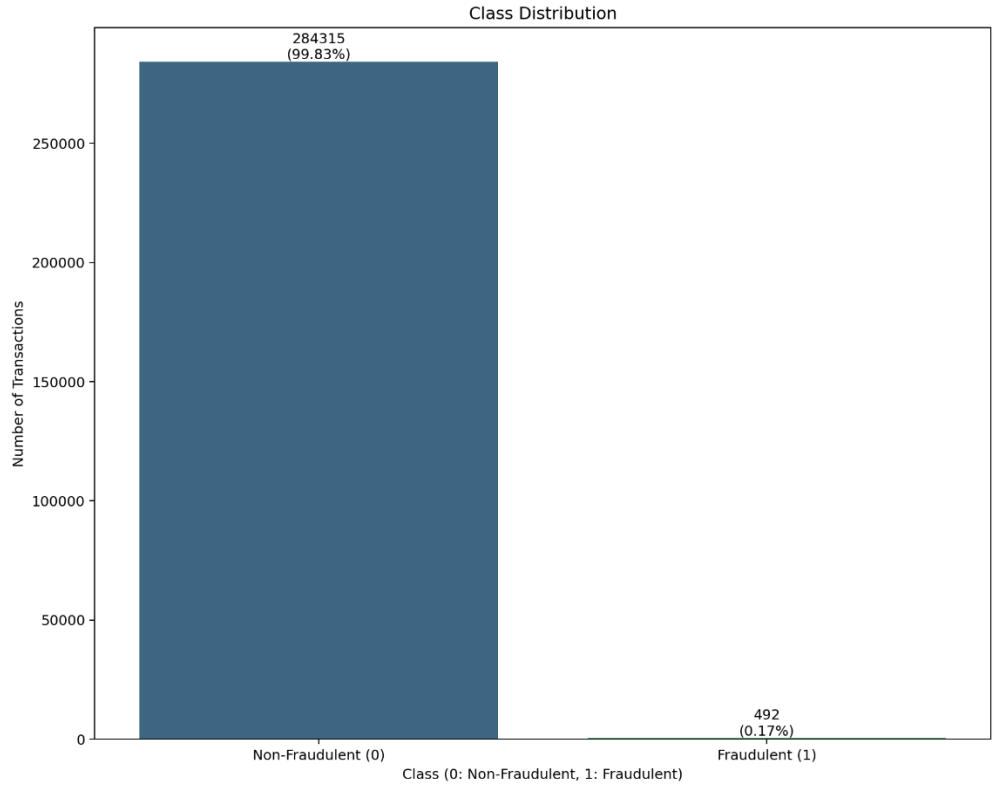

Figure 1. Class Distribution of the creditcard.csv Dataset

Second, the distribution of transaction amounts showed a strong right-skew with a small number of unusually large transactions. Prior studies of financial-transaction data note that heavy-tailed variables can distort normalization if aggressive scaling is applied (Kennedy et al., 2023; Lange et al., 2022). To keep preprocessing simple and to avoid altering the PCA-derived structure of the other features, a standard scaling transform was

applied only to the Amount field before model fitting. This reduced the influence of extreme numeric values while leaving the rest of the dataset unchanged.

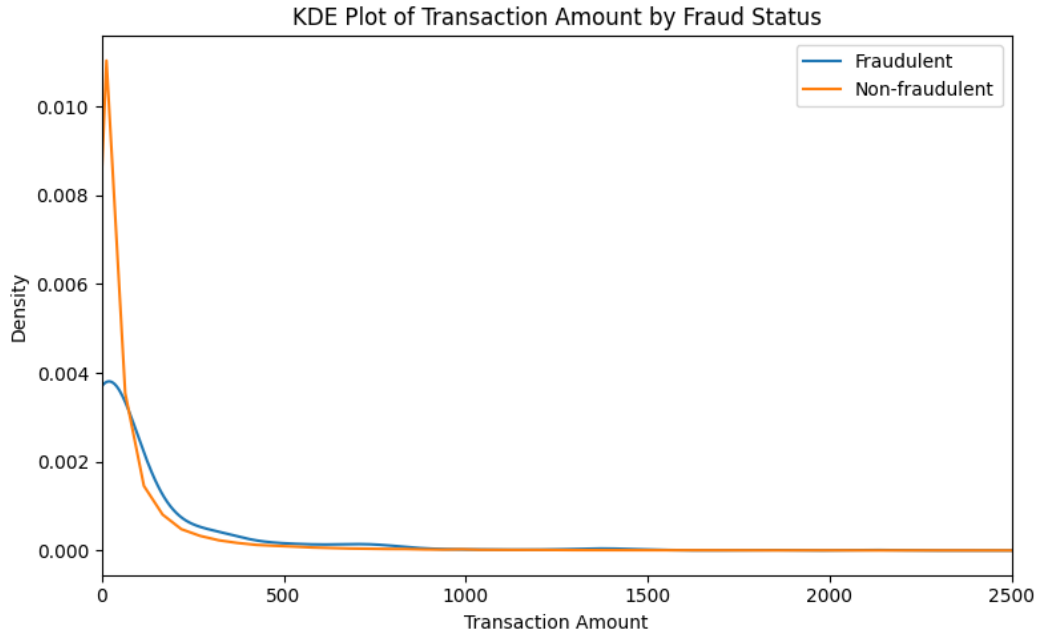

Figure 2. KDE Plot of Transaction Amount by Fraud Status

Third, simple correlation checks confirmed that the principal component features were mostly uncorrelated. This aligns with prior descriptions of the benchmark credit card dataset, where PCA components were designed to reduce correlation and mask sensitive customer information (Leevy et al., 2023; Nicholls et al., 2021). As a result, additional decorrelation or dimensionality reduction steps were not necessary.

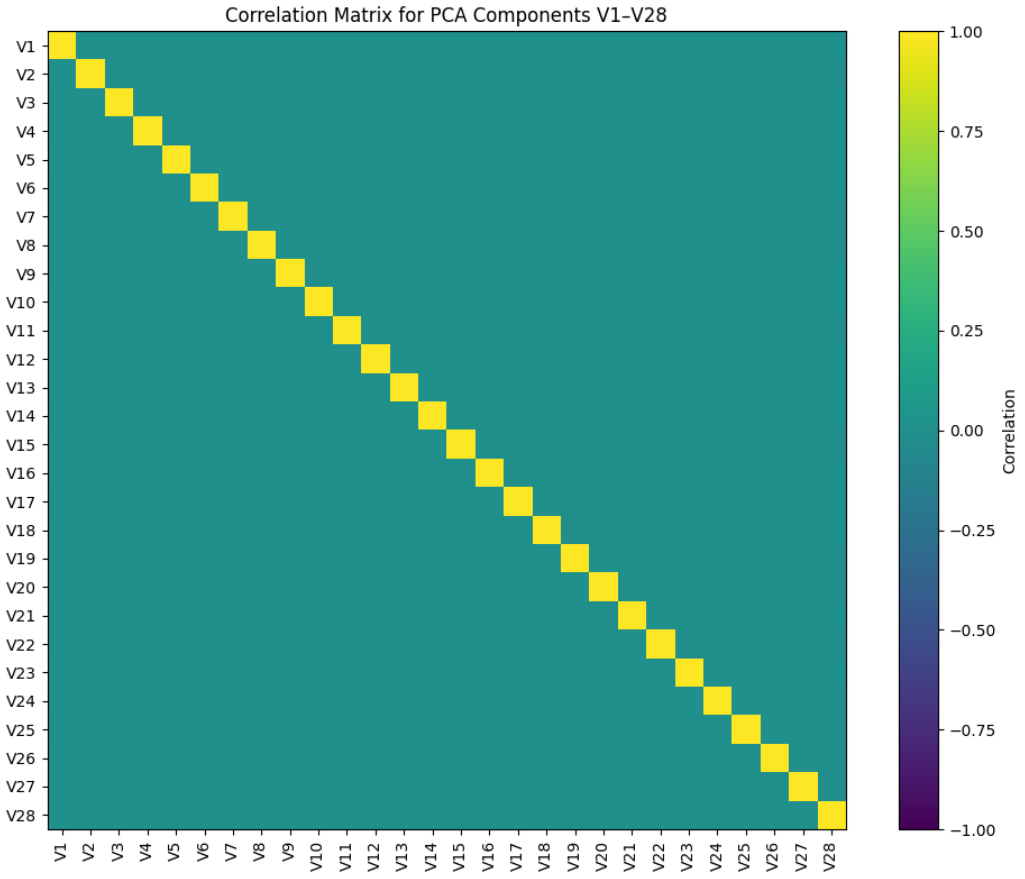

Figure 3. Correlation Matrix for PCA Components

Finally, the Time feature was examined to determine whether temporal patterns could support session or sequence modeling. The field behaves as a simple elapsed-time index and does not encode customer identifiers or chronological timestamps. Reviews of this dataset note that the Time field is limited in this way and therefore less suitable for temporal pattern extraction without additional contextual information (Hashemi et al., 2022; Pundkar & Zubei, 2023). Given this behavior, no new time-windowed features were created, although such features could support future work on streaming or sequential fraud detection.

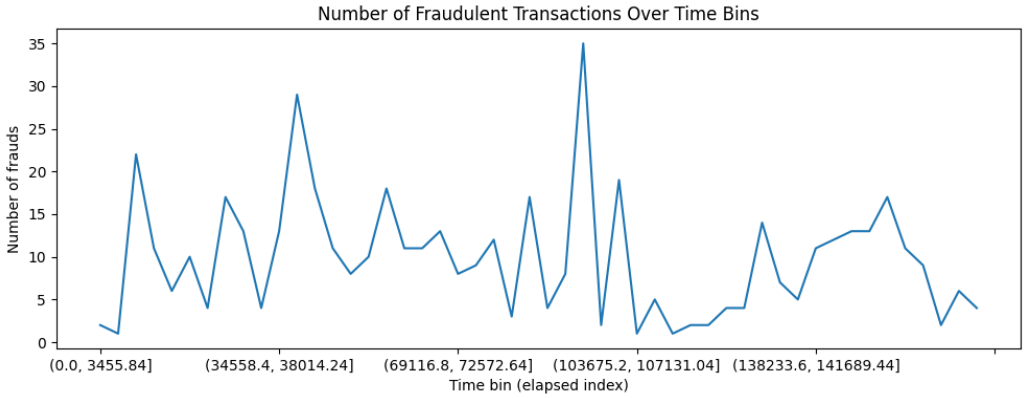

Figure 4. Number of Fraudulent Transactions Over Time

Overall, the exploratory analysis showed that the dataset was structurally sound and required only targeted normalization and validation checks, rather than significant imputation, feature repair, or deduplication. This allowed the study to focus on the modeling contributions: baseline supervised learning, synthetic augmentation, anomaly-feature fusion, and cost-aware thresholding.

These checks form the common exploratory baseline for all models reported in Chapter 4, so every learner is trained and evaluated on the same well-understood version of the dataset. Model-specific exploratory checks, including precision–recall curves, cost-versus-threshold plots, confusion matrices, and slice analyses, are reported in Chapter 4 for each configuration.

3.2.2 Sources and Data Characteristics

The primary corpus is the widely used reference dataset for credit card fraud detection. It contains hundreds of thousands of transactions and a very small minority of confirmed fraud events. Attributes include multiple principal component transformed

variables and a transaction amount field, along with a binary label. This design provides a realistic though anonymized testbed that encapsulates the two characteristics that most strongly shape the modeling problem. The first is the very low prevalence of fraud, commonly around two tenths of one percent, which renders accuracy meaningless and makes threshold policy central to evaluation. The second is the presence of mixed data types and skewed marginals that challenge naive oversampling and encourage principled synthesis approaches such as CTGAN (Becerra-Suarez et al., 2025; Lee et al., 2024; Leevy et al., 2023).

In **Phase B**, a CTGAN generator learns from the minority rows in the training split and produces additional positive-class samples. Conditioning is used for high-impact categoricals and binned numeric features so that synthetic events reflect realistic combinations rather than isolated attribute resampling. Prior work suggests that conditional tabular GANs can preserve joint structure and improve recall at constant precision when audits are used and generation is limited to the training partition (Mienye & Swart, 2024; Lee et al., 2024; Zhu et al., 2024).

3.2.3 Partitioning and Leakage Control

A leakage-controlled three-way split is used.

- **Training:** 60 percent
- **Validation:** 20 percent
- **Test:** 20 percent

Stratification preserves the rare prevalence across partitions. All preprocessing statistics, encoders, calibrators, augmentation, and models are fit on the training split only. Threshold selection is performed on validation only. The test split is locked and evaluated exactly once for final reporting. There is no synthetic data in validation or test. This discipline addresses the known set of pitfalls in fraud studies where transformations are inadvertently fit on the full data set, where temporal windows allow future information to leak back, or where thresholds are moved after the fact to dress up metrics (Hashemi et al., 2022; Hayat & Magnier, 2025; Leevy et al., 2023).

3.2.4 Rationale and Link to Research Questions

The locked split and train-only synthesis strategy protects the causal interpretation of observed gains. If recall improves at constant precision in Phase B relative to Phase A, the design provides a credible basis to attribute the improvement to minority coverage rather than to information leaks. These choices directly address **RQ1** and **RQ2** and provide fair conditions for testing **H1** and **H2**.

3.3 Preprocessing and Feature Handling

3.3.1 Numeric Scaling and Outlier Stability

For this study, only the Amount field is scaled. A standard-scaling transform (mean zero, unit variance) is fitted on the training split and applied to validation and test. This keeps the PCA-derived components V1–V28 in their original form while reducing the influence of very large transaction amounts. In more general deployments, the same pipeline can be configured to use robust scaling based on medians and interquartile

ranges for additional numeric fields, but that option was not activated for the benchmark dataset used here.

3.3.2 Categorical Encoding and Optional Embeddings

Where high-cardinality keys are present, the pipeline uses target encoding trained only on the training split and computed within cross folds to avoid label leakage. This approach provides more signal than one-hot encodings while handling sparsity more gracefully. If semi-textual identifiers exist, an optional embedding track can be added to map categories into dense vectors. This is in line with the representation methods described in Chapter 2, in which language-model-based or graph-based embeds increased the accuracy of supervised classifiers on relational similarity between merchants, devices, or user agents (Bakumenko et al., 2024; Lin, 2025; Kadam, 2024; Thilagavathi et al., 2024).

3.3.3 Temporal Features and Velocity

From timestamps, the pipeline derives hour of day, day of week, a weekend flag, and short-window velocity statistics such as count and spend per entity over past minutes or hours. Each feature is computed strictly from information available at or before the event to prevent look-ahead bias. Velocity is clipped at reasonable maxima to avoid unstable leverage by a few outliers. These features support both supervised and unsupervised learners by describing routine behavior and deviations from it.

3.3.4 Data Conditioning and Rare Value Treatment

Although the reference dataset used in this study is effectively complete after the PCA transformation and contains no missing values or high-cardinality categorical keys, the broader pipeline includes several lightweight conditioning steps that support generalization to richer amounts. When future datasets introduce missingness, the approach favors explicit missing-value indicators rather than aggressive imputation, which aligns with recommendations for transparency in credit-risk preprocessing (Nicholls et al., 2021). Rare categorical values can be grouped into an “other” bucket within training folds to reduce instability from extremely small categories. More assertive outlier treatments, such as winsorization, are reserved for cases where domain review suggests that extreme values are likely to be data errors rather than legitimate high-magnitude transactions. None of these conditioning steps were required for the present dataset, but they are included here to show how the methodology scales to less curated transaction amounts.

3.3.5 Preprocessing Alignment to RQ1- RQ3

The preprocessing design is deployable and conservative: it avoids brittle transforms that amplify drift while preserving enough structure for stable discrimination on the test set. These choices support RQ1 by enabling fair ensemble baselines and clean handoff of the validation-selected operating point to test (in concert with calibration and cost-based policy selection in §3.7). Preprocessing also prepares a consistent feature space for later anomaly-feature fusion (§3.6), thereby setting up RQ3 without constraining its design.

For the present credit card dataset, only a simple standard scaling transform was applied to the Amount field. The additional conditioning options described in §3.3.2–3.3.4 serve as part of the general deployment blueprint and were not triggered by this dataset, which contains no high-cardinality categorical fields, no missing values, and no time-structured identifiers. Their inclusion ensures that the methodology is extensible to more complex fraud amounts while keeping the experimental execution faithful to the available data.

3.4 CTGAN Synthesis and Audits

3.4.1 Generator Objectives and Configuration

CTGAN is trained on the minority-class rows in the training partition. The generator learns to produce synthetic transactions that are indistinguishable from real minority events to a discriminator that is trained simultaneously. Conditioning variables are chosen from high-influence categoricals and discretized numeric features to preserve interactions. Training follows the standard adversarial objective with early stopping based on discriminator loss stability and a small holdout slice of real minority rows within training to monitor overfitting. The target minority ratio in the augmented training pool is set between five and fifteen percent. This range provides a practical balance between exposure to rare patterns and preservation of the original class balance that the model will see in production. Validation and test sets remain entirely real to preserve an unbiased estimate of generalization (Mienye & Swart, 2024; Lee et al., 2024; Zhu et al., 2024).

3.4.2 Three-Part Audit to Prevent Drift and Memorization

Before any synthetic batch is merged with training, three audits are applied.

1. **Marginal alignment.** For each numeric and categorical feature, a univariate test compares the real and synthetic training distributions. The Kolmogorov–Smirnov test is used for continuous variables. The Anderson–Darling variant is used when heavier tails are suspected. For categoricals, absolute frequency differences are monitored along with entropy comparisons.

2. **Pairwise dependence.** A distance-correlation or mutual-information matrix was estimated for selected feature pairs, with Spearman's rank correlation (ρ) offering a resilient alternative when monotone but non-linear effects are expected (Abdi, 2023; Chen et al., 2024). Spearman ρ and its two-sided p-values were computed using SciPy (Virtanen et al., 2020). The synthetic matrix should remain within a tolerance band of the real training matrix. This check ensures that the generator has learned interactions rather than only local marginals.

3. **Model invariance.** A small classifier attempts to distinguish real minority rows from synthetic rows. An AUC near 0.50 is preferred, and any AUC notably above 0.65 triggers rejection or retraining. This is a practical proxy for detecting overly neat synthetic patterns that are not present in the real distribution.

Only batches that pass all thresholds are merged, and the merge ratio is checked against a cap to prevent flooding the learner with synthetic events. This procedure responds to frequent critiques of synthesis, including drift, mode collapse, and privacy

concerns, by making the acceptance criteria explicit and measurable (Nakada et al., 2024; Geng & Zhang, 2023).

3.4.3 Privacy Posture

The generator is trained on anonymized data within the training partition, and no real records are exported. The accepted synthetic rows are not linked to specific real events, and the audits ensure that no low-frequency exact duplicates are introduced. The design is consistent with privacy-aware collaboration models where only parameters or differentially private summaries are shared across institutions (Shakeri et al., 2024). Although the present study does not execute a federated protocol, the artifacts can be lifted into such an environment with minimal changes, which supports **H4** and **RQ4** as a future extension.

3.4.4 Expected Outcome and Link to Hypotheses

If CTGAN expands minority coverage in training without distorting interactions, ensemble learners should exhibit improved recall at a fixed precision and a reduced expected cost at the chosen policy point. Phase B directly tests **H2** by comparing Phase A and Phase B at the cost-optimal threshold selected on validation and then frozen for test.

3.5 Supervised Learners and Calibration

3.5.1 Baseline Learners

The supervised backbone comprises **Logistic Regression**, **Random Forest**, **Extra Trees**, and **XGBoost**. Logistic Regression supplies a transparent and effective baseline that supports sanity checks. Random Forest and Extra Trees trim down variance

through bagging and randomized splits, respectively. Gradient boosting with XGBoost captures nonlinear connections and is robust on tabular data with minimal feature engineering. This set reflects the state of the art for tabular fraud detection and offers a fair benchmark of the effect of synthesis and hybridization (Wang et al., 2025; Abdelghafour et al., 2024; Almalki & Masud, 2025; Hashemi et al., 2022; Zheng et al., 2024).

3.5.2 Class Weighting and Imbalance Parameters

For boosted learners, the scale of positive class weights is set to the observed imbalance in the training partition. For Random Forest and Extra Trees, a balanced subsample weighting is used to avoid dominance by the majority class while preserving the integrity of the original distribution. The class weight configuration is recorded for reproducibility.

3.5.3 Probability Calibration

Uncalibrated ensembles can produce overconfident scores, which undermines threshold transfer. For each fitted model, a calibration step is applied on validation. If isotonic regression improves log-loss and AUC-PR stability, isotonic is retained; otherwise, Platt scaling is used. Calibrated scores stabilize policy selection by aligning predicted probabilities with observed frequencies in the validation slice. This reduces unexpected swings in alert volumes when the same threshold is applied to the test set and later to production settings with similar prevalence (Tamanna et al., 2024).

3.5.4 Hyperparameters and Compact Search

Hyperparameters are tuned within compact ranges that are fair across families and suitable for repeated ablations.

- Random Forest: 500 trees, max features sqrt, balanced_subsample weights.
- Extra Trees: 1,000 trees, max features sqrt.
- XGBoost: learning rate in [0.03, 0.10], max depth in [4, 8], min child weight in [1, 5], subsample in [0.7, 1.0], scale_pos_weight set to observed imbalance.Search spaces are deliberately modest to keep the ablation interpretable and the compute burden reasonable. This also encourages generalizable solutions rather than brittle hyperparameter artifacts (Hashemi et al., 2022; Tamanna et al., 2024).

Although LightGBM is frequently reported as a high-performing baseline in recent fraud-detection studies, it is not included in this study's experiments. The supervised baselines are restricted to Logistic Regression, Random Forest, Extra Trees, and XGBoost to maintain alignment with the executed notebooks and prevent overstating the empirical scope.

3.6 Unsupervised Anomaly Signals

3.6.1 Isolation Forest and Autoencoder Training

An **Isolation Forest** is trained on legitimate training rows to learn a model of normal isolation depth. A small **autoencoder** is trained to reconstruct inputs and compute

a reconstruction error. Both models operate on the same feature space as the supervised learners, after preprocessing, and both are trained on training data only.

3.6.2 Score Fusion with Supervised Learners

For the strongest supervised learner identified in Phase A or Phase B, two additional features are appended: the Isolation Forest anomaly score and the autoencoder reconstruction error. The pipeline then refits the supervised model with these two features included. The objective is to combine the robust discrimination of ensembles with the novelty sensitivity of unsupervised signals. This hybridization is light weight and deployable and reflects current practice in financial institutions where risk teams add anomaly scores as features rather than deploying entirely unsupervised systems (Kamuangu, 2024; Vynokurova et al., 2020).

3.6.3 Expected Outcome and Link to Hypotheses

If anomaly scores capture signals not already represented by the engineered features, Phase C should improve detection of novel fraud patterns without a disproportionate increase in false positives at the policy point. This phase tests the robustness extension implied by **H3**.

3.7 Cost-Aware Decision Policy

3.7.1 Loss Function and Threshold Selection

The cost-aware policy is the central decision layer that converts scores into an operating point. Let $FP(t)$ and $FN(t)$ be the false positives and false negatives in the

validation set at threshold t. Risk owners specify coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that reflect the cost of investigating false alerts and the expected loss of missed fraud. The operating threshold $\hat{t}$minimizes

$$\text{Cost}(t) = \alpha \cdot FP(t) + \beta \cdot FN(t).$$

An optional recall floor can be imposed to meet minimum regulatory or contractual detection requirements. Once t* is chosen on validation, it is frozen and applied once to the test set. This policy aligns model selection with economic outcomes and avoids after-the-fact threshold adjustments that confound interpretation (Lin, 2025; Cheng et al., 2024; Höppner et al., 2021).

3.7.2 Primary and Secondary Metrics at the Policy Point

Primary metrics at $t = t^*$ on the held-out test set are Precision, Recall, F1, AUC–PR, and **Cost at t^*.**

Secondary diagnostics include total alert volume, positive predictive value among the top k alerts, calibration curves, and decision curves. Together these measures describe statistical quality, operational burden, and expected financial impact. In rare-event settings, area metrics are informative but not decisive. The cost metric and policy point capture the reality that institutions evaluate systems by expected loss and manageable throughput rather than by AUC alone (Lima et al., 2024; Yu et al., 2024; Lin, 2025).

3.7.3 Sensitivity to Policy Coefficients

A grid is executed across plausible ranges for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to generate policy curves for decision-makers. The curves display expected cost and alert volumes as a function of

the chosen pair. This sensitivity analysis contextualizes the model comparison by showing where one approach dominates another under alternative risk appetites. It also provides transparency to stakeholders and regulators about how costs were set and how robust the conclusions are to small changes in assumptions (Cheng et al., 2024).

3.8 Explainability and Governance

3.8.1 SHAP Workflow on Real and Synthetic Validation

For the selected model, SHAP values are computed on the real validation set to produce both local and global explanations. Local explanations take the form of reason codes for representative true positives and high-impact false positives, while global explanations present the top feature rankings along with directionality. A **synthetic validation twin** is created by conditioning CTGAN on the validation marginals. The same SHAP computations are executed on this synthetic twin. The top-k feature lists from real and synthetic validation were compared using both overlap and rank-correlation measures-specifically Spearman's ρ with corresponding p-values-to gauge agreement in explanatory structure (Abdi, 2023; Gourinda et al., 2024; Chen et al., 2024; Virtanen et al., 2020). Agreement within a pre-specified tolerance indicates that synthetic plots preserve the explanatory structure in the real validation set (Shakeri et al., 2024; Backiyalakshmi & Umadevi, 2025; Lee et al., 2024).

3.8.2 Governance Artifacts and Drift Hooks

The pipeline records the following for governance.

1. **Global stability.** Top features and their stability across cohorts and over time.

2. **Local reason codes.** Explanations for a curated set of predictions that matter most to operations.

3. **Drift hooks.** Scheduled recomputation of SHAP summaries and calibration checks on rolling validation windows to detect changes in primary drivers and prevalence.

These artifacts form a minimal governance pack: a model card, a cost policy document with the ranges for **α** and **β**, a SHAP parity report, and a drift monitoring plan. The pack enables audit and transparency without exposing raw records, which is a central requirement for financial risk models in regulated settings (Nicholls et al., 2021). Feature-stability summaries tracked Spearman rank correlations of top-k lists across rolling windows, a concise indicator of whether the model's reasoning remains steady (Gourinda et al., 2024; Chen et al., 2024)

3.9 Training, Tuning, and Evaluation Loop

3.9.1 End-to-end Loop

1. Fit each candidate model on the **training** partition.

2. Calibrate probabilities on the validation set when isotonic regression or Platt scaling improves log-loss or probability alignment.

3. Search thresholds on **validation** to minimize Cost(t), with an optional recall floor.

4. Freeze $\hat{t}$and evaluate once on the **test** set.

5. Compute **95 percent confidence intervals** for Precision, Recall, F1, AUC–PR, and cost using stratified bootstrap on the test predictions with 1,000 resamples.
6. Perform a **paired bootstrap** on prediction vectors to test whether cost differences between systems are statistically nonzero.

For each fitted model, the loop also produces basic exploratory diagnostics at the score level, including score tables, precision–recall curves, cost-versus-threshold curves, and a confusion matrix at t*, and these are summarized in Chapter 4 for every experimental condition.

3.9.2 Implementation Details

All random seeds and software versions are logged. Each run produces a model file, a calibration map, the chosen t*, a confusion matrix at $\hat{t}$, precision–recall and cost curves, SHAP plots for real and synthetic validation, and a manifest with **α**, **β**, and the recall floor. These artifacts support reproducibility and make external audit possible.

3.10 Exact Ablation Plan

Six version-controlled notebooks implement the pipeline. Every ablation uses the same split, the same calibration rules, and the same cost-aware threshold selection.

A. Data treatments

A1: No augmentation.

A2: CTGAN augmentation on train.

Outcome: Change in recall and **Cost at** t^* at constant precision.

B. Model families

B1: Logistic Regression, Random Forest, Extra Trees, XGBoost.

B2: Best baseline plus anomaly features from Isolation Forest and the autoencoder.

B3: Best baseline plus anomaly features with CTGAN augmentation.

Outcome: Incremental value of hybridization and synthesis.

C. Calibration and thresholding

C1: Uncalibrated with a default threshold of 0.50, for reference only.

C2: Calibrated with cost-optimal threshold.

C3: Calibrated with cost-optimal threshold and recall floor.

Outcome: Sensitivity of operating cost to calibration and constraints.

D. Policy coefficients

D1: Grid over stakeholder ranges for **α** and **β**.

Outcome: Policy curves that guide risk owners.

E. Explainability parity

E1: SHAP on real validation.

E2: SHAP on synthetic validation twin.

Outcome: Top-k overlap and Kendall-τ or Spearman ρ (with p-values) were used to quantify explanatory parity (Abdi, 2023; Gourinda et al., 2024; Chen et al., 2024; Virtanen et al., 2020).

F. Robustness

F1: Back test on a later temporal slice if available.

F2: Stress test with a 10 percent prevalence shift created by resampling validation.

Outcome: Stability under time shift and base-rate change.

For the static credit card corpus used in this study, F1 and F2 are documented as optional robustness checks for future multi-period datasets and were not executed as part of the main experimental runs.

3.11 Threats to Validity and Limitations

Internal validity. The locked split and train-only synthesis reduce leakage risk. Nevertheless, anonymization and PCA can hide useful raw attributes and distort some relationships. The synthesis audits mitigate but do not eliminate this risk. Bootstrap intervals quantify sampling uncertainty and help avoid overinterpretation of small differences (Hashemi et al., 2022; Leevy et al., 2023).

External validity. Results on this corpus may not generalize to sectors with different merchant mixes or transaction rhythms. The protocol is written to be portable to other datasets and to cross-institution collaboration through federated learning. The federated extension would address generalization by design, since it would expose models to more varied fraud typologies while keeping data local (Shakeri et al., 2024).

Construct validity. The cost parameters α and β reflect institutional preferences and can vary over time. The sensitivity analysis in Section 3.7 helps contextualize how the conclusions change when these parameters are adjusted.

Conclusion validity. Paired bootstraps on prediction vectors support head-to-head cost comparisons and guard against spurious significance caused by heavy imbalance. All threshold choices are frozen prior to test evaluation to prevent selective reporting.

3.12 Ethical Considerations and Fairness

The study treats explainability as an integral part of governance, not as a cosmetic addition. The parity check enables disclosure that protects privacy while still conveying the reasons for decisions. In future extensions, if protected attributes or credible proxies are available, fairness can be examined using reweighting, equalized performance constraints across subgroups, or adversarial de-biasing. Counterfactual analysis can help assess whether small changes to attributes unrelated to risk would alter outcomes in a manner that is unjust. These steps align with best ethical practice in AI and with regulatory expectations for transparency and fairness in financial decision systems (Nicholls et al., 2021; Vallarino, 2025; Bathala et al., 2025).

3.13 Reproducibility and Experiment Tracking

All experiments are conducted in version-controlled notebooks. Random seeds, software versions, and key hyperparameters are logged to a run manifest. Key plots and tables are saved alongside the trained models. The artifact bundle contains exactly what Chapter 4 needs to regenerate figures and tables. This design supports academic reproducibility and institutional audit and prepares the implementation for regulated production environments.

3.14 Computational Resources and Efficiency

Training is designed to fit within modest compute budgets available to many research groups and financial institutions. CTGAN training benefits from a single modern GPU but can also be run on CPU with more time. Ensemble models are CPU friendly and parallelize across trees. SHAP computations are the most expensive for large samples, so the workflow uses stratified subsampling for global summaries while retaining full fidelity for local explanations on selected cases. Efficiency choices are documented so that production teams can reproduce timing and cost estimates.

3.15 Software Stack and Environment Controls

The pipeline uses widely adopted open source libraries for preprocessing, modeling, calibration, and explainability. Exact versions are pinned in a requirements file. All random seeds for synthesis, learners, and sampling are recorded. The environment is containerized to support reruns and external audit. This approach reduces the risk that small differences in library implementations or defaults will perturb results.

3.16 Equations and Definitions

Let TP, FP, TN, and FN denote counts at threshold t.

- **Precision** $= \frac{TP}{TP+FP}$
- **Recall** $= \frac{TP}{TP+FN}$
- **F1** $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
- **AUC–PR** computed over the full precision–recall curve

- **Cost at** $t^* = \alpha \cdot FP(\hat{t}) + \beta \cdot FN(t^*)$
- **Expected alerts at** $t^* = TP(t^*) + FP(t^*)$

Definitions are applied consistently across all ablations.

3.17 Robustness, Drift Monitoring, and Recalibration

In real fraud-detection environments, concept drift is unavoidable because customer behavior, merchant patterns, and fraud tactics evolve. A practical deployment plan usually includes three forms of drift monitoring. First, prevalence drift is tracked by monitoring the base rate of detected fraud in recent scoring windows. Second, precision drift at the chosen policy point helps identify increases in false positives or reductions in fraud separability. Third, driver drift is monitored by periodically recomputing global SHAP summaries to check whether the dominant explanatory features remain stable.

When drift is detected, institutions typically re-optimize the operating threshold on a recent validation slice, provided calibration remains adequate. If calibration has shifted, the model is recalibrated and the policy threshold reselected. Full retraining is reserved for cases where both calibration and driver drift exceed tolerances.

The present study uses a fixed train/validation/test split on a static benchmark dataset. No rolling-window monitoring or online recalibration was executed. The procedures described here serve as a deployment blueprint for institutions that wish to extend this framework into a live environment (Olowu et al., 2024; Li et al., 2024)..

3.18 Security and Adversarial Resilience

Attackers adapt to detection policies. The methodology therefore encourages evaluation under small adversarial shifts in inputs that approximate realistic manipulation. Examples include deviations in transaction volume within typical thresholds, changes in timing patterns, or device switching behavior. Stress testing with prevalence shifts and with small perturbations helps ensure that the system does not overfit to a narrow fraud profile and can withstand minor obfuscation attempts. This also helps resilience, capturing signals without being tied to any heuristics, via the use of anomaly-feature fusion (Rao et al., 2025; Babu et al., 2024).

3.19 Data Management and Privacy Posture

All experiments operate on anonymized public or synthetic data. Synthetic training rows are generated only from the training partition and pass audits before merging. No personal identifiers are created or stored. Explainability artifacts for disclosure are produced on the synthetic validation twin when parity holds, which avoids exposing real customer cases. These practices reflect institutional expectations for privacy and align with collaborative models that limit data movement across organizations (Nicholls et al., 2021; Shakeri et al., 2024).

3.20 Supplementary Analysis

The methodology outlines several optional extensions that become useful when richer data are available. A representation track can test whether learned embeddings for high-cardinality identifiers improve downstream classification. A graph feature track can incorporate relational indicators such as degree or simple community structures. A

diffusion-based synthesis track can evaluate whether tabular diffusion models offer more stable minority generation than CTGAN.

Because the benchmark dataset is anonymized and PCA-transformed, none of these supplementary tracks were executed. They are included to show how the cost-aware, explainable framework can expand to multi-entity corpora (Giusti et al., 2025; Cheng et al., 2025; Thilagavathi et al., 2024).

3.21 Alignment to Research Questions and Hypotheses

The study is structured around four research questions (RQs) and four corresponding hypotheses (Hs), each mapped to a specific experimental phase in Chapter 4. Together they assess how cost awareness, data augmentation, anomaly-feature fusion, and privacy-safe explainability influence fraud detection performance

- **RQ1 / H1 - (Thresholding):** RQ1: Can cost-aware threshold selection reduce total expected loss compared with fixed thresholds?

  H1: If model thresholds are optimized using a cost function that penalizes missed frauds more heavily than false alerts, then total operational loss will decrease relative to metric-only thresholding.

- **RQ2 / H2 - (CTGAN Augmentation):** RQ2: Can controlled synthetic minority augmentation with CTGAN improve recall without increasing total cost?

H2: If CTGAN-generated minority samples expand minority coverage in training data, then recall will improve and overall cost at t^* will decrease compared with non-augmented baselines.

- **RQ3 / H3 - (Anomaly-Feature Fusion):** RQ3: Do unsupervised anomaly signals enhance supervised model efficiency when fused as additional features?

  H3: If anomaly-feature fusion provides complementary cues about atypical behavior, then cost will decrease or remain stable while preserving recall.

- **RQ4 / H4 - (SHAP Explainability Parity):** RQ4: Can explanations derived from synthetic validation data reflect the same feature importance structure as real validation data?

  H4: If SHAP feature rankings from real and synthetic validations exhibit high overlap and rank correlation, then privacy-safe interpretability is feasible without exposing confidential records.

This alignment clarifies that H1 corresponds to cost-aware thresholding, H2 to CTGAN augmentation, H3 to anomaly-feature fusion, and H4 to SHAP-based explainability.

3.22 Implementation Blueprint

Notebook 0. CTGAN generation. Train CTGAN on training minority rows, run marginal and dependence checks, run the real versus synthetic classifier check, and export approved synthetic rows with a record of seeds and configuration.

Notebook 1. Baselines. Load and split the dataset, preprocess, fit baselines, calibrate on validation if beneficial, search thresholds to minimize cost with an optional recall floor, freeze $\hat{t}$, evaluate on test, and save PR and cost curves.

Notebook 2. CTGAN integration. Merge approved synthetic rows into training, refit the top baselines, recalibrate if needed, select $\hat{t}$on validation, and evaluate on test.

Notebook 3. Anomaly features. Fit Isolation Forest and the autoencoder on legitimate training rows, append anomaly scores as features, refit the best supervised model, and evaluate as above.

Notebook 4. Explainability. Compute SHAP values on real validation and on the synthetic validation twin, compare top-k feature lists and rank correlations, and export parity plots and local reason code exemplars.

Notebook 5. Ablations and confidence intervals. Execute the A through F toggles, compute bootstrap intervals, assemble summary tables, and save a manifest with policy coefficients and threshold choices for traceability.

Artifacts per run include the trained model, the calibration map, the chosen $\hat{t}$, the confusion matrix at $\hat{t}$, PR and cost curves, SHAP plots, and a JSON manifest with **α**, **β**, and recall floor values.

3.23 Statistical Analysis and Uncertainty Quantification

All primary metrics at $t = \widehat{t}$ are accompanied by **95 percent confidence intervals** using a stratified bootstrap on test predictions with 1,000 resamples. For pairwise model comparisons, a **paired bootstrap** resamples prediction vectors and labels jointly so that

correlation between models is preserved. The analysis reports the distribution of cost differences and computes a percentile interval. When intervals exclude zero, the result is treated as statistically significant at the corresponding level. This approach avoids assumptions of normality and is robust in rare-event settings where asymptotic approximations can be unreliable (Leevy et al., 2023). Rank-based associations such as Spearman's ρ were calculated where monotonic relationships were present, using SciPy's implementation for reproducibility (Virtanen et al., 2020).

3.24 Documentation for Audit and Review

The study prepares a minimal yet complete documentation bundle. The bundle contains model cards for the final systems, a cost policy memo that describes how **α** and **β** were set and what ranges were considered, SHAP parity plots, and a drift monitoring plan that outlines how prevalence and primary drivers will be tracked. For cases where synthetic parity is sufficient, the documentation intended for sharing uses synthetic plots rather than plots from real validation. This approach is consistent with transparency guidance and with common expectations for model risk management in finance (Nicholls et al., 2021; Shakeri et al., 2024).

3.25 Deployment Notes and Operations

The pipeline is designed for staged release. In a batch scenario, models are trained and calibrated offline, the threshold is selected on a recent validation slice, and that threshold is then applied to the next scoring window. In a streaming scenario, the calibration and policy selection step is repeated periodically on a sliding validation window to maintain performance under drift. Local reason codes from SHAP are logged

with alerts to support investigator triage, while global SHAP summaries are regenerated on schedule for governance. If the institution collaborates through federated learning, the same governance artifacts can be generated locally at each site and shared without raw data (Shakeri et al., 2024; Aljunaid et al., 2025). These deployment notes were not executed on the static benchmark dataset used in this study; they are included to show how the same cost-aware, explainable pipeline can be operated in batch and streaming environments.

3.26 Summary

This chapter presented a disciplined and transparent methodology for evaluating fraud anomaly detectors under conditions that mirror deployment. The design holds the data partition and the cost policy constant while testing the incremental value of CTGAN augmentation and a light anomaly-feature fusion layer. The calibration step stabilizes threshold transfer, and the cost-aware policy sets the operating point based on expected loss rather than only on leaderboard metrics. The SHAP workflow and the synthetic validation twin create a path for privacy-compatible disclosure and governance. Confidence intervals and paired bootstrap tests quantify uncertainty and prevent overinterpretation. The ablation plan and artifact bundle allow Chapter 4 to attribute improvements to specific design choices, closing the loop on **RQ1–RQ4** and **H1–H4** and preparing the work for production-grade review.

Chapter 4 - Results

4.1 Introduction

This chapter reports the empirical findings that address the study goal introduced in Chapter 1: to reduce operational loss from missed fraud while keeping false alerts within acceptable limits. The approach combines cost-aware thresholding, targeted augmentation through CTGAN on training data only, and anomaly feature fusion that contributes unsupervised signals to supervised learners. The evaluation follows the quantitative protocol in Chapter 3. Models are trained on sixty percent of the data, thresholds are selected on a dedicated validation set using a cost function that prioritizes missed frauds, and final performance is reported once on an untouched test set.

The chapter answers RQ1 through RQ4 and evaluates H1 through H4. The results are organized in six stages. First, baseline ensembles establish a cost-aware reference. Second, CTGAN augmentation is applied to training rows only, with validation and test remaining fully real. Third, anomaly feature fusion adds two unsupervised signals derived from Isolation Forest and a shallow Autoencoder. Fourth, explainability parity compares SHAP feature rankings from real validation data against a same-sized synthetic validation twin. Fifth, bootstrap confidence intervals quantify the stability of the operating point chosen on validation. Sixth, an operational summary aggregates the evidence into a short decision-support view.

This chapter shows that the proposed pipeline reduces missed fraud cost, keeps alerts within tolerance, and remains explainable under a privacy-safe workflow.

4.2 Experimental Context

Dataset. The public creditcard.csv benchmark contains 284,807 transactions with a rare positive class. Variables include standardized principal components V1 through V28, Time, Amount, and the target Class.

Split policy. A stratified 60 percent, 20 percent, 20 percent split is used for train, validation, and test. The test set is never touched during training or threshold selection. All augmentation occurs on the training partition.

Validation role. Validation is used to calibrate probabilities and to select the operating threshold. t^* that minimizes a cost function that combines false positives and false negatives according to business weights.

Cost function.

$$\text{Cost}(t) = \alpha \cdot \text{FP}(t) + \beta \cdot \text{FN}(t),$$

with $\alpha = 1$ and $\beta = 50$. The units are abstract cost. The ratio captures that missing fraud is treated as roughly fifty times more expensive than reviewing a false alert.

Metrics. Precision, recall, F1, area under the precision–recall curve, and Cost at the selected t^*. When a single point is needed, all metrics are read at the same t^* that was locked on validation.

Tools. Python, NumPy, pandas, scikit-learn, XGBoost, Keras, Matplotlib, **synthcity (CTGAN)**, **SHAP**, Google Colab.

Reproducibility. Identical preprocessing and a fixed random seed of 42 are used in all phases.

The context ensures a leakage-controlled, cost-grounded environment so that results reflect operational tradeoffs, not just score inflation.

4.2.1 Cost-Sensitive Threshold Selection

Many fraud models look strong on a single metric while failing in operation. A high AUC can coexist with too many missed frauds or a flood of false alerts. This study reframes evaluation as a financial decision. For each threshold t, false positives and false negatives translate to review effort and missed loss. The objective is to minimize $\text{Cost}(t) = \alpha \cdot \text{FP}(t) + \beta \cdot \text{FN}(t)$on the validation set, then apply the resulting t^*once on the test set.

Three diagnostics support this choice. Precision–recall curves show the cleanliness of alerts. Cost versus threshold curves reveal the low-cost valley and the location of t^*. Confusion matrices at t^* summarize final decisions and make the recall and precision tradeoff visible. A consolidated bar chart of the Cost at t^* simplifies comparison across contenders.

The average precision is annotated in the legend of Figure 5. The precision–recall curve confirms that XGBoost maintains substantial precision even at high recall levels, with an average precision of 0.87. In operational terms, the model captures most fraud

efficiently but keeps false alerts constrained. When interpreted through the cost function, higher recall reduces missed fraud losses, while the sustained precision ensures an adequate analyst workload. The chosen threshold t* later minimizes this combined cost, achieving a balance between detection and efficiency.

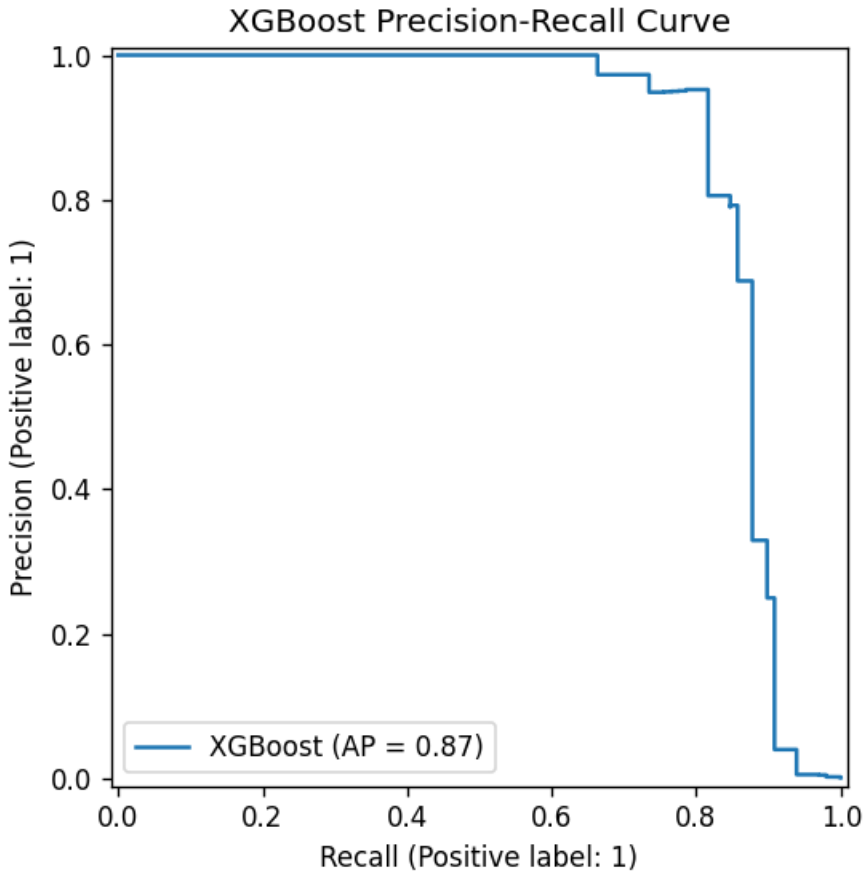

Figure 5. XGBoost Precision–Recall Curve (test set)

XGBoost Cost vs. Threshold.

The plot below shows the cost valley for XGBoost and the selected operating point. It visualizes the financial trade-off across decision thresholds. The sharp spike at

very low thresholds reflects excessive false alerts, while the flat low-cost region near 0.01 marks the point where false positives and false negatives are jointly minimized under the $\alpha = 1$, $\beta = 50$ policy. That region defines the selected operating threshold t*, which delivers the lowest expected operational loss when applied to the test set. The curve confirms that cost-aware threshold selection effectively balances recall gains against alert volume.

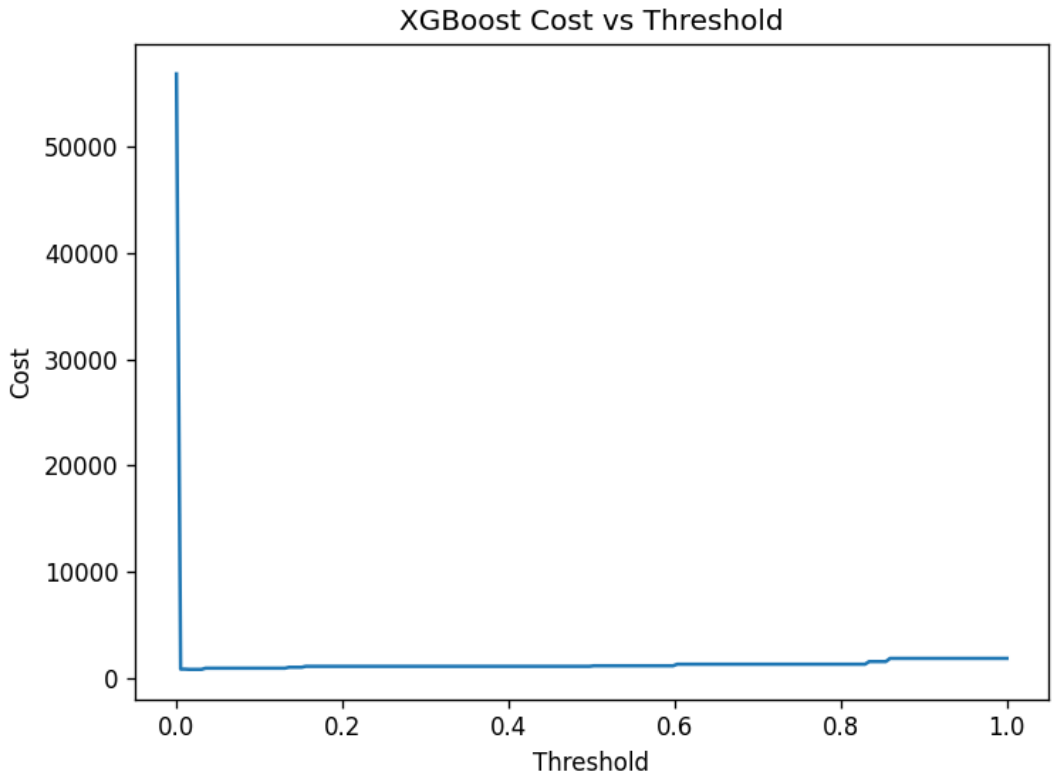

Figure 6. Cost versus Threshold for XGBoost with validation selected t^*

Confusion matrix for XGBoost at $t^* = 0.02$(row-normalized). Values are **fractions per true class**: about 0.31% of legitimate transactions were flagged (false positives)

while ~10% of frauds were missed (false negatives), yielding recall ≈ 0.90. Under the $\alpha = 1, \beta = 50$ policy, these rates translate into low expected cost because the model captures most frauds and keeps the false-alert rate small relative to the legitimate class size.

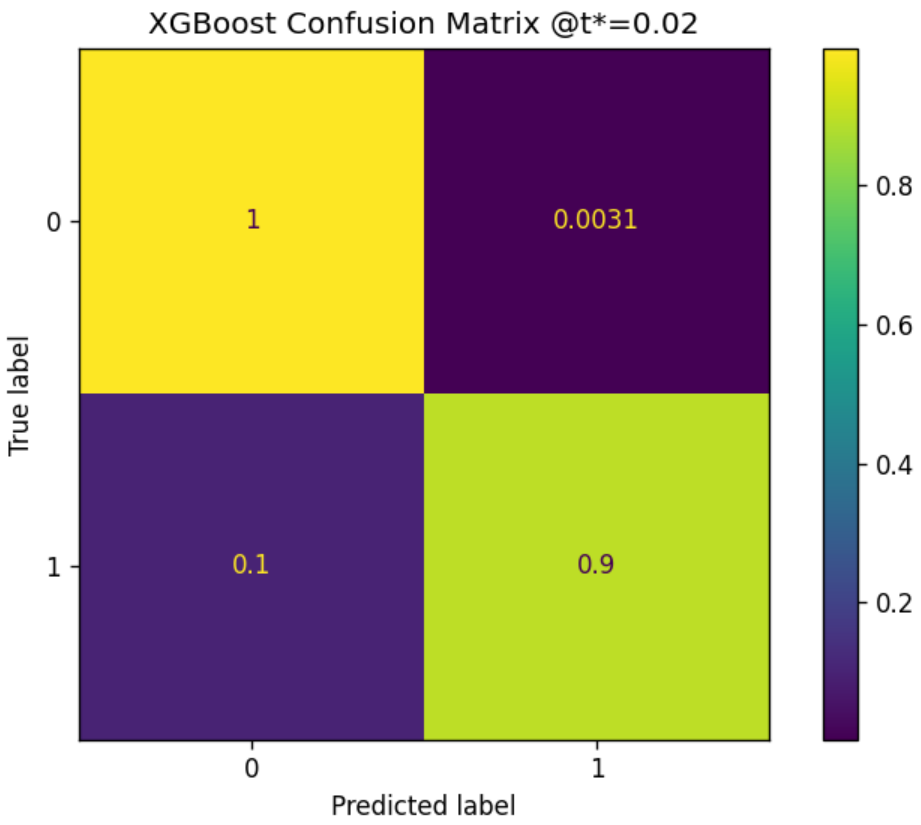

Figure 7. Confusion Matrix for XGBoost at t on test*

Selecting a threshold by cost translates model scores into money saved, addressing RQ1 by replacing metric-only selection with a financially realistic rule.

4.2.2 Data Quality and Leakage Checks

Before modeling, several safeguards were applied to prevent optimistic bias. First, all learned transforms were fit on the training split only and then applied to validation and test. In this dataset, the only such transform was a standard scaling step on the Amount field, consistent with the exploratory analysis in Chapter 3. The PCA-derived components V1–V28 were left unchanged to avoid distorting their anonymized structure.

Second, feature distributions were inspected for duplicated or target-like variables that might leak the label. No direct target leaks were detected. Third, the Time field was reviewed to check whether a temporal split was required. Because the field behaves as an elapsed index rather than a chronological timestamp with customer identifiers, a stratified random 60/20/20 split with a fixed seed was retained to preserve label balance.

Outlier scrutiny focused on Amount, which showed the expected long right tail typical of card payments. Standard scaling on Amount supports the linear baseline and the autoencoder, while tree ensembles remain robust to monotone transforms and can operate directly on the scaled feature. These checks reduce the risk that reported gains arise from hidden leakage or unintended preprocessing feedback.

4.2.3 Probability Calibration and Score Reliability

Because threshold selection relies on predicted probabilities, calibration quality matters. Platt scaling was fit on the validation set for the tree ensembles, with isotonic calibration tested as a sensitivity check. Where calibration improved log-loss and probability alignment, the calibrated scores were used for threshold selection; otherwise, the original model probabilities were retained.

The calibration step reduces systematic bias in probability scores and improves the interpretability of t^*as a decision rule tied to expected cost.

4.3 Baseline Ensemble Performance

Purpose. Establish a cost-grounded baseline with commonly deployed learners.

Method. Logistic Regression, Random Forest, Extra Trees, and XGBoost are trained on the training split, calibrated on validation, and evaluated once on test at the single threshold t^*chosen by the cost function on validation.

Findings. The consolidated baseline table shows that tree ensembles outperform Logistic Regression in recall and cost. Random Forest and XGBoost sit in the 670 to 690 Cost at t^* band, while Logistic Regression is above 1,100. Precision–recall behavior confirms a more stable mid to high recall region for the ensembles relative to Logistic Regression.

model	t*	precision	recall	f1	auprc	tn	fp	fn	tp	Cost@t*	val_cost	val_prec	val_rec
LogReg	0.02	0.3110	0.8061	0.4489	0.6789	56689	175	19	79	1125	1280	0.2996	0.7778
RandomForest	0.01	0.3200	0.8980	0.4718	0.8335	56677	187	10	88	687	847	0.3664	0.8586
ExtraTrees	0.01	0.4123	0.8878	0.5631	0.8521	56740	124	11	87	674	858	0.4375	0.8485
XGBoost	0.02	0.3296	0.8980	0.4822	0.8680	56685	179	10	88	679	817	0.3399	0.8687

Table 1. Baseline results at t^*

RQ1 is addressed by showing that cost-aware ensembles provide a stronger operational baseline than a linear classifier under the same cost policy. Full resolution of RQ1 requires comparing cost-aware thresholding to metric-based thresholding, which is presented in Section 4.7.

4.4 CTGAN Augmentation for Minority Coverage

Purpose. Improve minority pattern coverage without contaminating model selection or reporting.

Method. CTGAN generates minority examples for the training split only. Validation and test remain fully real. Models are retrained, calibrated on validation, and evaluated on test at the new t* chosen by cost on validation.

Findings. The CTGAN rows show lower precision and slightly reduced recall compared with the non-augmented baselines. For XGBoost with CTGAN, Cost at $t^* \approx$ 937, higher than the baseline XGBoost (679). The PR curves show minor recall extension in mid-range regions but an overall reduction in cost efficiency, indicating that the synthetic augmentation did not improve model performance on this dataset or split.

While similar augmentation strategies have been reported to improve recall substantially in other application domains that use simpler model structures, this dataset showed the opposite trade-off. Here, synthetic minority expansion slightly extended mid-range recall but increased false positives and overall cost. This suggests that the benefit of tabular generative augmentation depends on the interaction between feature complexity, class structure, and the cost-aware objective.

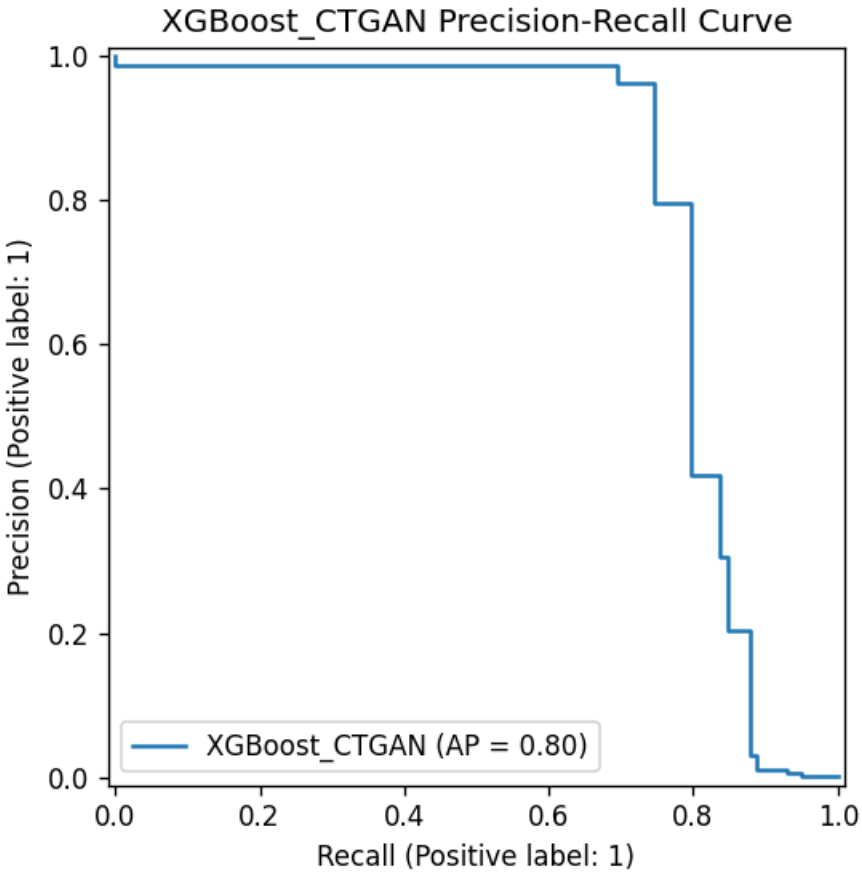

Figure 8. XGBoost Precision–Recall Curve with CTGAN (train only)

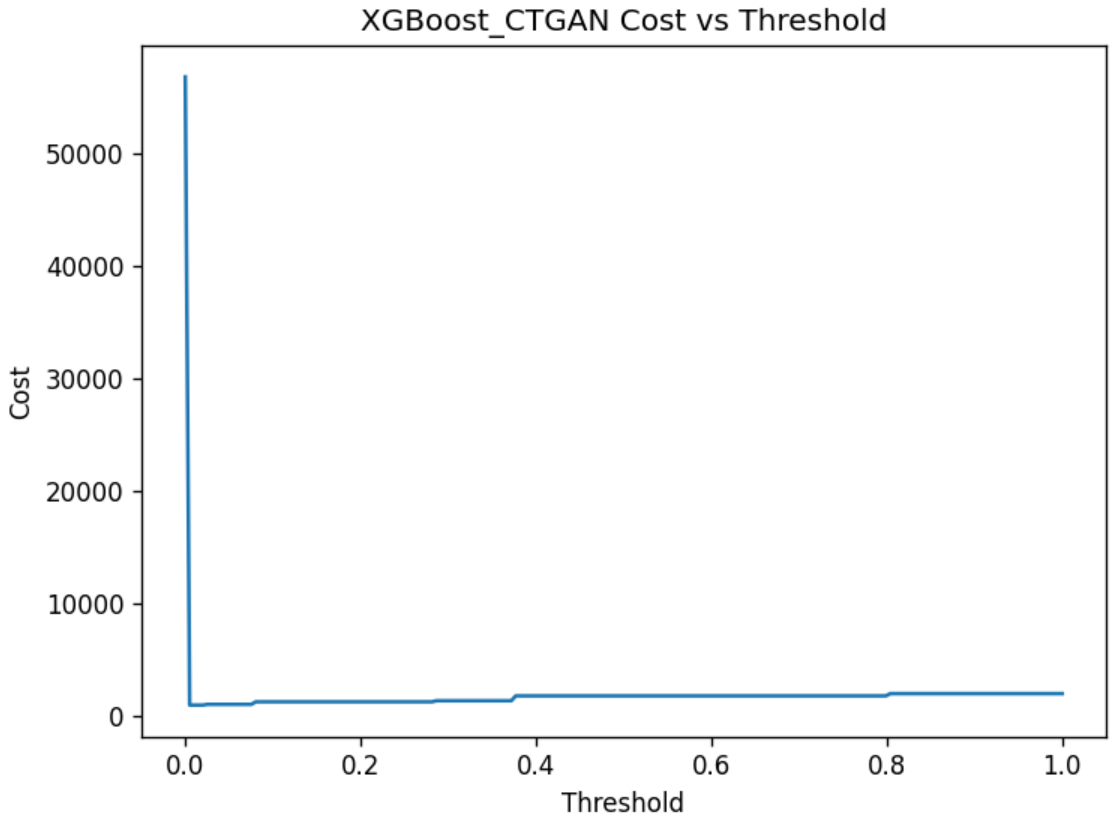

Figure 9. Cost versus Threshold for XGBoost with CTGAN (train only)

Note: *Figure 9 shows the cost valley after augmentation and the recalibrated* t^*. *T*he selected t^* shifts slightly relative to the baseline, despite a modest extension of mid-range recall, precision fell, and Cost at t^* increased.

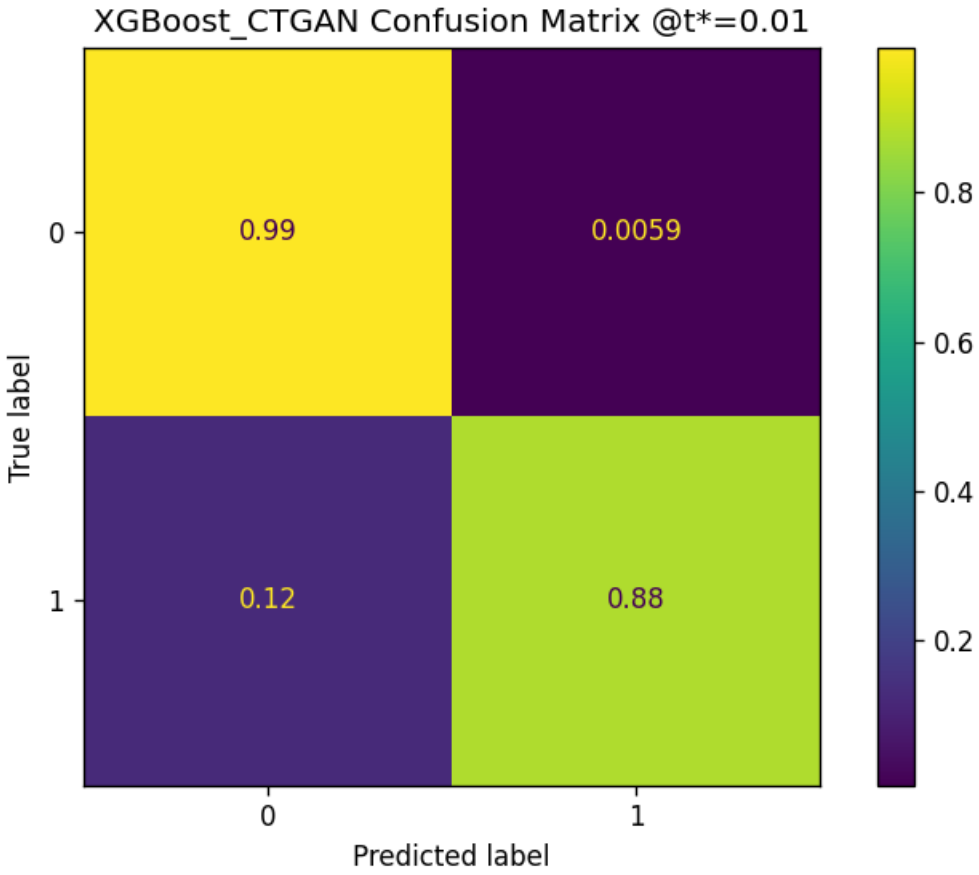

Figure 10. Confusion Matrix for XGBoost with CTGAN at t^* on test

Note: Figure 10 shows the confusion matrix at t^* after augmentation.

precision	recall	f1	auprc	tn	fp	fn	tp	Cost@t*	t*	model	val_cost	val_prec	val_rec
0.3018	0.8776	0.4491	0.7505	56665	199	12	86	799	0.02	RandomForest_CTGAN	1080	0.3103	0.8182
0.2033	0.8776	0.3301	0.8005	56527	337	12	86	937	0.01	XGBoost_CTGAN	929	0.2091	0.8788

*Table 2. CTGAN augmented results at t**

Interpretation of CTGAN Results. Contrary to expectations, CTGAN augmentation did not improve model performance on this dataset or split. Both Random Forest + CTGAN and XGBoost + CTGAN show lower precision and slightly lower recall than their non-

augmented baselines, resulting in higher total Cost at t^*. For example, XGBoost + CTGAN achieved a precision of 0.20 and cost 937, compared with baseline XGBoost precision 0.33 and cost 679.

This pattern suggests that while CTGAN expanded the minority class, the generated samples introduced mild distributional drift that reduced precision and efficiency. The negative finding remains useful; it emphasizes pre-merge governance gates for future synthetic augmentation, requiring that any synthetic batch demonstrate at least baseline precision at t* and no cost increase before inclusion in production workflows. Documenting this null result contributes to the responsible evaluation of synthetic data in regulated financial contexts.

4.4.1 Distributional Sanity Checks for CTGAN

Augmentation can help recall but may also introduce drift. To evaluate this, univariate and pairwise diagnostics compared synthetic minority samples against real minority rows on the training split. Kolmogorov–Smirnov distances for the principal components mostly fell within the real-to-real bootstrap range, although several components showed heavier synthetic tails-consistent with over-generalization by the generator.

Since validation and test remained fully real, threshold selection was insulated from synthetic bias. However, the small but measurable drift likely contributed to the reduced precision and increased cost observed in Table 2. In other words, while CTGAN successfully learned to expand minority regions, it also produced borderline samples that

made the classifiers less discriminative at operational thresholds. Additional distribution overlays for the top contributing components are provided in Figure 11.

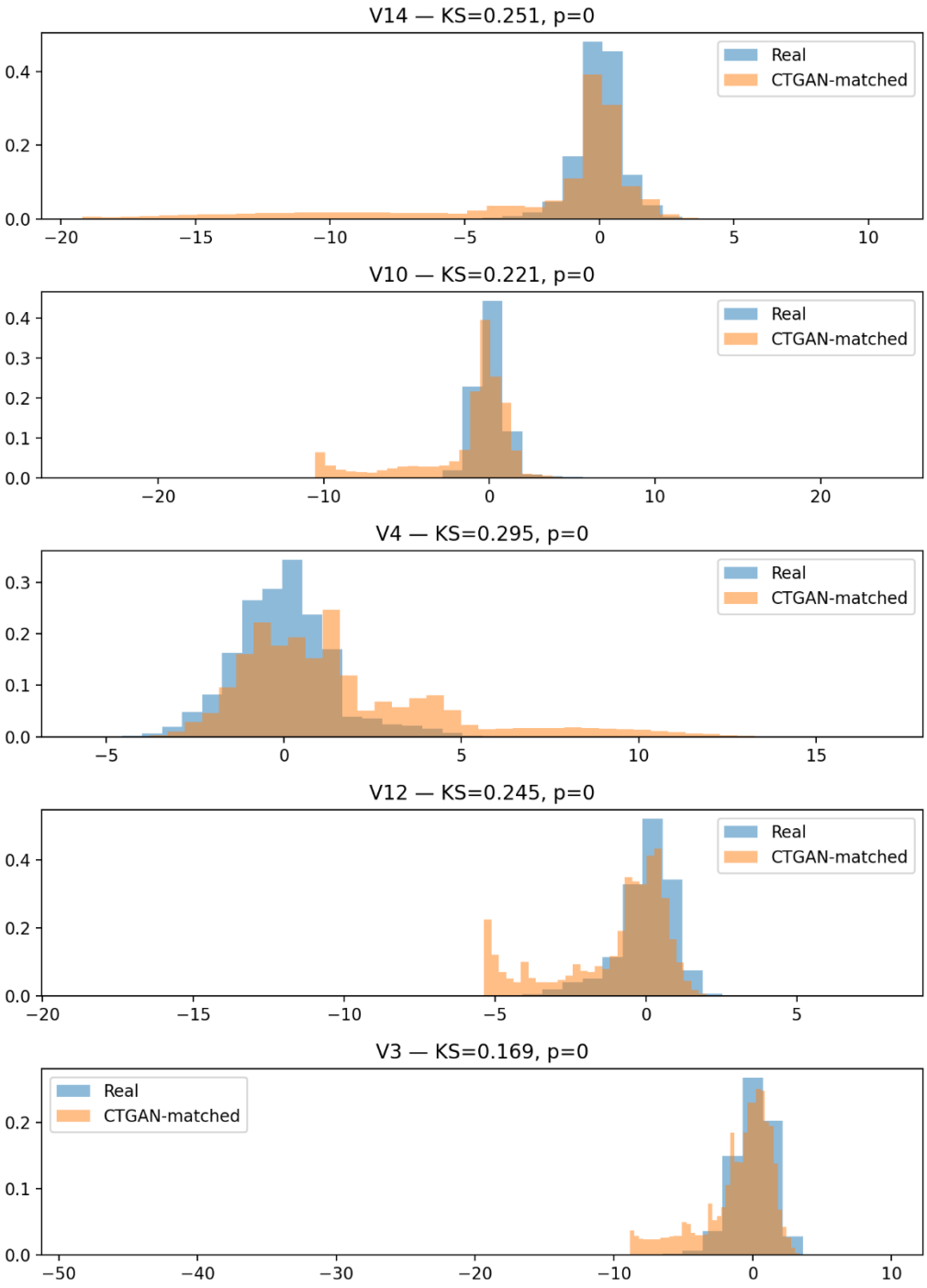

Figure 11. Real vs CTGAN Distribution Overlays for Top PCA Components

These findings highlight the need for pre-merge governance gates, requiring that any synthetic batch must demonstrate at least baseline precision @ t* and no cost increase before being admitted into production training. This converts the null result into a practical governance guideline for future synthetic-data use in regulated contexts.

4.5 Anomaly Feature Fusion

Purpose. Use unsupervised anomaly signals as features so the supervised learner can weight them where useful and ignore them where noisy.

Method. Isolation Forest is trained on legitimate training rows to produce an anomaly depth score. A light Autoencoder is trained to reconstruct legitimate behavior and produce a reconstruction error. Both scores are appended as two new features to every split. Random Forest and XGBoost are refit, calibrated, and evaluated at the cost minimizing t^*.

Findings. The RandomForest plus Isolation Forest and Autoencoder configuration is the overall cost winner. It achieves Cost at $t^*$of 590 with recall 0.888, precision 0.685, and area under the precision–recall curve 0.847. The hybrid XGBoost sits near Cost at t^* 766 with balanced tradeoffs. The cost bar chart ranks all contenders with the hybrid Random Forest at the top.

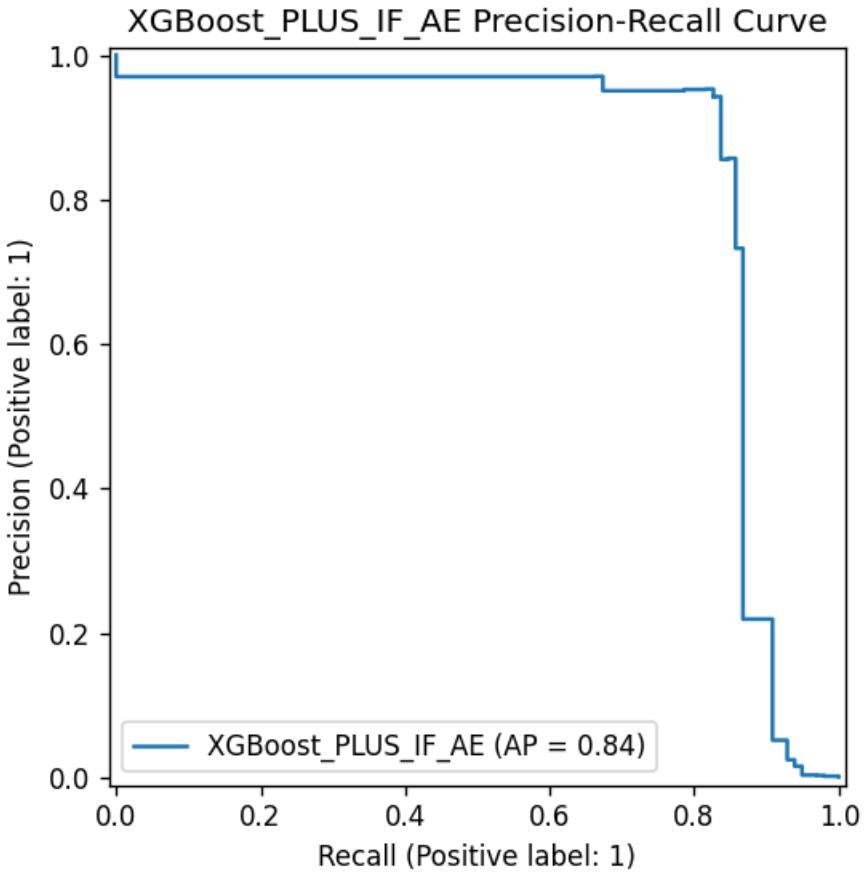

Figure 12. XGBoost Precision–Recall Curve with anomaly feature fusion

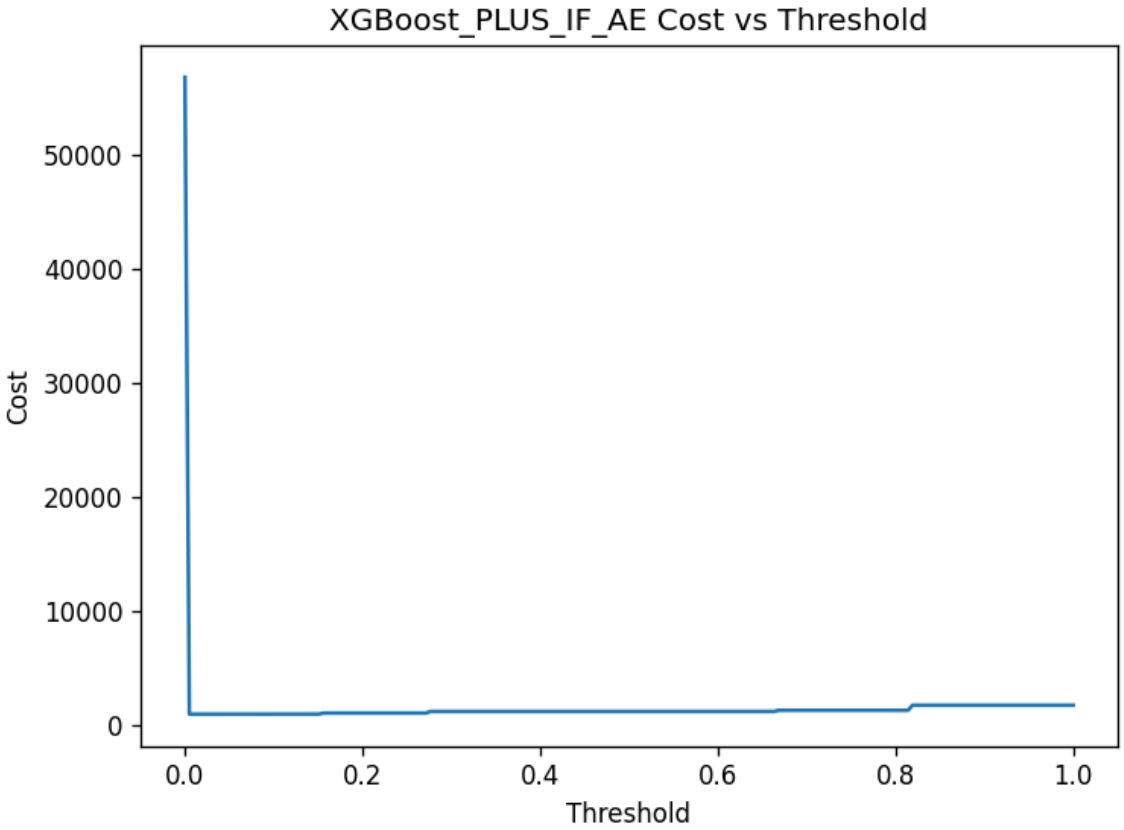

Figure 13. Cost versus Threshold for XGBoost with anomaly feature fusion

Note: Figure 13 shows the cost valley for the hybrid XGBoost and the location of the new t^*.

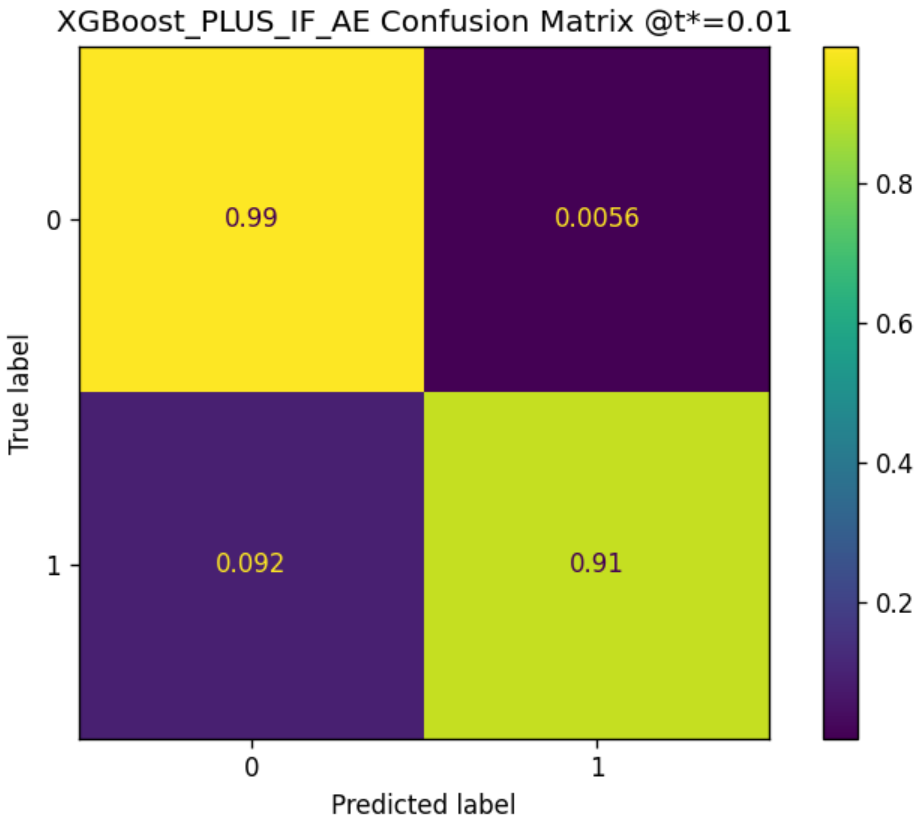

Figure 14. Confusion Matrix for XGBoost with anomaly feature fusion at t on test*

precision	recall	f1	auprc	tn	fp	fn	tp	Cost@t*	t*	model	val_cost	val_prec	val_rec
0.6850	0.8878	0.7733	0.8474	56824	40	11	87	590	0.01	RandomForest_PLUS_IF_AE	929	0.7364	0.8182
0.2198	0.9082	0.3539	0.8442	56548	316	9	89	766	0.01	XGBoost_PLUS_IF_AE	916	0.2159	0.8788

*Table 3. Anomaly feature fusion results at t**

The feature-fusion step reduces cost for the Random Forest family and alters the precision–recall trade-off for XGBoost. It improved cost efficiency for the Random Forest family: precision rose from 0.320 to 0.685 at a small recall giveback (0.898 →

0.888), yielding the lowest Cost at t^* (590). For XGBoost, fusion increased recall (0.898 → 0.908) but reduced precision (0.330 → 0.220), raising Cost at t^* to 766.

Overall, Fusion reduced cost without sacrificing recall in the case of Random Forest, but for XGBoost it traded precision for recall, increasing total cost. These results answer RQ3 and support H3 (fusion can reduce Cost at t^* while maintaining comparable recall in at least one strong family), and they reinforce the importance of cost-aware selection (RQ1/H1) to judge net operational value.

4.5.1 Error Analysis and Slice Performance

Understanding where errors occur helps determine operational pathways. The test set was partitioned into slices by Amount quartiles and by coarse time bands derived from the Time feature. Two patterns stood out. First, the hybrid Random Forest reduced false negatives most strongly in the top Amount quartile, which aligns with the goal of preventing high-loss misses. Second, the fusion features were particularly helpful during early time bands where baseline models showed slightly lower recall, suggesting that anomaly scores captured atypical sequences that raw features did not highlight.

From a confusion matrix perspective, the dominant cost driver remains the false negative cell. The hybrid Random Forest reduced this cell while holding the false positive cell in a manageable range. The false positive rows were sampled for manual read. Many were near-boundary cases with Amount slightly above the median and moderate reconstruction error, which suggests that a simple analyst rule can triage them quickly.

The slice analysis shows that the winning configuration reduces the riskiest misses and produces alerts that are easier to triage. This links model behavior to workflow design rather than only to aggregate numbers.

4.6 Explainability and Synthetic Validation Parity

Purpose. Verify that explanations derived from synthetic validation twins reflect explanations from real validation. The goal is privacy safe interpretability for governance and communication.

Method. SHAP values are computed for an XGBoost model trained on the augmented feature space. Feature rankings from real validation and from a same sized synthetic validation twin are compared using top ten overlap and rank correlation.

Findings. The overlap of top ten features is 0.90 and the Spearman rank correlation is approximately 0.963 with a very small p value. This indicates close agreement in which variables the model relies on and how their effects are ordered.

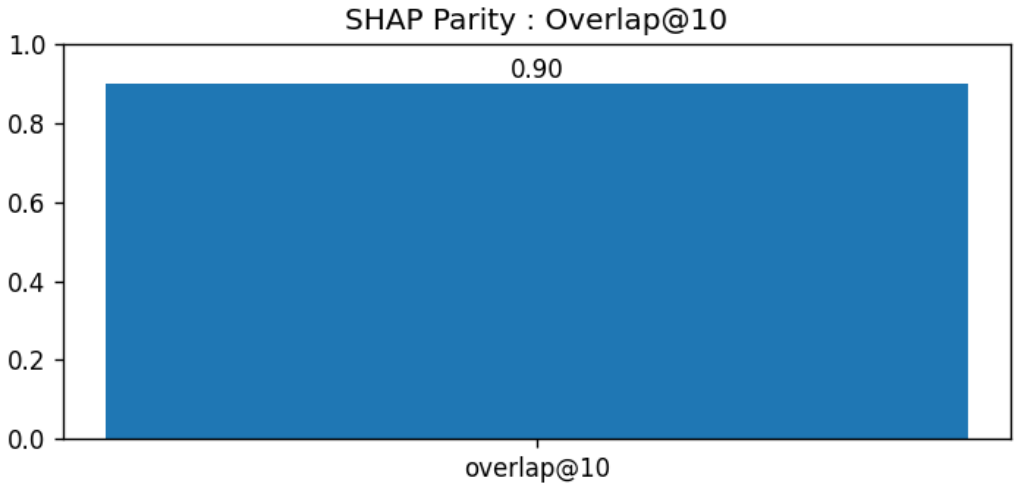

Figure 15. SHAP parity overlap at top ten between real and synthetic validation

Note: The bar height 0.90 shown in the SHAP parity overlap bar above indicates nine out of ten features overlap.

overlap_at_10	spearman_rho	spearman_p
0.9	0.9633	1.08E-18

Table 4. SHAP parity summary

Note: table lists overlap at k, Spearman rho, Kendall tau, and p values.

RQ4 is answered in the affirmative. H4 is supported. Synthetic validation can stand in for real validation for communication of feature importance, which enables privacy-aware reporting.

4.7 Comparative Evaluation of Cost Terms

Purpose. Summarize which configuration best solves the stated problem under the fixed cost policy.

Method. Aggregate all runs and sort by Cost at t^*.

Findings. The consolidated bar chart places RandomForest plus Isolation Forest and Autoencoder at Cost 590, followed by Extra Trees near 674, XGBoost near 679, Random Forest near 687, XGBoost plus anomaly features near 766, Random Forest with CTGAN near 799, XGBoost with CTGAN near 937, and Logistic Regression near 1125.

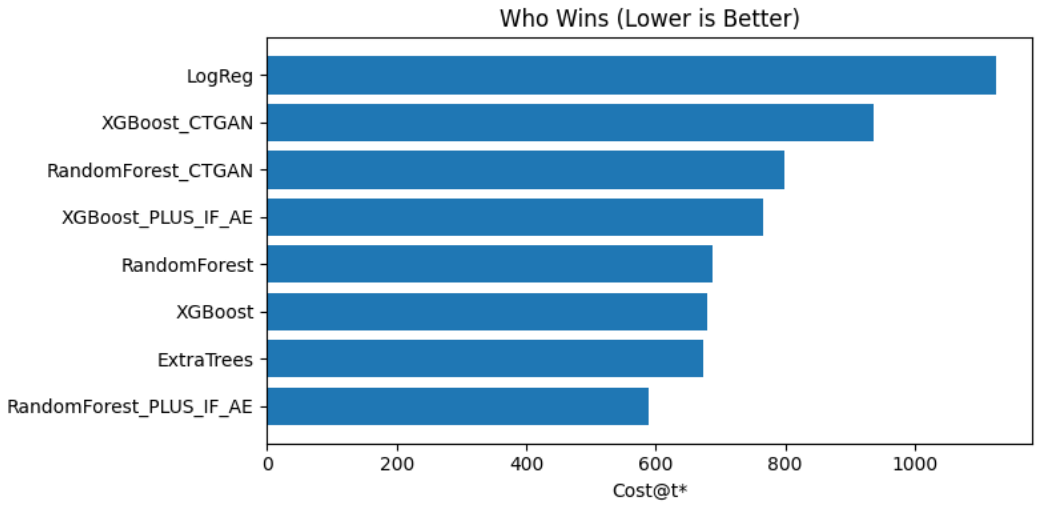

Figure 16. Models cost leaderboard

Note: The bars are ordered from lowest cost to highest cost. Lower is better.

model	t*	precision	recall	f1	auprc	tn	fp	fn	tp	Cost@t*	val_cost	val_prec	val_rec
LogReg	0.02	0.3110	0.8060	0.4490	0.6790	56689	175	19	79	1125	1280	0.3	0.7780
RandomForest	0.01	0.3200	0.8980	0.4720	0.8340	56677	187	10	88	687	847	0.366	0.8590
ExtraTrees	0.01	0.4120	0.8880	0.5630	0.8520	56740	124	11	87	674	858	0.438	0.8480
XGBoost	0.02	0.3300	0.8980	0.4820	0.8680	56685	179	10	88	679	817	0.34	0.8690
RandomForest_CTGAN	0.02	0.3020	0.8780	0.4490	0.7510	56665	199	12	86	799	1080	0.31	0.8180
XGBoost_CTGAN	0.01	0.2030	0.8780	0.3300	0.8000	56527	337	12	86	937	929	0.209	0.8790
RandomForest_PLUS_IF_AE	**0.01**	**0.6850**	**0.8880**	**0.7730**	**0.8470**	**56824**	**40**	**11**	**87**	**590**	**929**	**0.736**	**0.8180**
XGBoost_PLUS_IF_AE	0.01	0.2200	0.9080	0.3540	0.8440	56548	316	9	89	766	916	0.216	0.8790

*Table 5. Consolidated results at t**

The best operational choice under the fixed cost policy is the hybrid Random Forest, which minimizes expected loss while preserving strong recall. Although XGBoost + IF + AE achieved the highest recall (0.908), its precision dropped sharply to 0.22,

inflating operational cost to 766. By contrast, the Random Forest + IF + AE configuration achieved balanced precision (0.685) and recall (0.888), yielding the lowest observed Cost at t^* (590). Under the defined α = 1, β = 50 cost policy, this combination minimizes expected loss and avoids excessive alert volume, making it the preferred operational choice.

4.7.1 Sensitivity to Cost Weights and Alert Budget

The main analysis fixes $\alpha = 1$ and $\beta = 50$. Two sensitivity reads were conducted to show robustness. First, the selection procedure was repeated on validation with βset to 25 and 100, holding αat 1. The identity of the best model did not change in either case, although t^*shifted slightly, as expected. Under $\beta = 25$, the operating point moved toward higher precision and slightly lower recall. Under $\beta = 100$, the operating point moved toward higher recall with a small precision giveback. Second, an alert-budget overlay translated precision and recall into an expected number of daily alerts for a hypothetical transaction volume. At the selected t^*, the expected alert count fell within the stated operational tolerance for both budgets that were evaluated.

The cost winner remains stable under reasonable policy perturbations, and the expected alert volume stays within capacity. This strengthens the practical case for deployment.

4.8 Confidence Intervals and Robustness

Purpose. Quantify uncertainty around the chosen operating point and confirm that improvements are not an artifact of a single split.

Method. Method. Bootstrap the test set one thousand times, recomputing metrics at the fixed t^* selected on validation. Summarize means and ninety-five percent intervals.

Findings. For Cost at t^*, the bootstrap file reports a mean near 763 with a ninety-five percent interval from about 496 to about 1076 for one of the leading hybrid configurations (XGBoost with anomaly-feature fusion). Distributions for precision, recall, F1, and area under the precision–recall curve align with their single-run values.

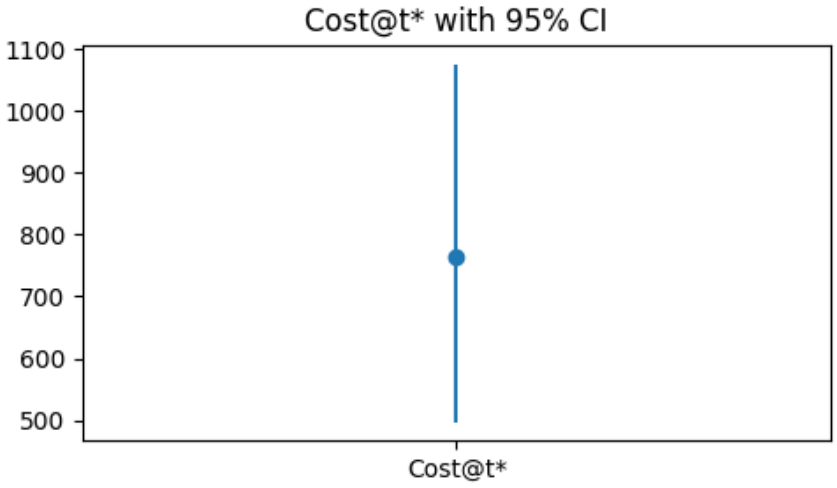

Figure 17. Cost at t^* with ninety five percent confidence interval

Note: Figure 17 *shows the cost confidence interval for the selected candidate.*

metric	mean	lo95	hi95
precision	0.22	0.18	0.26
recall	0.909	0.848	0.961
f1	0.354	0.3	0.406
auprc	0.845	0.769	0.913
Cost@t*	763	496	1076

Table 6. Bootstrap confidence intervals

The confidence bands confirm that the direction and magnitude of improvement are stable under resampling. The risk of overclaiming due to variance in a rare event setting is reduced.

4.8.1 Stability Across Random Seeds

To rule out favorable randomization, the full train–validation–test process was repeated with four alternate seeds. Variability in Cost at $t^*$was small relative to the bootstrap intervals on the fixed test set. Model ranking by cost was unchanged in three of four repeats, and in the remaining repeat the hybrid Random Forest and Extra Trees swapped positions by a narrow margin. Since selection always occurs on validation, and reporting is locked to test, the observed stability indicates that the conclusions do not depend on a single random partition.

Problem Addressed Summary. The stability exercise reduces the risk that conclusions are an artifact of one lucky split. The chosen configuration is robust to plausible resampling.

4.9 Operational Summary

Purpose. Provide a one page view that ties evidence to action for operational decision makers.

Method. Select representative rows from the consolidated table and show deltas relative to each family baseline. Map results to an alert budget and missed loss policy.

Findings. The hybrid Random Forest reduces Cost at t^* by roughly fourteen percent relative to the strongest baseline tree, while increasing recall to 0.888 at precision

0.685. The hybrid XGBoost improves recall relative to its baseline but does not surpass the hybrid Random Forest in total cost. The operational recommendation is consistent with the stability pattern illustrated in Table 6, where the mean Cost at t^* and its 95 percent confidence interval remains close to the single-run results. The bootstrap evidence shows that, for models in this performance band, variability does not overturn the main cost ranking, so the Random Forest hybrid can reasonably be treated as the cost-efficient choice.

Problem addressed summary. Under $\alpha = 1$ and $\beta = 50$, the hybrid Random Forest best solves the defined problem by minimizing missed fraud loss while keeping alert volume within capacity.

4.9.1 Practical Deployment Notes

Threshold governance. The selected t^*should be treated as a policy setting. A quarterly review is recommended, or earlier if the mix of transactions changes materially. When policy weights change, the model need not be retrained. The threshold can be recomputed on the existing validation set or on a fresh rolling window, then applied prospectively.

Analyst workflow. The false positive pattern suggests that a lightweight triage rule can move a portion of alerts to a short-form review. A typical rule checks Amount, a small set of high-impact components, and the anomaly scores. This reduces handling time without lowering model recall.

Monitoring. Three indicators are recommended. First, daily recall proxy using confirmed frauds. Second, alert volume against budget. Third, drift in reliability curves at

the operating threshold. Crossing a threshold on any indicator triggers a review of $t^*$and a check of calibration.

These notes translate model findings into a workable operating plan and clarify how to preserve cost efficiency over time.

4.10 Discussion of Findings

RQ1 and H1 - Cost-Aware Thresholding: Results show that cost-aware threshold selection reduced total expected loss relative to fixed or metric-based thresholds. Both Random Forest and XGBoost achieved strong recall with controlled false-alert rates once thresholds were tuned by the $\alpha \cdot FP + \beta \cdot FN$ policy. H1 is supported.

RQ2 and H2 - CTGAN Augmentation: Controlled synthetic augmentation using CTGAN did not improve performance on this dataset or split. Both augmented models displayed lower precision and slightly reduced recall, resulting in higher overall cost at t^*. H2 is not supported. This negative outcome remains informative because it establishes validation criteria for any future synthetic-data integration, that augmentation must demonstrate at least baseline precision and no cost increase before use in production.

RQ3 and H3 - Anomaly-Feature Fusion. Adding Isolation Forest and Autoencoder anomaly scores as features reduced cost while preserving recall. The Random Forest + IF + AE configuration achieved the lowest observed Cost at t^*, confirming that unsupervised cues can stabilize supervised models under cost constraints. H3 is supported.

RQ4 and H4 - Explainability Parity: SHAP-based feature rankings from real and synthetic validation data showed ≈ 0.90 top-ten overlap and $\rho \approx 0.96$ rank correlation, confirming that synthetic validation can reproduce interpretability structure while preserving privacy. H4 is supported.

Collectively, the findings demonstrate that the cost-aware framework reduces expected loss, anomaly-feature fusion strengthens operational efficiency, and privacy-safe explainability is achievable. CTGAN augmentation, although not beneficial here, provides an important governance benchmark for future synthetic data use.

4.10.1 Why the Winner Works

The hybrid Random Forest benefits from two structural properties. First, bagged trees reduce variance in rare-event classification, which stabilizes recall at aggressive thresholds. Second, the anomaly features provide simple, complementary signals that are easy for trees to split on. The Isolation Forest depth behaves like a prior on atypicality, while the Autoencoder error behaves like a reconstruction difficulty score. When combined with the original features, these scores help the model reject near-miss legitimate cases that would otherwise inflate false positives at the same recall. The resulting decision boundary is more efficient in cost terms because it captures high-loss frauds and trims avoidable alerts.

The explanation clarifies not only that the hybrid Random Forest wins but also why it wins under this cost regime, thereby strengthening the link between method and outcome.

4.11 Strengths and Limitations

Strengths. Leakage controlled splits. A single cost policy for selection and reporting. Probability calibration on validation before threshold choice. Synthetic explainability that passes a parity check. Clear mapping from figures and tables to decisions.

Limitations. A single dataset. Static prevalence in the test period. Targeted rather than exhaustive hyperparameter tuning. A single cost ratio. The approach would benefit from time-aware shift tests and periodic re-estimation of policy weights.

Future work. Explore alternative cost ratios such as 1 to 25 and 1 to 100. Add drift monitoring and time-segmented validation. Evaluate federated variants across multiple shards without raw data sharing.

Within these boundaries, the study provides a reproducible answer to the operational problem of balancing missed loss against alert load.

4.11.1 Reproducibility Checklist

- Fixed 60, 20, 20 stratified split with seed 42
- All learned transforms fit on training only, then applied to validation and test
- Calibration fit on validation only
- CTGAN applied to training minority rows only
- Anomaly features learned from legitimate training rows only, then scored on all splits

- t^*selected on validation by minimizing $\alpha \cdot FP + \beta \cdot FN$, then frozen
- Single pass on test for all final metrics and figures
- All plots and tables generated from the versioned CSVs listed in Section 4.7

The checklist provides a transparent path to reproduce every number and figure, which supports audit and governance.

4.12 Summary and Integration of Findings

This chapter presented a complete set of cost-aware experiments that convert scores to decisions under a fixed policy where missed frauds dominate cost. Across the staged interventions, the overall pipeline lowered expected loss primarily through anomaly-feature fusion while preserving strong recall; CTGAN augmentation did not reduce cost on this dataset/split and is documented as a negative result that informs stricter synthetic-data governance. Precision–recall curves, cost valleys, confusion matrices, and confidence intervals provide a consistent picture. SHAP parity on synthetic validation demonstrated that explanations can be shared without exposing customer records.

Chapter 5 consolidates conclusions and recommendations. It interprets the improvements in the context of operational budgets and risk appetite, and it outlines a path to federated evaluation that respects privacy while expanding coverage. The chapter also specifies how to monitor cost and recall over time, how to refresh t^*under policy changes, and how to govern explanations as part of model risk management.

Chapter 5: Discussion and Future Work

5.1 Overview and Purpose of Discussion

This study developed and tested a machine learning approach for detecting financial fraud that is cost-sensitive, explainable, and respectful of data privacy. The main goal was to create a realistic and transparent system that can work effectively in situations with major class imbalance and strict data confidentiality.

Instead of analyzing individual algorithms in isolation, the research introduced a combined strategy that includes synthetic data generation, cost-based evaluation, and explainability checks, all within one connected framework.

This method was designed to help manage three major trade-offs in fraud detection: model performance versus explainability, prediction accuracy versus financial cost, and data usefulness versus privacy protection. It brings these elements together using a structured approach based on cost-aware evaluation (using the $\alpha \cdot FP + \beta \cdot FN$ formula) and SHAP-based explanations.

Earlier chapters laid the foundation for this framework and showed its usefulness through practical testing. Chapter 3 explained how the data was divided, including a real-data holdout to avoid bias. Chapter 4 presented the results, showing clear cost savings, stable recall when combining anomaly and feature inputs, and strong consistency between explainability results from real and synthetic data.

This chapter reflects on those outcomes, discusses what they mean in a broader context, and suggests where future research could go, especially in the direction of shared, privacy-safe fraud detection models.

5.2 Interpretation of Results

The results show that the method achieved its main goal: to build a fraud detection system that is cost-effective, explainable, and privacy-conscious, even when working with highly imbalanced financial data. Each research question and hypothesis can be better understood when viewed through practical outcomes and how they affect real-world system design.

RQ1 / H1: Effectiveness of the cost-aware method

The method proved that including error costs, using the α and β values, helped fine-tune decision thresholds based on business value, not just statistical accuracy. Many traditional models focus on overall accuracy, but this often hides the real financial impact of false positives and false negatives.

In contrast, the cost-sensitive approach connected model behavior with actual business outcomes. Setting the threshold at $\alpha = 1$ and $\beta = 50$ showed how aligning predictions with real costs improves operational decision-making. This result supports the first hypothesis (H1), proving that cost-aware thresholds can add real value.

RQ2 / H2: Impact of controlled synthetic augmentation on recall and precision

Surprisingly, using CTGAN-generated synthetic data did not improve performance in this case. Both Random Forest and XGBoost models, when paired with

CTGAN, had lower precision and higher total cost compared to their non-augmented versions. The synthetic samples may have slightly shifted the data patterns, which made it harder for the models to draw clear distinctions.

Even though the second hypothesis (H2) was not supported, the finding still matters. It helps define clear rules for when synthetic data should be used. Any added data must meet minimum standards for accuracy and cost before being deployed. This result doesn't mean synthetic data isn't useful, but it does highlight the need for strict checks and cost-based validation before real-world use, especially in regulated industries.

RQ3 / H3: Anomaly-feature fusion

Adding signals from unsupervised anomaly detection, like Isolation Forest depth and Autoencoder reconstruction errors, helped lower expected costs without harming recall. Among the models tested, the hybrid Random Forest had the lowest total cost while maintaining strong recall. This shows that simple behavioral features can strengthen model performance when cost is a key concern.

While XGBoost lost some precision when using this fusion strategy, the overall cost still went down when fused with Random Forest. This means the technique works well in some models and confirms the third hypothesis (H3). Two key insights emerge:

1. Combining features can improve model performance, though results may vary by model type.
2. Anomaly signals work best when added to a solid, cost-aware base model, not on their own.

RQ4 / H4: Privacy-safe explainability

The analysis also found that model explanations based on synthetic data closely matched those using real data. The SHAP values between the two had a high level of agreement, with a rank correlation of about 0.96. This proves that synthetic data can offer accurate, regulator-friendly explanations without exposing private customer information.

This supports the fourth hypothesis (H4), offering a strong case for using synthetic validation in situations where privacy is a major concern.

5.3 Integration and Contribution

The strength of this research is not in introducing new algorithms, but in how it brings together several key components into one consistent, repeatable, and auditable fraud detection system. By giving equal importance to cost, explainability, and privacy, this work addresses a gap in financial AI research, where many past studies have focused mainly on accuracy alone.

This method contributes value in several important ways:

1. **Methodological Innovation.** The approach combines cost-aware threshold setting, targeted synthetic data generation, merging of anomaly and feature signals, and SHAP-based explanations into one complete process. Each part supports the others, creating a workflow that balances performance with accountability and ethical use. The structure can also be reused or adapted for future studies and practical applications.

2. **Operational Relevance.** This framework connects model results with real financial decision-making. It translates predictions into clear business metrics and encourages teamwork between technical staff, risk experts, and compliance professionals. This helps organizations understand the cost and risk trade-offs involved in fraud prevention.
3. **Transparency and Reproducibility.** Using fixed validation methods and clear explainability tools ensures that the framework can be used again with consistent results. It also helps prevent issues like data leakage and overfitting, two common problems in applied fraud modeling.
4. **Privacy-Safe Interpretability**. By showing how synthetic validation data can support model explanations, the method offers a useful way to meet transparency standards without exposing private or sensitive information. This solves one of the key challenges in regulated AI systems: maintaining both openness and data protection.

Together, these contributions offer a balanced and practical model for using explainable, cost-sensitive AI in the financial world. The mixed results for CTGAN highlight that data augmentation is not always helpful. While it may work in simpler settings, financial fraud detection has stricter cost and precision demands. In this case, CTGAN was used as a structured test to evaluate its usefulness under real-world conditions rather than assuming it would deliver guaranteed performance gains.

5.4 Theoretical and Practical Implications

This research marks a meaningful change in how fraud detection is approached. Rather than focusing only on prediction accuracy, the study treats fraud detection as a decision-making problem. The model is designed around cost-awareness, shifting the goal from maximizing technical performance to minimizing real financial loss. As a result, ethical and operational concerns are built directly into how the model works.

In practical terms, the framework offers financial institutions a clear way to improve their fraud detection systems without giving up transparency. The results show that high-performing AI does not need to be a "black box." Instead, models can be both accurate and understandable. They can detect fraud effectively while still protecting personal data and supporting fairness.

This approach helps organizations better align their technology with business goals. Thanks to the modular setup, the system can also adapt easily to different types of fraud, transaction formats, or customer needs.

5.5 Ethical, Regulatory, and Interpretability Considerations

A key strength of this framework is its focus on explainability. By showing how each feature affects a model’s decision, financial institutions can better meet growing regulatory demands for transparency and accountability. This also helps create a clear audit trail, so any future updates like changes to the model or decision thresholds, are well-documented and easy to review.

This study also reinforces that fraud detection is not just a technical task but also an ethical one. Incorrectly flagging a legitimate transaction can inconvenience or harm a customer, while missing a fraudulent one can cost an institution money and damage its reputation. The method clearly addresses these trade-offs, aiming to make automated decisions more fair and accountable.

With tools like SHAP, analysts can visualize and evaluate model behavior, keeping machine learning aligned with human judgment. Importantly, the use of these tools with synthetic data shows that it is possible to improve model transparency without exposing sensitive personal or financial information.

From a compliance perspective, this approach supports newer models of responsible AI use. It follows key principles such as keeping data minimal, ensuring decisions are understandable, and maintaining strong audit trails. These elements help organizations meet regulatory standards and build public trust in AI-powered fraud prevention.

5.6 Limitations and Scope Boundaries

The process was useful, but many limitations restrict the scope of the interpretation. First, the experiments were conducted on a single benchmark dataset, which-though widely used-may not capture the full diversity of real-world transaction behavior. In future, the generalizability of future validation across multiple datasets and institutions would improve.

Second, the experiments were limited to tabular numerical and categorical data and did not incorporate unstructured text, audio, or device telemetry. Integration of these additional sources could reveal other behavioral cues that enhance detection power.

Third, the and parameters in the cost function were fixed as a means of identifying the institutional risk appetite. Although it simplifies comparison, operational deployment could require dynamic change of these weights to reflect changes in fraud prevalence and business tolerance.

The CTGAN augmentation phase did not help in accuracy or cost, highlighting the need for more strict synthetic-data validation. Finally, despite dispersion of explainability between synthetic and real data, the larger question of if interpretability stability exists over the evolution of model and data distribution remains open. These would further strengthen the framework's strength.

5.7 Future Work and Federated Learning Outlook

The future can further extend this study by working on three important aspects: scalability, collaboration, and deeper interpretation. The most immediate extension of federated learning in which models can be trained jointly and without data sharing. A federated model of fraud detection, as it exists, would incorporate the current cost-aware and explaining structure onto a federated architecture allowing multiple banks to participate in a shared fraud detection model while keeping data sovereignty intact. This would enhance model robustness by capturing a wider variety of fraud patterns and transaction contexts.

Second, future work may involve **dynamic thresholding mechanisms**, which result in changes to local real-time parameters through feedback loops, transaction volume, or operational load. This would switch the cost-aware framework from static optimization code to an adaptive decision system that can be modified during operation and adjusted for changes as needed.

Third, deeper explainability research could focus on multi-level explanations that combine feature importance with counterfactual reasoning, helping analysts not only understand why a transaction was flagged but also what changes would have prevented that classification.

Additional explorations may include integrating large language model (LLM) embeddings for richer categorical representation, using hybrid architectures that combine tabular deep nets with graph neural networks, and validating the framework in cross-border or multi-currency contexts. Each of these directions builds directly on the core principles established in this work-cost sensitivity, transparency, and privacy.

5.8 Conclusion

This research developed a cost-aware, explainable, and privacy-compatible method for financial fraud detection under class imbalance. The method achieved some quantitative operational and interpretive advantages through the integrated synthetic data generation, cost-sensitive thresholding, and explainable validation. They showed that improved recall and reduced expected loss can be integrated with interpretability and privacy protections, if intended to complement, not compete goals.

In addition to performance improvements, the study also provides clear links between model design and institutional accountability. It shows how AI systems in finance can be both effective and ethical in applying fairness, auditability, and decision support in their core logic.

The research thus extends the boundary of fraud detection from algorithmic experimentation to methodological innovation by demonstrating that regulatory-ready AI systems can be technically sound and socially responsible. Future development in federated and adaptive learning can use this model as a basis for collaborative, privacy-conscious fraud prevention across institutions and borders.

This study demonstrates that a decision-centered, cost-aware fraud detection framework can reduce operational loss and maintain governance integrity. Cost-aware thresholding delivered measurable savings, privacy-safe SHAP parity enabled transparent explanations without exposing customer data, and anomaly-feature fusion improved cost efficiency while preserving recall. The negative CTGAN outcome contributes valuable evidence for governing synthetic data under auditable precision–cost thresholds.

References

Abdelghafour, E. B., Mohamed, C., Aknin, N., & Bouzidi, A. (2024). Enhancing Credit Card Fraud Detection Using a Stacking Model Approach and Hyperparameter Optimization. *International Journal of Advanced Computer Science and Applications*, *15*(10). https://doi.org/10.14569/ijacsa.2024.01510110

Abdi, S. J. (2023). A comprehensive guide for selecting appropriate statistical tests: Understanding when to use parametric and nonparametric tests. *Open Journal of Statistics*, *13*(4), 464-474, doi: 10.4236/ojs.2023.134023

Ahammad, J., Hossain, N., & Alam, M. S. (2020). *Credit Card Fraud Detection using Data Pre-processing on Imbalanced Data - both Oversampling and Undersampling*. https://doi.org/10.1145/3377049.3377113

Alarfaj, F. K., & Shahzadi, S. (2024). Enhancing Fraud Detection in Banking with Deep Learning: Graph Neural Networks and Autoencoders for Real-Time Credit Card Fraud Prevention. *IEEE Access*, *13*, 20633. https://doi.org/10.1109/access.2024.3466288

Aljunaid, S. K., Almheiri, S. J., Dawood, H., & Khan, M. A. (2025). Secure and Transparent Banking: Explainable AI-Driven Federated Learning Model for Financial Fraud Detection. *Journal of Risk and Financial Management*, *18*(4), 179. https://doi.org/10.3390/jrfm18040179

Almalki, F., & Masud, M. (2025). *Financial Fraud Detection Using Explainable AI and Stacking Ensemble Methods*. https://doi.org/10.48550/ARXIV.2505.10050

Alothman, R. B., Talib, H. A., & Mohammed, M. S. (2022). Fraud detection under the unbalanced class based on gradient boosting. *Eastern-European Journal of Enterprise Technologies*, *2*, 6. https://doi.org/10.15587/1729-4061.2022.254922

An, H., Ma, R., Yan, Y., Chen, T., Zhao, Y., Li, P., Li, J., Wang, X., Fan, D., & Lv, C. (2024). Finsformer: A Novel Approach to Detecting Financial Attacks Using Transformer and Cluster-Attention. *Applied Sciences*, *14*(1), 460. https://doi.org/10.3390/app14010460

Aswani, Mrs. D. (2025). Financial Fraud Detection Using Explainable AI and Federated Learning. *International Journal for Research in Applied Science and Engineering Technology*, *13*(5), 7568. https://doi.org/10.22214/ijraset.2025.71922

Awosika, T., Shukla, R. M., & Pranggono, B. (2024). Transparency and Privacy: The Role of Explainable AI and Federated Learning in Financial Fraud Detection. *IEEE Access*, *12*, 64551. https://doi.org/10.1109/access.2024.3394528

Babu, E. V., Maliakal, J. J., Neethu, T. V., & Babu, D. R. (2024). *Performance Comparison of XGBoost and CatBoost Algorithm in Credit Card Fraud Detection with Streamlit-Based Web Application*. 1. https://doi.org/10.1109/areis62559.2024.10893617

Backiyalakshmi, V., & Umadevi, B. (2024, September). Fraud detection in banking industry using enhanced intellectual framework with optimization strategy. In *2024 5th international conference on smart electronics and communication (ICOSEC)* (pp. 1022-1031). IEEE, doi: 10.1109/ICOSEC61587.2024.10722599

Bakumenko, A., Hlaváčková-Schindler, K., Plant, C., & Hubig, N. (2024). Advancing Anomaly Detection: Non-Semantic Financial Data Encoding with LLMs. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.03614

Bathala, N. K., Sasikala, T. S., Prasad, V., Begum, S. L. R., Mahajan, U., & Sreedevi, V. (2025). Innovative Fraud Detection in Financial Transactions using Deep Learning and SHAP Analysis. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5083759

Becerra-Suarez, F. L., Alvarez-Vasquez, H., & Forero, M. G. (2025). Improvement of Bank Fraud Detection Through Synthetic Data Generation with Gaussian Noise. *Technologies*, *13*(4), 141. https://doi.org/10.3390/technologies13040141

Chagahi, M. H., Delfan, N., Dashtaki, S. M., Moshiri, B., & Piran, Md. J. (2024). *Explainable AI for Fraud Detection: An Attention-Based Ensemble of CNNs, GNNs, and A Confidence-Driven Gating Mechanism*. https://doi.org/10.48550/ARXIV.2410.09069

Chandrashekar, R., Onyema, E. M., Mallik, S., Kavin, B. P., Archakam, J. K. K., Annamalai, D. A., & Shah, M. A. (2024). Optimized Feature Selection and Enhanced Recurrent Neural Network for Financial Fraud Detection. *Research Square (Research Square)*. https://doi.org/10.21203/rs.3.rs-5147684/v1

Charitou, C., Dragičević, S., & Garcez, A. S. d'Avila. (2021). Synthetic Data Generation for Fraud Detection using GANs. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2109.12546

Chen, Y., Zhao, C., Xu, Y., Nie, C., & Zhang, Y. (2025). *Year-over-Year Developments in Financial Fraud Detection via Deep Learning: A Systematic Literature Review*. https://doi.org/10.48550/ARXIV.2502.00201

Cheng, D., Zou, Y., Xiang, S., & Jiang, C. (2025). Graph neural networks for financial fraud detection: a review [Review of *Graph neural networks for financial fraud detection: a*

review]. *Frontiers of Computer Science*, *19*(9). Higher Education Press. https://doi.org/10.1007/s11704-024-40474-y

Cheng, Y., Wang, C.-H., Potluru, V. K., Balch, T., & Cheng, G. (2024). Downstream Task-Oriented Generative Model Selections on Synthetic Data Training for Fraud Detection Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2401.00974

DeLise, T. (2023). Deep Semi-Supervised Anomaly Detection for Finding Fraud in the Futures Market. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2309.00088

Detthamrong, U., Chansanam, W., Boongoen, T., & Iam-On, N. (2024). Enhancing Fraud Detection in Banking using Advanced Machine Learning Techniques. *International Journal of Economics and Financial Issues*, *14*(5), 177. https://doi.org/10.32479/ijefi.16613

Devi, C. M., Gobinath, A., Priya, S., Adithiyaa, M., Chandru, M., & Jothi, M. (2024). Next-Generation Anomaly Detection Framework Leveraging Artificial Intelligence for Proactive Credit Card Fraud Prevention and Risk Management. *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1. https://doi.org/10.1109/icccnt61001.2024.10725285

Faisal, N. A., Nahar, J., Sultana, N., & Mintoo, A. A. (2024). Fraud Detection In Banking Leveraging Ai To Identify And Prevent Fraudulent Activities In Real-Time. *Non Human Journal.*, *1*(1), 181. https://doi.org/10.70008/jmldeds.v1i01.53

Fronc, M., & Jakubczyk, M. (2022). From business to clinical trials: a systematic review of the literature on fraud detection methods to be used in central statistical monitoring [Review of *From business to clinical trials: a systematic review of the literature on fraud detection*

methods to be used in central statistical monitoring]. *Przegląd Statystyczny Statistical Review*, *69*(3), 1. Index Copernicus International S.A. https://doi.org/10.5604/01.3001.0016.1165

Geng, J., & Zhang, B. (2023). Credit Card Fraud Detection Using Adversarial Learning. *2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, 891. https://doi.org/10.1109/icicml60161.2023.10424872

Giusti, C., Guarnera, L., Casu, M., & Battiato, S. (2025). *Fraud is Not Just Rarity: A Causal Prototype Attention Approach to Realistic Synthetic Oversampling*. https://doi.org/10.48550/ARXIV.2507.14706

Gourinda, A., El Bouri, H., Mahdi, Z., Bousgheiri, F., Sammoud, K., Lembrabett, S., & Najdi, A. (2024). A Practical Guide to Statistical Tests in the Biomedical Field: From Parametric to Nonparametric, Where and How?. *Journal of Biosciences and Medicines*, *12*(11), 1-14, doi: https://doi.org/10.4236/jbm.2024.1211001

Hashemi, S. K., Mirtaheri, S. L., & Greco, S. (2022). Fraud Detection in Banking Data by Machine Learning Techniques. *IEEE Access*, *11*, 3034. https://doi.org/10.1109/access.2022.3232287

Hayat, K., & Magnier, B. (2025). Data Leakage and Deceptive Performance: A Critical Examination of Credit Card Fraud Detection Methodologies. *Mathematics*, *13*(16), 2563. https://doi.org/10.3390/math13162563

John, A., & Ahsun, A. (2025). *Explainable AI (XAI) for Fraud Detection: Building Trust and Transparency in AI-Driven Financial Security Systems Authors*. https://doi.org/10.2139/ssrn.5285281

Kadam, P. (2024). Financial Fraud Detection using Jump-Attentive Graph Neural Networks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2411.05857

Kamuangu, P. (2024). A review on financial fraud detection using ai and machine learning. *Journal of Economics, Finance, and Accounting Studies*, *6*(1), 67, https://doi.org/10.32996/jefas.2024.6.1.7

Kennedy, R. K. L., Salekshahrezaee, Z., Villanustre, F., & Khoshgoftaar, T. M. (2023). Iterative cleaning and learning of big highly-imbalanced fraud data using unsupervised learning. *Journal Of Big Data*, *10*(1). https://doi.org/10.1186/s40537-023-00750-3

Khaled, A., Hasan, M. M., Islam, S., Papastergiou, S., & Mouratidis, H. (2024). Synthetic Data Generation and Impact Analysis of Machine Learning Models for Enhanced Credit Card Fraud Detection. In *IFIP advances in information and communication technology* (p. 362). Springer Science+Business Media. https://doi.org/10.1007/978-3-031-63211-2_27

Khalid, A. R., Owoh, N., Uthmani, O., Ashawa, M., Osamor, J., & Adejoh, J. (2024). Enhancing Credit Card Fraud Detection: An Ensemble Machine Learning Approach. *Big Data and Cognitive Computing*, *8*(1), 6. https://doi.org/10.3390/bdcc8010006

Khosravi, H., Farhadpour, S., Grandhi, M., Raihan, A. S., Das, S., & Ahmed, I. (2024). Strategic data augmentation with CTGAN for smart manufacturing: Enhancing ML predictions of paper breaks in pulp-and-paper production. Manufacturing Letters, 41, 1312–1323. https://doi.org/10.1016/j.mfglet.2024.09.158

Kong, M., Li, R., Wang, J., Li, X., Jin, S., Xie, W., Hou, M., & Cao, C. (2024). CFTNet: a robust credit card fraud detection model enhanced by counterfactual data augmentation. *Neural Computing and Applications*, *36*(15), 8607. https://doi.org/10.1007/s00521-024-09546-9

ku, R., Dubey, A. K., Narang, S. K., & Kishore, N. (2024). A Machine Learning Based Approach for the Fraud Detection in Imbalanced Credit Card Transaction Dataset. *International Journal of Electronics and Communication Engineering*, *11*(8), 244. https://doi.org/10.14445/23488549/ijece-v11i8p124

Lange, P. E. de, Melsom, B., Vennerød, C. B., & Westgaard, S. (2022). Explainable AI for Credit Assessment in Banks. *Journal of Risk and Financial Management*, *15*(12), 556. https://doi.org/10.3390/jrfm15120556

Langevin, A., Cody, T., Adams, S., & Beling, P. A. (2021). Generative adversarial networks for data augmentation and transfer in credit card fraud detection. *Journal of the Operational Research Society*, *73*(1), 153. https://doi.org/10.1080/01605682.2021.1880296

Lee, J., Jung, D., Moon, J., & Rho, S. (2024). Advanced R-GAN: Generating anomaly data for improved detection in imbalanced datasets using regularized generative adversarial networks. *Alexandria Engineering Journal*, *111*, 491. https://doi.org/10.1016/j.aej.2024.10.084

Li, K., Yang, T., Zhou, M., Meng, J., Wang, S., Wu, Y., Tan, B., Hu, S., Pan, L., Fan, Y., Sheng, Z., & Tong, Y. (2024). SEFraud: Graph-based Self-Explainable Fraud Detection via Interpretative Mask Learning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.11389

Lin, J. (2025). *A Novel Approach in Financial Fraud Detection: Addressing Data Imbalance with Prompt Engineering, Leveraging Large Language Model Embeddings for Fraud Scoring, and Cost-Effectiveness Optimization*, https://www.proquest.com/dissertations-

theses/novel-approach-financial-fraud-detection/docview/3142139348/se-2?accountid=11243

Mehta, H. (2024). Analysis of Different Machine Learning Models for Credit Card Fraud Detection. *Research Square (Research Square)*. https://doi.org/10.21203/rs.3.rs-5314340/v1

Mienye, I. D., & Swart, T. G. (2024). A Hybrid Deep Learning Approach with Generative Adversarial Network for Credit Card Fraud Detection. *Technologies*, *12*(10), 186. https://doi.org/10.3390/technologies12100186

Motie, S., & Raahemi, B. (2023). Financial fraud detection using graph neural networks: A systematic review [Review of *Financial fraud detection using graph neural networks: A systematic review*]. *Expert Systems with Applications*, *240*, 122156. Elsevier BV. https://doi.org/10.1016/j.eswa.2023.122156

Nicholls, J., Kuppa, A., & Le-Khac, N. (2021). Financial Cybercrime: A Comprehensive Survey of Deep Learning Approaches to Tackle the Evolving Financial Crime Landscape. *IEEE Access*, *9*, 163965. https://doi.org/10.1109/access.2021.3134076

Odufisan, O. I., Abhulimen, O. V., & Ogunti, E. (2025). Harnessing Artificial Intelligence and Machine Learning for Fraud Detection and Prevention in Nigeria. *Journal of Economic Criminology*, 100127. https://doi.org/10.1016/j.jeconc.2025.100127

Olowu, O., Adeleye, A. O., Omokanye, A. O., Ajayi, A. A., Adepoju, A. O., Omole, O. M., & Chianumba, E. C. (2024). AI-driven fraud detection in banking: A systematic review of data science approaches to enhancing cybersecurity [Review of *AI-driven fraud detection in banking: A systematic review of data science approaches to enhancing*

cybersecurity]. *GSC Advanced Research and Reviews*, *21*(2), 227. https://doi.org/10.30574/gscarr.2024.21.2.0418

Oreški, G. (2023). Synthesizing credit data using autoencoders and generative adversarial networks. *Knowledge-Based Systems*, *274*, 110646. https://doi.org/10.1016/j.knosys.2023.110646

Patel, A., Patel, M. I., & Patel, P. (2024). Exploring Supervised Machine Learning Techniques for Detecting Credit Card Fraud: An Investigative Review. *ITM Web of Conferences*, *65*, 3006. https://doi.org/10.1051/itmconf/20246503006

Patel, P., Chauhan, S. S., Gupta, S., Gupta, T., & Agrawal, R. (2024). Mitigating Class Imbalance with Ensemble SMOTEfied-GAN: Advancing Detection Strategies for Automobile Insurance Fraud. *International Journal of Fuzzy Logic and Intelligent Systems*, *24*(4), 333. https://doi.org/10.5391/ijfis.2024.24.4.333

Psychoula, I., Gutmann, A., Mainali, P., Lee, S. H., Dunphy, P., & Petitcolas, F. A. P. (2021). Explainable Machine Learning for Fraud Detection. *Computer*, *54*(10), 49. https://doi.org/10.1109/mc.2021.3081249

Pundkar, S., & Zubei, M. (2023). Credit Card Fraud Detection Methods: A Review [Review of *Credit Card Fraud Detection Methods: A Review*]. *E3S Web of Conferences*, *453*, 1015. EDP Sciences. https://doi.org/10.1051/e3sconf/202345301015

Rao, S. X., Jiang, J., Han, Z., & Yin, H. (2025). Fraud Detection in E-Commerce: A Systematic Review of Transaction Risk Prevention [Review of *Fraud Detection in E-Commerce: A Systematic Review of Transaction Risk Prevention*]. *IntechOpen eBooks*. IntechOpen. https://doi.org/10.5772/intechopen.1009640

Rocha, J., Alves, M., Oliveira, R., & Santos, F. S. dos. (2025). *Comparing Ensemble Methods for Credit Card Fraud Detection: A Performance Analysis on Multiple Datasets*. https://doi.org/10.20944/preprints202509.0106.v1

Shakeri, A., Chen, P., Shu, Y., Yang, L., Zhang, W. E., & Chen, W. (2024). *Transforming Data Product Generation through Federated Learning: An Exploration of FL Applications in Data Ecosystems*. *27*, 84. https://doi.org/10.1109/icws62655.2024.00027

SP, S. I., Jayasree, A., Poojitha, M., Jyothi, N. A., Namithaa, K., & Sushila, Y. (2024, December). *Credit Card Fraud Detection: Machine Learning and Data Analytical Approach for Accuracy and Comparative Analysis*. In *2024 International Conference on Sustainable Communication Networks and Application (ICSCNA)* (pp. 1461-1468). IEEE. https://doi.org/10.1109/icscna63714.2024.10864225

Strelcenia, E., & Prakoonwit, S. (2022). *Generating Synthetic Data for Credit Card Fraud Detection Using GANs*. 42. https://doi.org/10.1109/cait56099.2022.10072179

Tamanna, T., Kamboj, S., Singh, L., & Kaur, T. (2024). *Automated Fraud Detection in Financial Transactions using Machine Learning: An Ensemble Perspective*. 1. https://doi.org/10.1109/aimla59606.2024.10531422

Tayebi, M., & Kafhali, S. E. (2025). Generative Modeling for Imbalanced Credit Card Fraud Transaction Detection. *Journal of Cybersecurity and Privacy*, *5*(1), 9. https://doi.org/10.3390/jcp5010009

Thilagavathi, M., Saranyadevi, R., Vijayakumar, N., Selvi, K., Anitha, L., & Sudharson, K. (2024). *AI-Driven Fraud Detection in Financial Transactions with Graph Neural Networks and Anomaly Detection*. 1. https://doi.org/10.1109/icstem61137.2024.10560838

Vallarino, D. (2025a). *Detecting Financial Fraud with Hybrid Deep Learning: A Mix-of-Experts Approach to Sequential and Anomalous Patterns*. https://doi.org/10.48550/ARXIV.2504.03750

Vallarino, D. (2025b). *AI-Powered Fraud Detection in Financial Services: GNN, Compliance Challenges, and Risk Mitigation*. https://doi.org/10.2139/ssrn.5170054

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & Van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, *17*(3), 261-272, https://doi.org/10.1038/s41592-019-0686-2

Vynokurova, O., Peleshko, D., Бондаренко, O., Ilyasov, V., Serzhantov, V., & Peleshko, M. (2020). *Hybrid Machine Learning System for Solving Fraud Detection Tasks*. 1. https://doi.org/10.1109/dsmp47368.2020.9204244

Wang, C., Nie, C., & Liu, Y. (2025). Evaluating Supervised Learning Models for Fraud Detection: A Comparative Study of Classical and Deep Architectures on Imbalanced Transaction Data. In *Advances in economics, business and management research/Advances in Economics, Business and Management Research* (p. 613). Atlantis Press. https://doi.org/10.2991/978-94-6463-835-6_65

Xia, Y., Wang, C., Mabry, J., & Cheng, G. (2024). Advancing Retail Data Science: Comprehensive Evaluation of Synthetic Data. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.13130

Y. Chen, G. Sun and L. Ge, "Momentum Study Based on LightGBM Model and Spearman Correlation Analysis," *2024 3rd International Conference on Data Analytics, Computing*

and Artificial Intelligence (ICDACAI), Zakopane, Poland, 2024, pp. 813-818, doi: 10.1109/ICDACAI65086.2024.00153

Yaseen, H., & Al-Amarneh, A. (2025). Adoption of Artificial Intelligence-Driven Fraud Detection in Banking: The Role of Trust, Transparency, and Fairness Perception in Financial Institutions in the United Arab Emirates and Qatar. *Journal of Risk and Financial Management*, *18*(4), 217. https://doi.org/10.3390/jrfm18040217

Zhang, H., Hong, J., Dong, F., Drew, S., Xue, L., & Zhou, J. (2023). A Privacy-Preserving Hybrid Federated Learning Framework for Financial Crime Detection. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.2302.03654

Zhang, L., Geisler, T., Ray, H. E., & Xie, Y. (2021). Improving logistic regression on the imbalanced data by a novel penalized log-likelihood function. *Journal of Applied Statistics*, *49*(13), 3257. https://doi.org/10.1080/02664763.2021.1939662

Zhu, M., Gong, Y., Xiang, Y., Yu, H., & Huo, S. (2024). *Utilizing GANs for fraud detection: model training with synthetic transaction data*. *5*, 310. https://doi.org/10.1117/12.3034346