EAI Endorsed Transactions
on Intelligent Systems and Machine Learning Applications
Research Article EAI.EU

Enhancing Credit Card Fraud Detection under Severe Class Imbalance using Cost-Sensitive Learning and Threshold Optimization

Vu Ngoc Thanh Sang[1,*]

[1]ICIP-Lab, Faculty of Information Technology, Saigon University, Ho Chi Minh City, Vietnam

Abstract

INTRODUCTION: Credit card fraud detection remains challenging because fraudulent transactions are rare, fraud patterns evolve over time, and precision–recall trade-offs vary under different class prevalences. While synthetic oversampling methods such as SMOTE are widely used, they may alter the observed feature distribution and complicate deployment interpretation. OBJECTIVES: This study develops a cost-sensitive and decision-threshold-calibrated fraud detection framework that preserves the original data distribution under severe class imbalance. METHODS: The framework combines cost-sensitive XGBoost and a Multi-Layer Perceptron trained with weighted Binary Cross-Entropy and Focal Loss. Hyperparameters are tuned using Optuna within a leakage-conscious validation protocol, and class imbalance is handled through scale-aware weighting rather than synthetic resampling. Decision thresholds are selected using minority-class $F1$ and an illustrative amount-aware cost criterion. SHAP analysis and a chronological split of the 2013 dataset are used to examine transformed-feature auditability and near-future generalization. RESULTS: On the imbalanced 2013 dataset, the optimized XGBoost model improves test-set PR-AUC from 0.7809 to 0.8815 and minority-class $F1$ from 0.7919 to 0.8497, with false positives decreasing from 21 to 13 on a test partition containing 98 fraud cases. On the balanced 2023 dataset, ranking performance is near-saturated and threshold selection mainly reduces false alarms. Under the illustrative assumption $C_{FP} = 1$, amount-aware thresholding yields a lower simulated cost than the $F1$-optimized threshold, but with substantially more false positives. Chronological validation on the 2013 dataset yields lower PR-AUC than random stratified evaluation. CONCLUSION: The results suggest that fraud detection under class imbalance benefits from separating ranking optimization from decision-threshold selection. However, cost-aware and temporal findings should be interpreted as benchmark-based, deployment-motivated analyses rather than direct evidence of operational deployment readiness.

Received on 03 March 2026; accepted on 19 May 2026; published on 01 June 2026

Keywords: credit card fraud detection, class imbalance, cost-sensitive learning, threshold calibration, XGBoost, neural networks

doi:10.4108/eetismla.12078

1. Introduction

Credit card fraud remains a major challenge for global financial systems, causing substantial economic losses annually. The rapid expansion of digital payment infrastructures, combined with increasingly sophis-ticated fraudulent strategies, requires detection sys-tems that are not only accurate but also computa-tionally efficient and operationally stable. In both aca-demic research and industrial practice, publicly avail-able benchmark datasets are widely used to evaluate machine learning (ML) models for fraud detection. Nev-ertheless, a persistent gap exists between benchmark-driven performance improvements and deployment-ready reliability under real-world operating conditions.

*Corresponding author. Email: vntsang@sgu.edu.vn

Unlike conventional binary classification problems, credit card fraud detection presents distinct computational characteristics. Most notably, the task is governed by extreme class imbalance, where legitimate transactions vastly outnumber fraudulent ones. Such skewed distributions complicate model training and evaluation, as aggregate metrics such as accuracy may obscure poor minority-class performance. Moreover, fraud patterns evolve over time, leading to shifts in feature distributions and increasing the difficulty of maintaining stable precision–recall trade-offs. From an operational perspective, fraud detection systems must simultaneously minimize false positives to reduce customer friction and maximize recall to prevent financial loss, all while satisfying strict latency requirements.

To address class imbalance, data-level approaches such as the Synthetic Minority Oversampling Technique (SMOTE) are frequently adopted. While these methods increase minority-class representation, they may alter the original data distribution through interpolation in high-dimensional feature space. In settings where features are anonymized or transformed (e.g., via Principal Component Analysis), synthetic sample generation may introduce additional uncertainty in the learned decision boundary. Although oversampling techniques can improve validation metrics in controlled experiments, their robustness under varying class prevalences and deployment constraints remains uncertain. These limitations motivate the exploration of model-level and decision-level strategies that preserve the natural data distribution while explicitly accounting for imbalance.

This study proposes a cost-sensitive and decision-threshold-calibrated fraud detection framework built upon Extreme Gradient Boosting (XGBoost) with a complementary neural baseline. Rather than modifying the training distribution through synthetic resampling, class imbalance is handled via scale-aware weighting and systematic hyperparameter optimization using Optuna within a leakage-conscious validation protocol designed to reduce information leakage during model selection. A dynamic threshold calibration strategy is further employed to align decision boundaries with minority-class objectives, including the $F1$ score, precision–recall trade-offs, and amount-aware financial cost. The framework is evaluated on two benchmark datasets representing distinct class distribution regimes, including an extremely imbalanced dataset (0.17% fraud prevalence) and a class-balanced dataset.

In this study, the term calibration refers to decision-threshold calibration rather than probability calibration. No post-hoc probability calibration method, such as Platt scaling, isotonic regression, or temperature scaling, is applied. Therefore, model scores are used for ranking and threshold selection, but they should not be interpreted as fully calibrated probabilities. The main contributions of this work are as follows:

(i) A reproducible cost-sensitive learning pipeline that preserves the original data distribution without reliance on synthetic oversampling.

(ii) A decision-level calibration mechanism that explicitly separates ranking optimization from operational threshold selection, enabling stable performance across heterogeneous class distributions.

(iii) A deployment-motivated evaluation that integrates SHAP-based transformed-feature auditability, amount-aware threshold analysis, and chronological validation to better approximate operational concerns in benchmark-based fraud detection.

2. Literature Review

Table 1 summarizes representative credit card fraud detection (CCFD) studies, highlighting differences in dataset characteristics, imbalance mitigation strategies, evaluation metrics, and treatment of decision thresholds. A recurring pattern across the literature is the discrepancy between benchmark-centered experimentation and operational deployment settings, where extreme rarity, non-stationarity, and asymmetric misclassification costs dominate real-world requirements.

2.1. Benchmark Dependence and Class Rarity

A substantial proportion of CCFD studies rely on the 2013 European cardholder dataset, which is anonymized and heavily preprocessed. While this dataset enables standardized comparison, several studies report near-saturated performance on this benchmark [1, 2]. Such results may not reflect deployment scenarios in which fraudulent events are extremely rare and the data-generating process evolves over time.

Empirical evidence from large-scale fraud contexts suggests that class rarity itself strongly influences achievable performance. Bauder et al. [10] demonstrate consistent degradation in detection effectiveness as minority prevalence decreases across multiple datasets and learning algorithms. Herland et al. [9] further show that rarity affects evaluation outcomes and may alter conclusions drawn from conventional validation protocols. These findings indicate the importance of evaluation procedures that explicitly account for extreme imbalance rather than relying solely on benchmark-driven comparisons.

Table 1. Strategic comparison of recent credit card fraud detection approaches

Ref	Dataset Type	Imbalance Handling	Evaluation Focus	Threshold / Calibration	Key Limitation
[1]	ULB + Synthetic	SMOTE / ROS	Accuracy, Precision, Recall	Not reported	Benchmark-heavy results; limited deployment evidence
[2]	ULB + IEEE-CIS	WGAN-based oversampling	AUC, Recall	Not reported	Synthetic augmentation; calibration not discussed
[3]	ULB + PaySim	SMOTE-ENN	Accuracy, F1-score	Not reported	Resampling-centric; threshold analysis absent
[4]	Proprietary Banking Data	Cost-sensitive trees	Cost-aware metrics	Incorporated in rule design	Threshold optimization not systematically studied
[5]	Credit scoring datasets	Cost-sensitive boosting	AUC, G-Mean, Type I/II errors	Implicit via loss	Limited emphasis on PR-AUC and post-hoc calibration
[6]	ULB Dataset	Threshold optimization	AUPRC, TPR/TNR constraints	Explicit tuning	Not integrated with cost-sensitive boosting
[7]	ULB	SMOTE + weight tuning	Accuracy, AUC, F1	Not explicit	Calibration policy not formalized
[8]	Proprietary Data	ROS, RUS, SMOTE, ADASYN	F1-score	Not reported	Minority performance remains challenging
[9]	Medicare Big Data	RUS analysis	Train/Test vs CV comparison	Not threshold-focused	Demonstrates instability under severe rarity

2.2. Imbalance Handling: Data-Level and Algorithm-Level Strategies

Imbalance mitigation in CCFD is generally categorized into data-level and algorithm-level strategies. Data-level approaches include random oversampling and SMOTE-family techniques [3, 11]. More recent studies explore generative augmentation methods, such as GAN-based sampling [2]. Although such methods often improve recall on benchmark datasets, their robustness under transformed feature spaces and operational distributions is less consistently validated. In addition, the interaction between synthetic augmentation and deployment-level threshold selection is not always examined.

Algorithm-level strategies modify the learning objective while preserving the observed data distribution. Cost-sensitive learning introduces asymmetric penalties for false negatives and false positives. Elkan [12] established foundational principles for cost-sensitive decision-making, emphasizing that optimal decisions under unequal costs require explicit consideration of decision policies in addition to model training. Subsequent studies applied cost-sensitive designs to semi-supervised learning [13], tree-based ensembles [4], cascade forests [14], and focal-aware boosted trees [5]. These approaches aim to enhance minority-class utility without exclusive reliance on aggressive resampling.

2.3. Evaluation Under Extreme Imbalance and Decision Calibration

Evaluation methodology significantly influences conclusions in rare-event detection. Under extreme class imbalance, aggregate metrics such as accuracy and ROC-AUC may provide optimistic signals due to the dominance of true negatives. Davis and Goadrich [15] demonstrate that Precision–Recall (PR) analysis more accurately reflects performance trade-offs in highly skewed datasets. Consequently, recent studies increasingly prioritize PR-AUC and minority-class $F1$ as model selection criteria.

Reliable evaluation further requires strict separation between training, validation, and testing stages. Improper hyperparameter tuning or threshold adjustment can introduce optimistic bias. Empirical analyses [9, 10] show that prevalence variation alone can substantially alter performance estimates. Leakage-resistant protocols, such as nested stratified cross-validation, are therefore recommended to decouple model selection from final assessment.

Operational fraud detection also requires translating probabilistic scores into actionable decisions. Threshold calibration offers a principled mechanism for adjusting the decision boundary without modifying the training distribution [6]. Classical criteria such as Youden's index [16] and decision-theoretic approaches such as Neyman–Pearson optimization [17] illustrate that ranking optimization and threshold selection constitute distinct stages in detection system design.

Finally, fraud detection operates in dynamic environments where class prevalence and transaction patterns may drift over time [18]. Empirical evidence indicates that prevalence variation alone can destabilize evaluation and degrade minority performance [9, 10]. These observations highlight the importance of separating score optimization from decision calibration to maintain robustness across evolving deployment conditions.

Despite these advances, three gaps remain important for deployment-oriented CCFD evaluation. First, many benchmark studies report predictive performance without model explanations, limiting analyst trust and auditability. Second, minority-class $F1$ is frequently used as a summary metric even though losses from false negatives and operational costs from false positives are asymmetric. Third, random stratified validation can mix earlier and later transactions, potentially overstating future-period performance. These gaps motivate the present study, which combines imbalance-aware training with SHAP-based explanation, amount-aware threshold analysis, and chronological validation.

3. Materials and Methods

3.1. Data Description

Two publicly available benchmark datasets are used to evaluate the proposed framework: the 2013 European cardholder credit card dataset [19] and a 2023 credit card fraud dataset [20]. These datasets represent contrasting class distribution regimes and serve as complementary testbeds for assessing robustness under heterogeneous imbalance conditions.

The 2013 dataset, provided by the Machine Learning Group at Université Libre de Bruxelles, contains 284,807 transactions, of which 492 are fraudulent (0.172% prevalence). Due to confidentiality constraints, the original attributes were transformed via Principal Component Analysis (PCA), yielding 28 anonymized features ($V1$–$V28$), along with the raw variables `Time` and `Amount`. This dataset represents an extreme-imbalance scenario with latent feature representation.

In contrast, the 2023 dataset exhibits a substantially more balanced class distribution and includes engineered transactional features. It is used as a balanced-reference setting to examine model behavior when prevalence skew is reduced. Because the dataset produces near-saturated ranking performance across all evaluated models, results on this dataset are interpreted as evidence of threshold behavior under a highly separable benchmark setting rather than as evidence of real-world deployment difficulty.

For both datasets, a stratified 66/14/20 split is adopted for the main training, validation, and testing protocol. Stratification preserves class proportions across splits. The validation set is used exclusively for hyperparameter tuning and threshold calibration, while the test set remains untouched until final evaluation. To probe near-future generalization, an additional chronological split is conducted on the 2013 dataset by sorting transactions using the `Time` variable and assigning the first 66%, next 14%, and final 20% to training, validation, and testing, respectively. The resulting fraud prevalences are 0.196% (368/187,972), 0.123% (49/39,873), and 0.132% (75/56,962) for the training, validation, and test windows, respectively. Because the 2013 dataset covers only an approximately 48-hour observation window, this experiment should be interpreted as a short-horizon chronological validation rather than a full concept-drift analysis. The 2023 dataset is not evaluated chronologically because it does not provide an explicit transaction-time variable; its `id` column is therefore not assumed to represent temporal order.

3.2. Preprocessing and Feature Engineering

All preprocessing operations are fitted exclusively on training data and subsequently applied to validation and test partitions using learned parameters to reduce the risk of information leakage.

Identifier columns (e.g., `id`) and the `Time` variable are removed from model inputs. In the chronological validation experiment, `Time` is retained only for ordering transactions before splitting and is not used as a predictive feature. When used as a predictive input, the `Amount` feature is transformed as $\log(1 + \texttt{Amount})$ to reduce heavy-tailed skewness. Separately, the original untransformed transaction amount is retained only for the amount-aware cost analysis, where it serves as a proxy for missed-fraud exposure. Thus, the model input representation and the cost-evaluation amount are kept distinct. No additional PCA is applied to the 2013 dataset, as features are already provided in transformed latent space.

Neural network models employ `StandardScaler`, which standardizes features to zero mean and unit variance, stabilizing gradient-based optimization. For gradient boosting models, scaling is not strictly required due to the scale-invariant nature of tree-based splits. However, `RobustScaler` and a normal-output `QuantileTransformer` are evaluated to examine whether distribution smoothing enhances minority-class separability. Scaling parameters are computed only on training folds and applied to validation subsets.

Synthetic oversampling methods, including SMOTE, are not used in the final reported results. This design choice preserves the observed training distribution and allows the study to focus on model-level weighting and decision-threshold selection. A systematic comparison against oversampling methods under distribution shift is left for future work. For the 2013 dataset, inverse-frequency class weights are incorporated into neural

network training to mitigate gradient bias toward the majority class. For the 2023 dataset, imbalance-aware weighting is not required due to near-balanced prevalence.

3.3. Model Architectures

Two modeling paradigms are investigated: a gradient boosting framework and a deep neural network architecture. Both models are evaluated under identical data partitions and threshold calibration procedures.

Gradient Boosting Framework Extreme Gradient Boost-ing (XGBoost) is selected for its effectiveness on struc-tured tabular data and its ability to model nonlinear feature interactions via additive tree ensembles. The baseline configuration uses default hyperparameters and a probability threshold of $\tau = 0.5$.

An optimized variant, denoted XGB (Optuna+τ), is obtained through automated hyperparameter tuning using Optuna with a Tree-structured Parzen Estimator sampler and median pruning. The search space includes tree depth, minimum child weight, learning rate, subsampling ratios, regularization coefficients, and the `scale_pos_weight` parameter. Final predictions are generated using a calibrated threshold τ, selected to maximize $F1$ on validation data.

Deep Learning Framework A Multi-Layer Perceptron (MLP) is employed to model nonlinear decision boundaries in high-dimensional feature spaces. Dropout regularization is applied to reduce overfitting under class imbalance. The baseline configuration consists of fully connected layers with ReLU activation and Binary Cross-Entropy loss using $\tau = 0.5$.

An enhanced variant, denoted MLP (BCE+Focal+τ), integrates weighted Binary Cross-Entropy with Focal Loss to emphasize hard minority instances. Class weights and focusing parameters are tuned to improve minority discrimination. As with XGBoost, a calibrated threshold τ is selected to maximize $F1$.

Hyperparameter Optimization and Threshold Calibration Hyperparameter selection is performed using 5-fold stratified c ross-validation w ithin t he 6 6% training partition. Optuna maximizes the mean validation $F1$ across the inner folds, while all preprocessing steps are fitted only on the corresponding training folds to reduce information leakage.

After hyperparameter selection, the final m odel is retrained on the full 66% training partition. The 14% validation partition is then used exclusively for decision-threshold selection. For the main classification results, the threshold τ is selected by maximizing minority-class $F1$ on the validation partition. For the amount-aware analysis, τ is selected by minimizing the validation-set TotalCost(τ). Final metrics are reported once on the untouched 20% test partition. Thus, the reported test results are single held-out test estimates rather than averages over outer cross-validation folds.

Amount-Aware Cost Evaluation Because false positives and false negatives have asymmetric consequences in fraud detection, an additional amount-aware cost analysis is introduced for the revised evaluation. For a threshold τ, the simulated financial cost is defined as

$$\text{TotalCost}(\tau) = C_{FP} \cdot FP(\tau) + \sum_{i \in FN(\tau)} \text{Amount}_i,$$

where C_{FP} denotes the assumed cost of reviewing one false-positive alert, expressed in the same normalized unit as the transaction amount. In all reported experiments, $C_{FP} = 1$ is used as an illustrative exchange rate between one manual-review action and one unit of missed-fraud exposure. Each false negative is weighted by its original transaction amount as a proxy for unrecovered fraud exposure. Because chargeback recovery, customer reimbursement, investigation cost, customer-friction cost, and institution-specific r isk p olicies are unavailable in the public datasets, TotalCost(τ) should not be interpreted as an absolute monetary loss. Instead, it is used as a relative decision-policy indicator under a fixed c ost a ssumption. T he cost-optimized threshold is therefore illustrative and would require institution-specific c alibration o f C_{FP} b efore operational deployment. The amount-aware component is introduced at the decision-policy stage rather than the model-training stage. For the optimized XGBoost model, three decision policies are compared: $\tau = 0.5$, the validation $F1$-optimized threshold, and the validation cost-optimized threshold.

Explainability Analysis To reduce black-box behavior, Tree SHAP explanations are computed for the best-performing XGBoost models [21]. The 2023 explanation is used only as a balanced-reference setting, while the 2013 explanation is used as the primary severe-imbalance setting. The analysis reports global feature contribution rankings and a local explanation for one detected fraudulent transaction in each setting. Since the datasets use anonymized transformed variables, SHAP values are interpreted as evidence over latent dimensions rather than direct financial b usiness rules. Due to anonymization, these explanations cannot be mapped to original business variables such as merchant type, transaction channel, customer profile, or location; therefore, they support model auditability and internal consistency rather than full regulatory explainability.

3.4. Evaluation Metrics

Given the extreme imbalance inherent in fraud detection, evaluation prioritizes minority-sensitive metrics.

AUROC measures ranking performance across all thresholds. However, under severe imbalance, it may

overestimate performance due to dominance of true negatives.

PR-AUC evaluates the trade-off between Precision and Recall and is more informative under skewed distributions.

Precision is defined as

$$\text{Precision} = \frac{TP}{TP + FP},$$

while Recall is defined as

$$\text{Recall} = \frac{TP}{TP + FN}.$$

The primary classification summary metric is

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

which balances precision and recall but does not encode asymmetric financial loss.

A calibrated threshold τ is selected to maximize $F1$ on validation data for the main classification results. False Positives (FP) and False Negatives (FN) are explicitly monitored to assess operational risk. In the revised cost analysis, thresholds are also selected by minimizing TotalCost(τ) so that false negatives can be weighted by transaction amount.

4. Results

4.1. Performance on the Imbalanced 2013 Dataset

The 2013 dataset represents an extreme rarity regime (0.17% fraud prevalence), where model behavior is governed primarily by minority-class discrimination rather than aggregate accuracy. Under such skew, PR-AUC and $F1$ provide a more faithful assessment of practical utility than AUROC. The results in Table 2 reveal a clear performance gap between baseline and optimized boosting configurations. After hyperparameter tuning and threshold calibration, XGBoost improves PR-AUC from 0.7809 to 0.8815 and increases $F1$ from 0.7919 to 0.8497. Notably, both false positives and false negatives decline simultaneously (21/20 to 13/16), indicating that the improvement reflects enhanced score discrimination rather than a simple threshold-induced precision shift. The neural models exhibit a different behavior pattern. Although the baseline MLP achieves strong AUROC (0.9835), its PR-AUC remains constrained under severe skew. Incorporating BCE+Focal loss increases precision and reduces false positives (16 to 9) while leaving false negatives unchanged. This asymmetry suggests that the gain primarily arises from a stricter operating threshold rather than a substantive improvement in ranking quality. Because focal loss changes the distribution of neural-network scores, this high threshold should be interpreted as an operating point selected on validation data rather than as a calibrated probability cutoff.

The distinction between ranking enhancement and decision calibration becomes more evident when examining the PR and ROC curves (Figures 1 and 2). The optimized XGBoost curve shifts upward across a broad recall range, particularly beyond recall 0.7, demonstrating improved ordering of fraudulent instances throughout the score spectrum. In contrast, the MLP variants display localized precision improvements at selected operating points rather than consistent global displacement. The graphical evidence therefore supports the interpretation derived from Table 2: under extreme imbalance, meaningful gains require improved score separation rather than merely conservative calibration.

Error decomposition in Figure 3 provides additional insight into the source of these improvements. For XGBoost, hyperparameter optimization combined with calibrated thresholding reduces both false positives (21 to 13) and false negatives (20 to 16), thereby increasing correctly detected fraud cases from 78 to 82. The simultaneous decline in both error types indicates enhanced score discrimination rather than a conservative shift of the decision boundary. In contrast, the neural enhancement follows a different pattern. While the baseline MLP records 16 false positives and 20 false negatives, the BCE+Focal variant reduces false positives to 9 without changing the number of missed fraud cases. This asymmetric improvement suggests that the gain is driven primarily by stricter threshold calibration ($\tau = 0.997$) rather than improved minority-class separability. Taken together, the confusion matrices align with the Precision–Recall analysis: under extreme skew, the optimized boosting model achieves balanced error reduction through improved ranking quality, whereas the neural enhancement mainly refines precision via decision-level adjustment.

4.2. Performance on the Balanced 2023 Dataset

The balanced 2023 dataset presents a fundamentally different operating regime. Here, class prevalence no longer dominates evaluation, and ranking performance approaches saturation. As shown in Table 3, PR-AUC and AUROC values are nearly unity across all configurations, and recall remains perfect in every case. Consequently, model comparison is determined almost entirely by the control of false positives. Under this regime, threshold calibration becomes the primary mechanism of improvement. The tuned XGBoost model lowers false positives from 38 to 23 while maintaining full recall, yielding a marginal but consistent increase in $F1$. The enhanced MLP reduces false positives even further (81 to 15), achieving the highest $F1$ among all configurations. Unlike the 2013 scenario, these gains

Table 2. Test-set performance on the 2013 dataset (0.17% fraud prevalence)

Model	PR-AUC	AUROC	Precision	Recall	$F1$	τ	FP/FN
XGB (base)	0.7809	0.9262	0.7879	0.7959	0.7919	0.50	21/20
XGB (Optuna+τ)	0.8815	0.9799	0.8632	0.8367	0.8497	0.58	13/16
MLP (base)	0.7798	0.9835	0.8298	0.7959	0.8125	0.50	16/20
MLP (BCE+Focal+τ)	0.8630	0.9757	0.8966	0.7959	0.8432	0.997	9/20

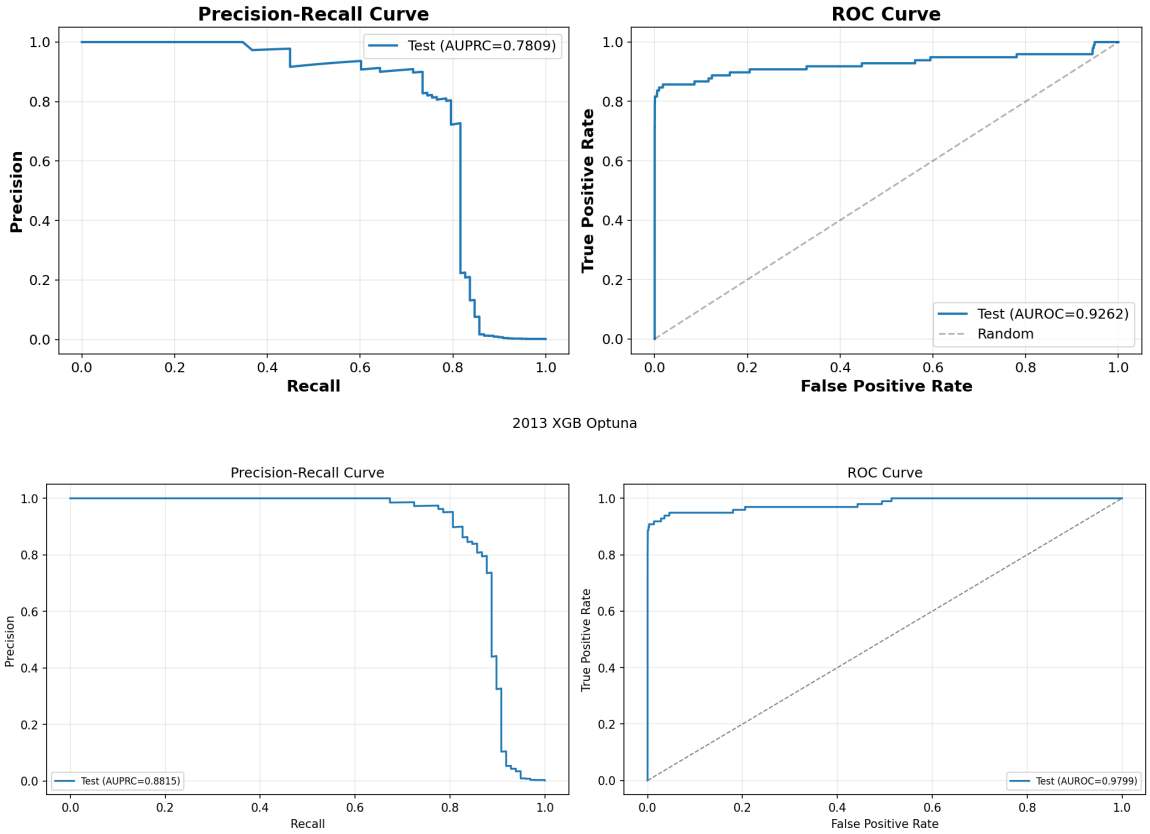

Figure 1. Precision–Recall and ROC curves for XGBoost on the 2013 dataset (0.17% fraud prevalence). The optimized model exhibits a clear upward shift in the high-recall region, reflecting improved minority-class ranking quality

Table 3. Test-set performance on the 2023 dataset (class-balanced)

Model	PR-AUC	AUROC	Precision	Recall	$F1$	τ	FP/FN
XGB (base)	0.999962	0.999974	0.999332	1.000000	0.999666	0.50	38/0
XGB (Optuna+τ)	0.999977	0.999984	0.999596	1.000000	0.999798	0.78	23/0
MLP (base)	0.999910	0.999921	0.998578	1.000000	0.999288	0.50	81/0
MLP (BCE+Focal+τ)	0.999970	0.999982	0.999736	1.000000	0.999868	0.924	15/0

do not reflect improved ranking structure but rather refined operating thresholds.

The near-saturated nature of ranking performance is visually evident in Figures 4 and 5. All curves cluster tightly near the upper boundary of the PR space, with minimal separation across recall levels. Incremental upward shifts are observable for the tuned variants, yet these differences remain subtle and do not indicate structural changes in score ordering. In this balanced context, separability is already strong; calibration primarily determines operational precision.

The confusion matrices in Figure 6 reinforce this interpretation. All models detect every fraudulent instance, and performance differences are attributable solely to the number of legitimate transactions incorrectly flagged. The optimized configurations achieve fewer false alarms without sacrificing sensitivity. Thus, unlike the imbalanced dataset where ranking quality drives improvement, the balanced regime highlights the

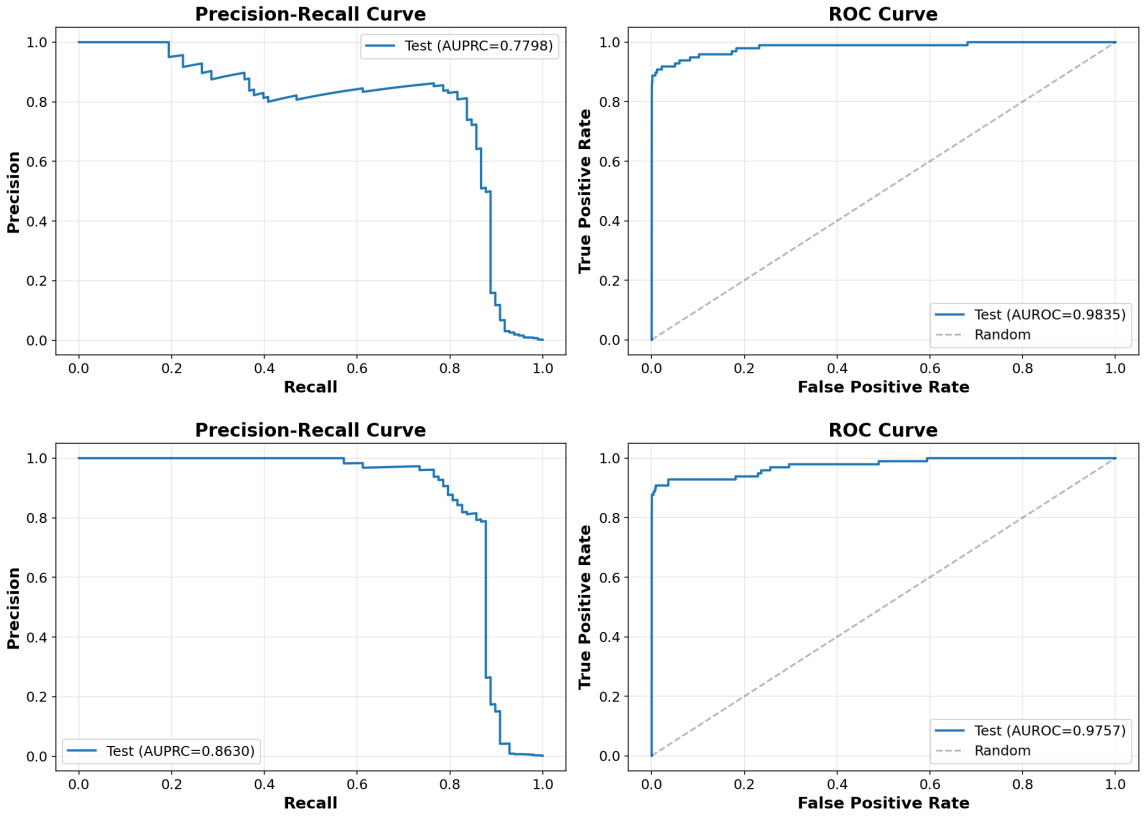

Figure 2. Precision–Recall and ROC curves for MLP variants on the 2013 dataset (0.17% fraud prevalence)

practical importance of threshold selection for managing investigation burden.

4.3. Amount-Aware Cost Analysis on the 2013 Dataset

Table 4 evaluates the same Optuna-tuned 2013 XGBoost model under three threshold policies. The validation $F1$-optimized threshold achieves the highest $F1$ score, whereas the cost-optimized threshold minimizes the illustrative amount-aware objective under the fixed assumption $C_{FP} = 1$. The cost-optimized threshold increases recall from 0.8367 to 0.9082 and reduces missed fraud cases from 16 to 9, lowering the simulated cost from 1969.61 to 1656.71. However, this reduction is obtained by increasing false positives from 13 to 363.

The low precision of the cost-optimized policy is a direct consequence of the very low decision threshold. At $\tau = 0.004$, the model flags a much larger number of transactions as fraudulent. This aggressive operating point increases recall and reduces amount-weighted missed fraud, but it also substantially increases false positives. In the random stratified test split, the cost-optimized policy produces 89 true positives and 363 false positives, yielding the Precision of 0.1969. Thus, the low precision does not indicate a deterioration in ranking performance; rather, it reflects a deliberately aggressive decision policy selected by the illustrative cost objective under $C_{FP} = 1$.

Therefore, this result should be interpreted as evidence that $F1$-optimal and cost-optimal thresholds can differ under a specified cost assumption, not as evidence that the cost-optimized threshold is universally preferable in practice. In real fraud monitoring systems, banks would select τ subject to review capacity, acceptable customer friction, regulatory requirements, and institution-specific risk appetite.

4.4. Chronological Validation on the 2013 Dataset

To evaluate deployment realism, the 2013 dataset is additionally split chronologically using the `Time` variable. The first 187,972 transactions are used for training, the next 39,873 for validation, and the final 56,962 for testing. The corresponding fraud counts are 368, 49, and 75, respectively. As shown in Table 5, the chronological split yields lower PR-AUC than the random stratified split for the optimized XGBoost model (0.7902 versus 0.8815). This result suggests that random stratification may provide optimistic estimates for this benchmark dataset when earlier and later transactions are mixed across training, validation, and test partitions. However, because the chronological

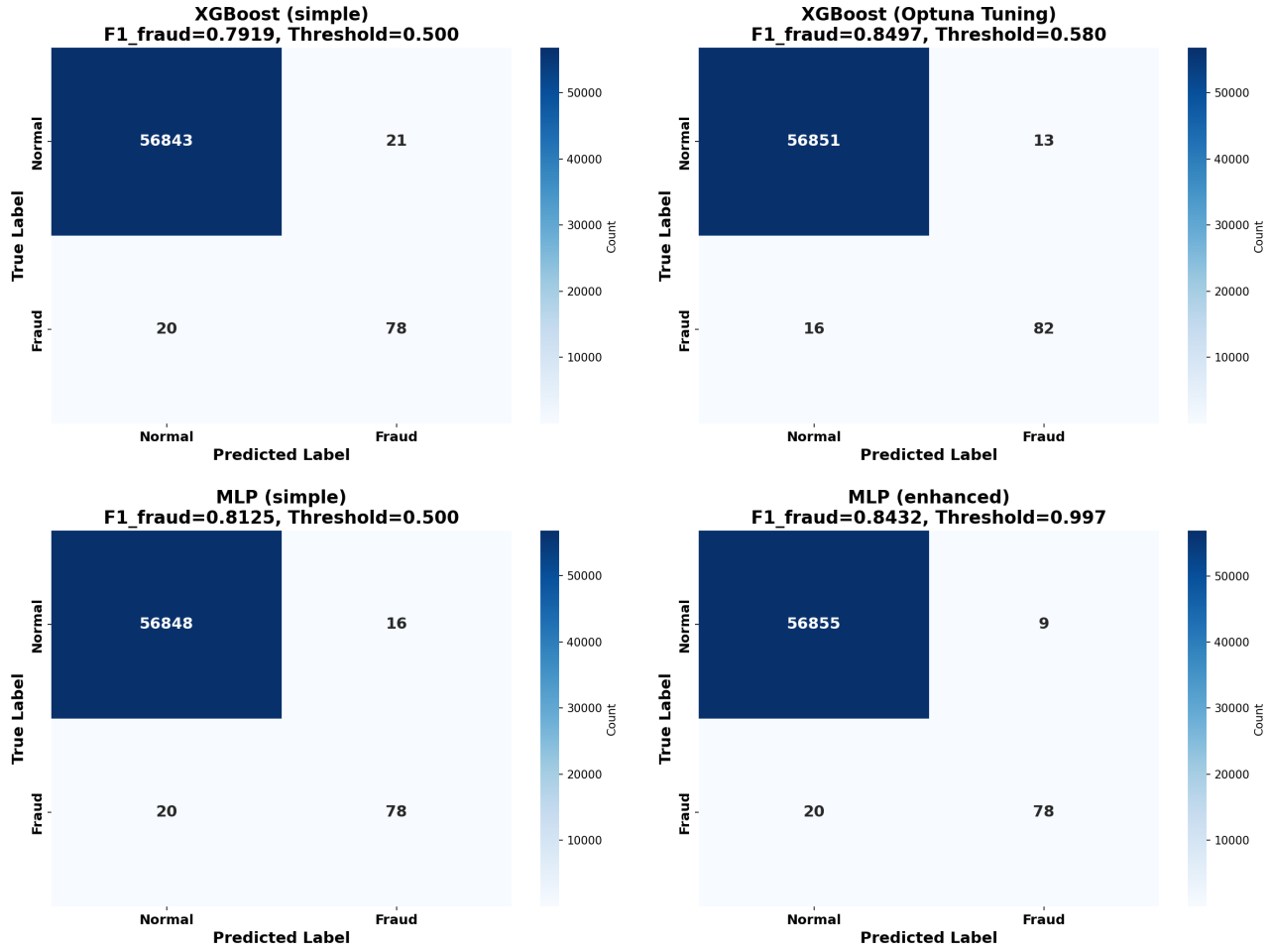

Figure 3. Confusion matrices on the 2013 dataset (0.17% fraud prevalence)

Table 4. Amount-aware cost evaluation on the 2013 random stratified test split. All rows use the same Optuna-tuned XGBoost model; only the decision threshold differs. The $\tau = 0.5$ row is therefore not equivalent to the untuned XGB (base) row in Table 2

Strategy	τ	Precision	Recall	$F1$	PR-AUC	FP/FN	Total Cost
Optimized XGBoost @ $\tau = 0.5$	0.500	0.8400	0.8571	0.8485	0.8815	16/14	1847.72
Optimized XGBoost @ $F1$-optimal τ	0.580	0.8632	0.8367	0.8497	0.8815	13/16	1969.61
Optimized XGBoost @ cost-optimal τ	0.004	0.1969	0.9082	0.3236	0.8815	363/9	1656.71

experiment uses a single short observation window, it should be interpreted as a limited near-future validation probe rather than as conclusive evidence of long-term temporal robustness.

The time-based results also show that threshold objectives affect operational behavior. The $F1$-optimized threshold achieves the highest temporal $F1$ score (0.8085) and sharply reduces false positives, while the cost-optimized threshold yields the lowest simulated financial cost by increasing recall to 0.8800. This trade-off is consistent with real fraud monitoring systems, where banks may prefer lower thresholds when the expected loss from missed fraud exceeds the cost of additional manual review. The 2023 dataset is not evaluated chronologically because it lacks an explicit transaction-time variable; its `id` column is therefore not assumed to represent temporal order.

4.5. Explainability Analysis

Tree SHAP explanations are computed for the optimized XGBoost models to examine which transformed features contribute most strongly to model decisions. Because the datasets use anonymized or engineered variables, the analysis is interpreted as transformed-feature auditability rather than direct business-level explainability.

Figure 7 presents the SHAP analysis for the optimized 2013 XGBoost model under severe class

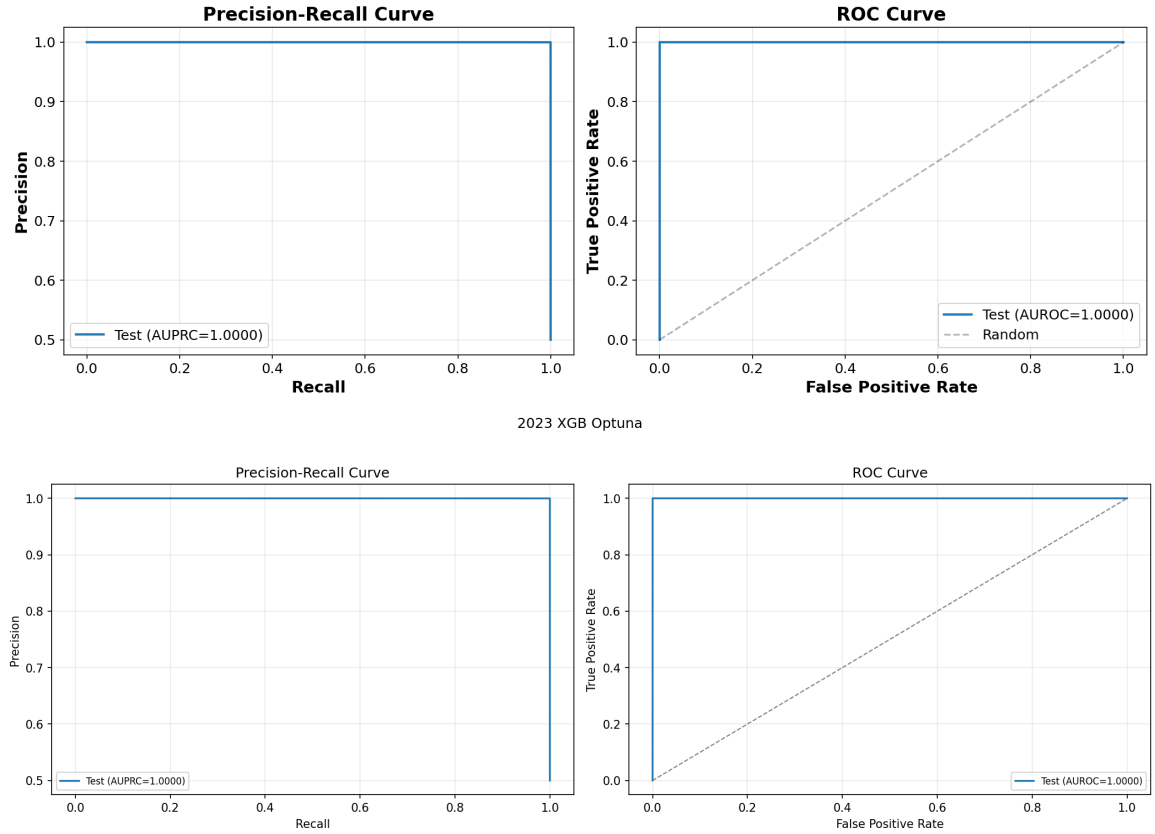

Figure 4. Precision–Recall and ROC curves for XGBoost on the 2023 dataset (balanced)

Table 5. Chronological validation on the 2013 dataset using transaction time. All rows use the same Optuna-tuned XGBoost model; only the decision threshold differs. The chronological split is used as a short-horizon near-future validation probe rather than a full concept-drift experiment

Strategy	τ	Precision	Recall	$F1$	PR-AUC	FP/FN	Total Cost
Optimized XGBoost @ $\tau = 0.5$	0.500	0.6316	0.8000	0.7059	0.7902	35/15	2432.79
Optimized XGBoost @ $F1$-optimal τ	0.726	0.8636	0.7600	0.8085	0.7902	9/18	2647.27
Optimized XGBoost @ cost-optimal τ	0.072	0.1038	0.8800	0.1857	0.7902	570/9	2285.59

imbalance. The global explanation identifies $V14$, $V4$, $V10$, $V12$, and $V8$ as the strongest transformed-feature contributors within this dataset. A local explanation for one detected fraudulent transaction shows that several transformed dimensions push the prediction toward fraud, while others reduce the fraud score. These explanations help inspect whether the model relies on a small set of influential latent dimensions, but they cannot be mapped to original financial variables because the dataset is anonymized.

Figure 8 provides a companion SHAP analysis for the optimized 2023 XGBoost model in the balanced-reference setting. The global explanation identifies $V14$, $V4$, $V12$, $V3$, and $V7$ among the leading contributors. Although some variable indices overlap with those observed in the 2013 analysis, the two datasets are independently anonymized or engineered. Therefore, index-level overlap should be treated as a descriptive observation rather than evidence that the same underlying business variables drive predictions across datasets.

4.6. Cross-Regime Interpretation

Across the two datasets, two distinct performance mechanisms become apparent, reflecting the influence of class prevalence on model behavior and evaluation outcomes. Under extreme imbalance (2013), improvement is fundamentally tied to ranking quality. In this regime, the rarity of fraudulent transactions amplifies the importance of accurately ordering minority instances in score space. The optimized XGBoost model demonstrates this effect through substantial PR-AUC

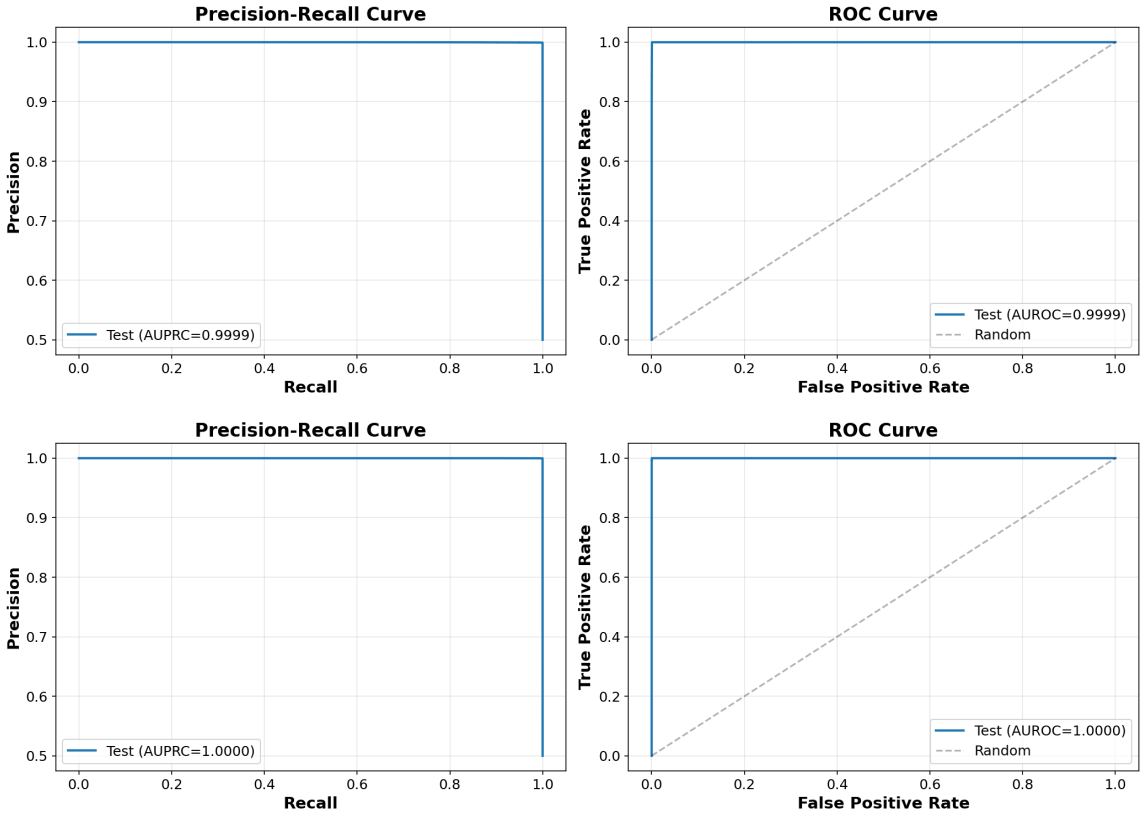

Figure 5. Precision–Recall and ROC curves for MLP variants on the 2023 dataset (balanced)

gains and simultaneous reductions in both false positives and false negatives. Such balanced error reduction indicates that the model does not merely adjust its operating threshold, but rather achieves improved separability between fraudulent and legitimate transactions across the score distribution. By contrast, the balanced 2023 dataset exhibits a different structural dynamic. Here, ranking performance approaches saturation across all configurations, as reflected by near-perfect PR-AUC and AUROC values. Because separability is already strong, additional gains cannot meaningfully arise from improved score ordering. Instead, improvements primarily reflect threshold-level control of false alarms.

5. Discussion

On the 2013 dataset, hyperparameter tuning combined with imbalance-aware weighting yields substantial improvements in PR-AUC and $F1$. The simultaneous reduction in both false positives and false negatives indicates that performance gains originate from improved ranking quality in score space rather than from conservative threshold shifts alone. This behavior suggests that cost-sensitive boosting enhances minority-instance separability under extreme rarity.

Across both datasets, decision-threshold selection consistently outperform the default cutoff $\tau = 0.5$ when evaluated by classification metrics. Under severe imbalance, threshold optimization aligns model outputs with minority-oriented objectives, particularly maximizing $F1$. Under balanced conditions, calibration primarily reduces false positives while maintaining full recall. The amount-aware analysis illustrates that $F1$-optimal and cost-optimal thresholds need not coincide once false negatives are weighted by transaction amount. Under the illustrative assumption $C_{FP} = 1$, the cost-optimized threshold attains a lower simulated cost but substantially increases false positives. The magnitude of this trade-off depends directly on the assumed review cost, and a higher C_{FP} would move the cost-optimal operating point toward a more conservative region. Therefore, the cost analysis should be interpreted as evidence that threshold selection is an explicit policy choice, not as evidence that the reported cost-optimal threshold is universally preferable.

The framework demonstrates interpretable behavior across two benchmark prevalence regimes, although these results should not be interpreted as direct evidence of deployment readiness. Under extreme skew,

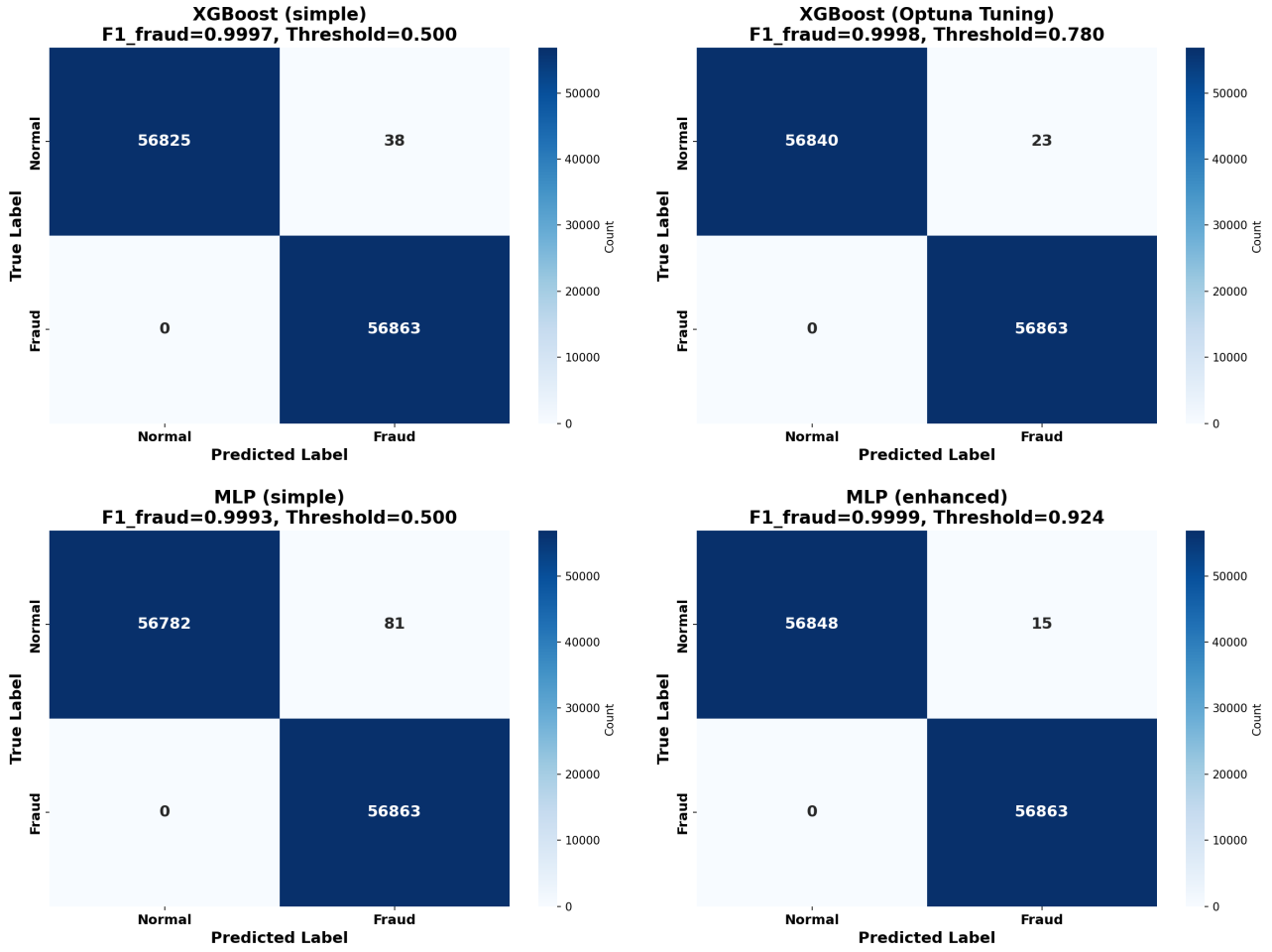

Figure 6. Confusion matrices on the 2023 dataset (balanced)

improvements require both enhanced ranking and calibrated decision boundaries. Under balanced conditions, ranking performance approaches saturation, and threshold calibration becomes the primary mechanism for refining operational trade-offs. This cross-regime consistency indicates that preserving the original data distribution while applying cost-sensitive learning and decision-level adjustment offers a reproducible alternative to synthetic oversampling.

In practical fraud detection systems, institutional objectives frequently require explicit control over the trade-off between investigation burden and missed fraud. The proposed pipeline enables such control without modifying the underlying training distribution. This separation reduces reliance on heuristic data augmentation and provides a clearer transition from score-based ranking to decision-policy selection. Nevertheless, real deployment would require institution-specific cost calibration, prospective validation, monitoring of temporal drift, and assessment of review capacity. The SHAP analyses support technical model auditing by identifying transformed dimensions that most strongly influence model decisions in both balanced and imbalanced settings. However, because the public datasets use anonymized or engineered variables, these explanations do not provide direct business-level interpretations such as merchant category, transaction channel, customer profile, or geographic location. The 2023 explanation provides a reference point, while the 2013 explanation highlights the features driving decisions under severe rarity. Because the variables are anonymized, these explanations should be used for model audit and cross-regime consistency checks rather than direct financial rule extraction.

Several limitations should be acknowledged. First, the evaluation relies on publicly available benchmark datasets with anonymized or engineered features, which limits direct interpretation of original banking variables. Second, the amount-aware cost analysis is based on an illustrative cost assumption

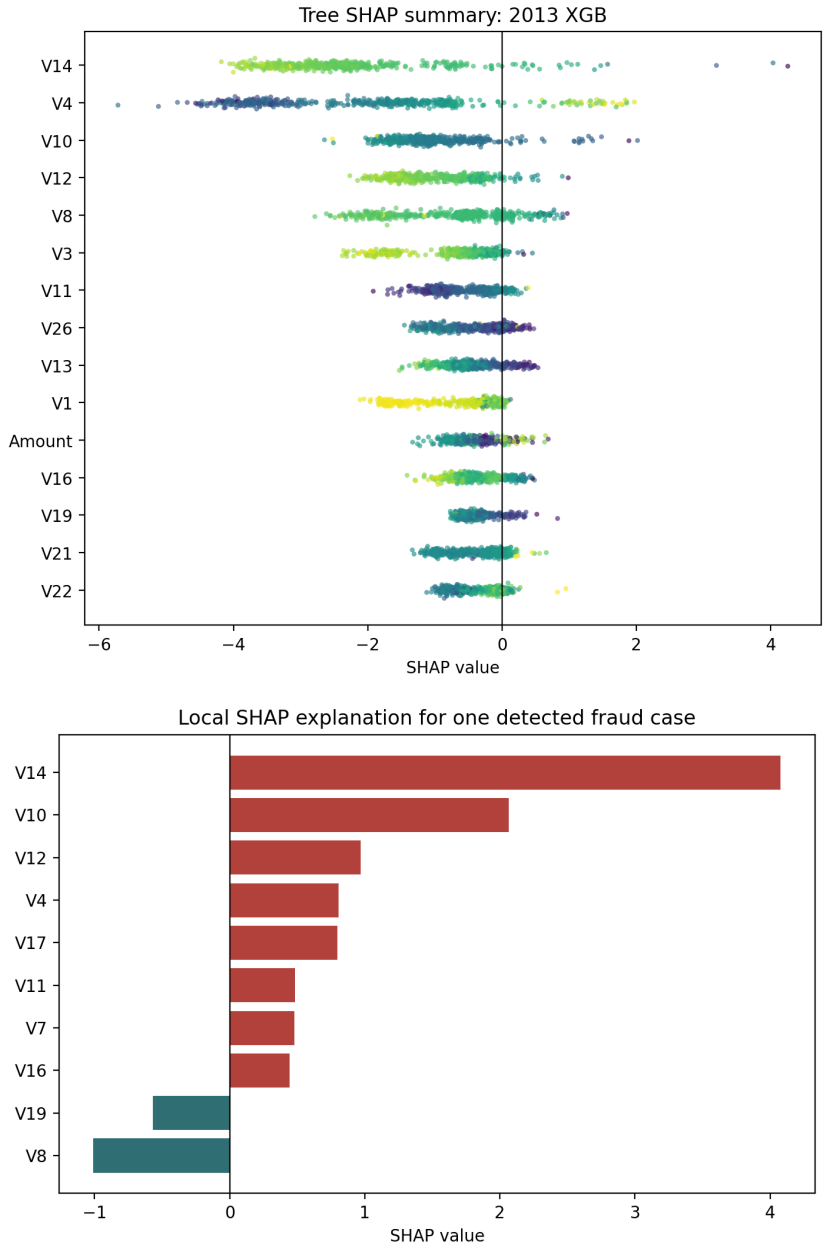

Figure 7. Tree SHAP explanations for the optimized 2013 XGBoost model. Left: global contribution distribution over sampled test transactions. Right: local contribution directions for one detected fraud case

because institution-specific review costs, customer-friction costs, chargeback recovery rates, reimbursement policies, and fraud-loss models are unavailable. Consequently, the reported cost-optimal thresholds should be interpreted as relative policy indicators under fixed assumptions rather than deployment recommendations. Third, the chronological validation is conducted only on the 2013 dataset and covers an approximately 48-hour observation window; therefore, it serves as a short-horizon validation probe rather than a full concept-drift experiment. Fourth, the reported improvements on the 2013 test split are descriptive comparisons based on a single held-out partition containing 98 fraud cases. Formal uncertainty analysis, such as bootstrap confidence intervals or paired significance testing, was not conducted in this revision. Finally, SHAP explanations are limited to transformed-feature auditability and cannot be mapped directly to business-level fraud rules.

Future research should evaluate the framework on longer-horizon institutional transaction streams, incorporate calibrated economic cost matrices into both

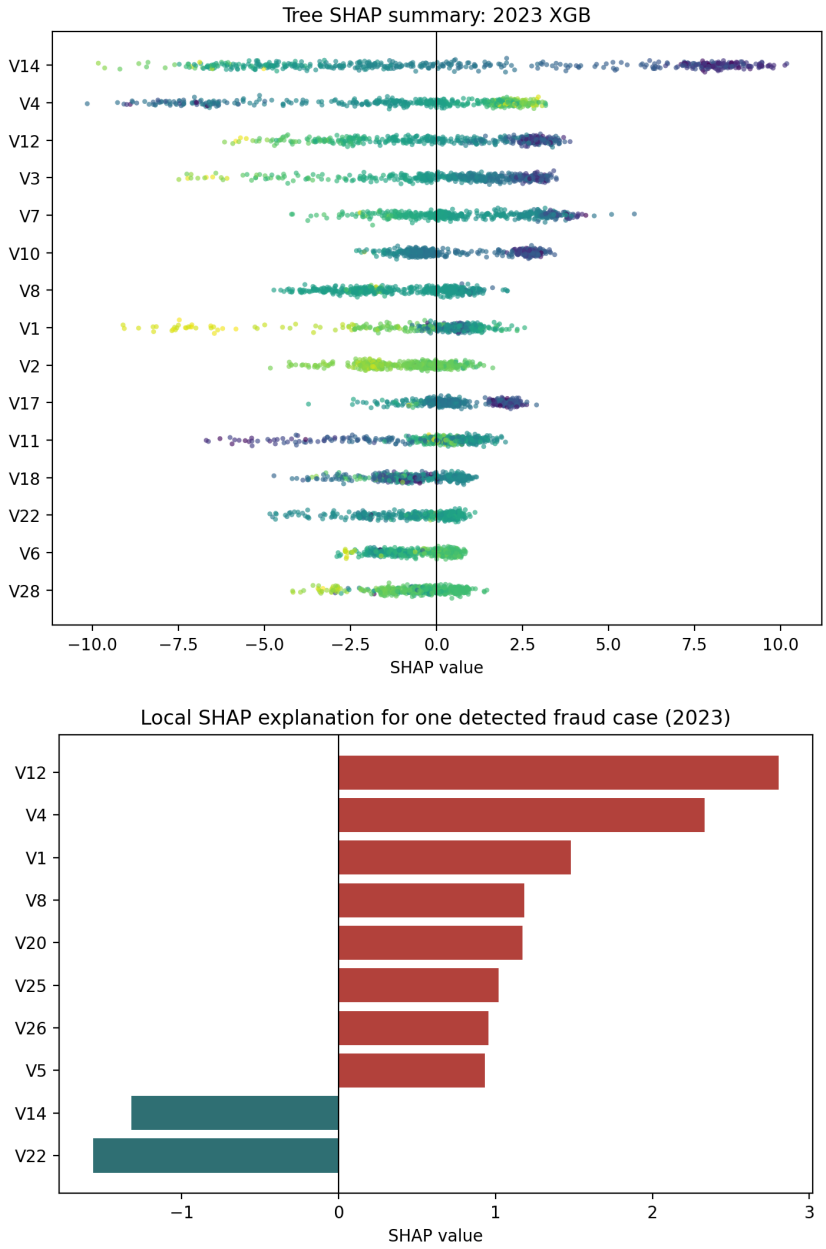

Figure 8. Tree SHAP explanations for the optimized 2023 XGBoost model used as a balanced-reference setting. Left: global contribution distribution over sampled test transactions. Right: local contribution directions for one detected fraud case

training and threshold optimization, and examine robustness under concept drift and prior-probability shift. Additional work should also include uncertainty estimation for benchmark comparisons, sensitivity analysis over false-positive review costs, and probability-calibration assessment when model scores are used for risk estimation rather than ranking alone. Extending the framework to real-time streaming environments would further assess the robustness of separating ranking optimization from threshold calibration under dynamic operational constraints.

6. Conclusion

This study presents a cost-sensitive and decision-threshold-calibrated framework for credit card fraud detection that operates directly on the original data distribution without synthetic oversampling. By integrating imbalance-aware model training with systematic threshold selection, the proposed approach separates score-based ranking from operational decision policy, enabling flexible adaptation across heterogeneous class distributions.

Empirical evaluation on both extreme-imbalance (0.17% prevalence) and balanced datasets demonstrates that performance improvements arise through different mechanisms depending on class prevalence. While severe rarity requires enhanced ranking discrimination, balanced regimes benefit p r imarily f r om calibrated decision thresholds that control false alarms. The revised amount-aware cost analysis further shows that $F1$-optimal and cost-optimal operating points can differ substantially, emphasizing the importance of aligning threshold policy with financial risk.

Rather than modifying class proportions through data-level augmentation, the framework emphasizes model-level cost sensitivity and decision-level calibration as complementary components of fraud detection system design. SHAP analyses across the balanced-reference and severe-imbalance settings improve transformed-feature auditability, while chronological validation on the 2013 dataset suggests that random stratified s p lits m a y p r oduce optimistic estimates in this benchmark setting.

Future work may extend the framework to streaming environments, incorporate explicit economic cost matrices, and evaluate robustness under longer-term temporal drift and prior probability shift. Overall, the study suggests that fraud detection under class imbalance should be evaluated not only through ranking metrics, but also through decision-threshold policy, transformed-feature auditability, and temporally aware validation. Further validation on institution-specific and longer-horizon data is required before drawing strong conclusions about operational robustness.

Declaration on the Use of Generative AI Tools

In accordance with the EAI Artificial Intelligence Policy, the author declares that generative AI tools were used for manuscript preparation support, including language refinement and clarity improvement. All AI-assisted content was reviewed and approved by the author, who remains fully responsible for the manuscript.

Acknowledgements This work is a part of the research project (CS.2025.B2.028) funded by Saigon University.

References

[1] K. E. Hachimi, G. Orhanou, and L. Ballihi, "Credit card fraud detection model based on optimized long short-term memory: Imbalanced class solution and overfitting reduction," *Expert Systems with Applications*, p. 130871, 2025.

[2] M. RezvaniNejad and A. S. Yameqani, "Wdae-gan: A hybrid dual autoencoder and generative adversarial framework with wavelet denoising for credit card fraud detection," *Expert Systems with Applications*, p. 130078, 2025.

[3] R. K. Gupta, A. Hassan, S. K. Majhi, N. Parveen, A. T. Zamani, R. Anitha, B. Ojha, A. K. Singh, and D. Muduli, "Enhanced framework for credit card fraud detection using robust feature selection and a stacking ensemble model approach," *Results in Engineering*, vol. 26, p. 105084, 2025.

[4] G. Metzler, X. Badiche, B. Belkasmi, E. Fromont, A. Habrard, and M. Sebban, "Tree-based cost sensitive methods for fraud detection in imbalanced data," in *International Symposium on Intelligent Data Analysis*, pp. 213–224, Springer, 2018.

[5] W. Liu, H. Fan, M. Xia, and M. Xia, "A focal-aware cost-sensitive boosted tree for imbalanced credit scoring," *Expert Systems with Applications*, vol. 208, p. 118158, 2022.

[6] J. L. Leevy, J. M. Johnson, J. Hancock, and T. M. Khoshgoftaar, "Threshold optimization and random undersampling for imbalanced credit card data," *Journal of Big Data*, vol. 10, no. 1, p. 58, 2023.

[7] R. Ango, R. K. Masih, C. K. K. Reddy, M. Shuaib, M. Singh, and S. Alam, "Fraud detection in banking using the kaggle credit card dataset and xgboost model," in *2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS)*, pp. 968–973, IEEE, 2024.

[8] H. Huang, B. Liu, X. Xue, J. Cao, and X. Chen, "Imbalanced credit card fraud detection data: A solution based on hybrid neural network and clustering-based undersampling technique," *Applied Soft Computing*, vol. 154, p. 111368, 2024.

[9] M. Herland, R. A. Bauder, and T. M. Khoshgoftaar, "The effects of class rarity on the evaluation of supervised healthcare fraud detection models," *Journal of Big Data*, vol. 6, no. 1, p. 21, 2019.

[10] R. A. Bauder, T. M. Khoshgoftaar, and T. Hasanin, "An empirical study on class rarity in big data," in *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pp. 785–790, IEEE, 2018.

[11] J. S. Sundar, R. Abinaya, M. Venkatesh, Y. Suthari, H. Koganti, and S. K. R. Katta, "Machine learning algorithms for real-time fraud detection in digital banking transactions," in *2025 3rd International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, pp. 787–792, IEEE, 2025.

[12] C. Elkan, "The foundations of cost-sensitive learning," in *International joint conference on artificial intelligence*, vol. 17, pp. 973–978, Lawrence Erlbaum Associates Ltd, 2001.

[13] S. Elshaar and S. Sadaoui, "Cost-sensitive semi-supervised classification for fraud applications," in *International Conference on Agents and Artificial Intelligence*, pp. 173–187, Springer, 2020.

[14] L. Huang, A. Abrahams, and P. Ractham, "Enhanced financial fraud detection using cost-sensitive cascade forest with missing value imputation," *Intelligent Systems in Accounting, Finance and Management*, vol. 29, no. 3, pp. 133–155, 2022.

[15] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, 2006.

[16] J. Singla *et al.*, "Comparing roc curve based thresholding methods in online transactions fraud detection system using deep learning," in *2021 international conference on computing, communication, and intelligent systems (ICCCIS)*, pp. 9–12, IEEE, 2021.
[17] L. A. Dalton, "Optimal roc-based classification and performance analysis under bayesian uncertainty models," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 4, pp. 719–729, 2015.
[18] C. Wang, S. Chai, and H. Zhu, "Opendrift: Online evolving fraud detection for open-category and concept-drift transactions," in *2023 IEEE International Conference on Web Services (ICWS)*, pp. 605–614, IEEE, 2023.
[19] M. L. G. ULB, "Credit card fraud detection: Anonymized credit card transactions labeled as fraudulent or genuine." Kaggle Data Repository, 2013. Accessed: 2026-02-23.
[20] K. Shenoy, "Credit card fraud detection dataset 2023: A comprehensive dataset for credit card fraud detection." Kaggle Data Repository, 2023. Accessed: 2026-02-23.
[21] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.