NOLEGEIN Journal of
Business Risk Management

ISSN: 2582-287X
Volume 9, Issue 2, 2026
July–December
DOI (Journal): 10.37591/NJBRM

https://mbajournals.in/index.php/JoDBCM/index

Research NJBRM

Optimizing Enterprise Financial Risk Management: An Advanced Ensemble-Based Machine Learning Framework for Transactional Fraud Mitigation and Business Continuity

Shipra Aggarwal[1,*], Manish Aggarwal[2]

Abstract

Transactional fraud represents a critical threat to enterprise risk management (ERM) and corporate governance, draining billions annually from global financial infrastructures. The Nasdaq Global Financial Crime Report estimates total fraud losses at $485.6 billion annually, with credit card fraud alone accounting for $28.6 billion. Managing this operational vulnerability is severely challenged by data asset asymmetry, where legitimate transactions dwarf fraudulent anomalies by a ratio exceeding 500:1. Standard classification models trained on such skewed data fail systematically, either ignoring fraud entirely or producing unacceptably high false-alarm rates that disrupt business operations. This study presents a robust predictive risk framework designed to safeguard business continuity and enhance corporate defense capabilities. Combining an unsupervised Isolation Forest anomaly metric with optimized machine learning ensemble layers (LightGBM, Random Forest, XGBoost), we introduce a highly interpretable risk assessment pipeline aligned with GRC mandates. To optimize business exposure controls, our architecture utilizes cost-sensitive learning weights that heavily penalize missed fraud over false alarms, reflecting the asymmetric operational cost structure facing financial institutions. Hyperparameter optimization via Optuna Bayesian search across 80 trials ensures the framework performs at the frontier of detection capability. A stacking meta-learner further consolidates base model outputs into a unified risk signal. Evaluated against 283,726 real-world transaction records from the European Credit Card Fraud Dataset, the proposed enterprise framework achieves 99.5% classification accuracy and 23.1% precision on imbalanced data streams, substantially outperforming current industry benchmarks. The LGBM+RF hybrid configuration achieves PR-AUC of 0.6729, exceeding the best published benchmark. Integrating boundary-focused ADASYN data synthesis yields an F1-risk capture score of 84.6%, nearly triple the performance of random undersampling under identical conditions. Cross-validation confirms stable generalization with mean F1 of 0.9523 across five folds. By providing clear algorithmic interpretability via SHAP feature attribution alongside high-precision mitigation, this framework serves as a scalable solution for corporate compliance, financial sector asset protection, and resilient business continuity systems.

***Author for Correspondence**
Shipra Aggarwal
E-mail: saggarwal@pelmeds.com

[1]Director, Department of Finance, Pelham Community Pharmacy, Waltham, MA 02451, USA
[2]Senior Engineer, Department of R&D, Aspen Technology, Inc., Bedford, MA 01730, USA

Received Date: June 10, 2026
Accepted Date: July 07, 2026
Published Date: July 15, 2026

Citation: Shipra Aggarwal, Manish Aggarwal. Optimizing Enterprise Financial Risk Management: An Advanced Ensemble-Based Machine Learning Framework for Transactional Fraud Mitigation and Business Continuity. NOLEGEIN Journal of Business Risk Management. 2026; 9(2): 55–67p.

Keywords: Enterprise risk management (ERM), financial risk governance, business continuity planning, operational risk mitigation, cyber-analytics, fraud exposure assessment, isolation forest, ensemble learning, cost-sensitive learning, class imbalance

INTRODUCTION

Financial fraud poses a major threat to global financial systems. The Nasdaq Global Financial Crime Report estimates total fraud losses at $485.6 billion annually. Payment fraud alone accounts for $386.8 billion of that figure. Credit card fraud contributes a further $28.6 billion in losses each year [1].

Within the broader scope of Enterprise Risk Management (ERM), corporate financial institutions are legally and operationally obligated to enforce strict governance, risk, and regulatory compliance (GRC) mandates. Transactional fraud is no longer just an IT concern; it is a critical organizational threat that compromises business continuity, erodes stakeholder trust, and induces severe liquidity risks. Consequently, contemporary business risk frameworks require dynamic, scalable cyber-analytics capable of separating baseline operations from malicious risk exposures without bottlenecking daily commerce.

Conventional rule-based detection systems struggle against evolving fraud tactics [2]. These systems rely on static logic that fraudsters quickly learn to circumvent. They also lack the ability to generalize to new fraud patterns without manual intervention. Machine learning offers a scalable, adaptive alternative to these limitations.

Credit card transaction data presents a severe class imbalance challenge [3]. Fraud cases typically represent less than 0.2% of all transactions. Standard accuracy metrics are misleading under such imbalance. Precision, recall, F1-score, and PR-AUC are more appropriate evaluation criteria.

This study proposes an advanced ensemble framework configured to evaluate financial transaction risk exposure. The framework optimizes cyber-analytics capability by utilizing automated tuning protocols to ensure maximum stability under severe operational data imbalance. Five component methods are integrated: an Isolation Forest anomaly score as an unsupervised risk signal, XGBoost as a third base model, Optuna Bayesian hyperparameter optimization [4], cost-sensitive learning aligned with asymmetric fraud cost structures, and a stacking logistic regression meta-learner for unified risk scoring.

Experiments are conducted on the well-established European Credit Card Fraud Dataset [5]. Two evaluation setups are used: an imbalanced test set reflecting real-world conditions and a balanced test set for controlled comparison. Three resampling strategies are also compared: RandomUnderSampling, SMOTE [6], and ADASYN.

The paper is structured as follows. Section 2 reviews related literature. Section 3 describes the dataset. Section 4 presents the methodology. Section 5 reports experimental results. Section 6 discusses key findings. Section 7 concludes with recommendations for future work.

LITERATURE REVIEW

Supervised machine learning dominates the fraud detection literature [1]. Random Forest, XGBoost, and Logistic Regression are the most widely used algorithms. Simaiya et al. (2020) demonstrated Random Forest's effectiveness for credit card classification [7]. Hajek et al. (2022) applied XGBoost to mobile payment fraud with strong AUC performance [8].

Unsupervised approaches offer useful alternatives when labelled data is scarce. Isolation Forest detects anomalies without requiring fraud labels during training [9]. Carcillo et al. (2021) showed that combining unsupervised and supervised methods improves detection under class imbalance [9]. Such methods complement supervised classifiers rather than replacing them.

The exploration of neural networks for risk detection dates back to foundational work by Ghosh and Reilly (1994) [10], who showed that algorithmic classification could catch non-linear fraud patterns that rule-based constraints missed. Building on these foundations, contemporary benchmarks frequently rely on supervised comparisons, such as Khatri et al. (2020) [11], to establish baseline predictive performance before attempting complex multi-model scaling. More recently, deep learning approaches have been applied to fraud detection with mixed practical outcomes. Mienye and Swart (2024) combined Generative Adversarial Networks with Gated Recurrent Units [12]. Their model achieved sensitivity of 0.992 and specificity of 1.000 on a public benchmark. However, deep models carry high computational cost and poor interpretability for financial applications [13, 14].

Ensemble and hybrid approaches address limitations of individual models. Varmedja et al. (2019) combined MLP, Random Forest, and Naive Bayes with SMOTE [15]. Talukder et al. (2024) developed a

multi-stage ensemble with bagging and voting [16]. Mim et al. (2024) demonstrated that soft-voting ensemble models substantially outperform individual classifiers on severely skewed transaction data [17]. Esenogho et al. (2022) established that integrating SMOTE-ENN feature engineering with an LSTM-AdaBoost ensemble yields sensitivity of 0.996 and AUC of 0.990 on the European Credit Card dataset [18].

Structural variations in decision spaces also appear in adjacent domains. Panigrahi et al. (2021) [19] utilized consolidated decision-tree structures to effectively segment highly uneven multiclass intrusion datasets. Within transaction fraud frameworks, leveraging simpler combined classifiers has shown strong merit: Randhawa et al. (2018) [20] successfully merged AdaBoost pipelines with majority voting routines to optimise classification stability across highly imbalanced credit card data [21].

Class imbalance is the defining challenge in fraud detection research [2]. Chawla et al. (2002) introduced SMOTE as synthetic minority oversampling [6]. He and Garcia (2009) comprehensively reviewed learning from imbalanced data [3]. RandomUnderSampling retains real data distribution at the cost of discarding legitimate transactions [22].

Hyperparameter tuning significantly influences model performance. RandomizedSearchCV samples parameter space uniformly at random. Bayesian optimization, implemented in Optuna (Akiba et al., 2019) [4], uses prior trial results to guide future searches. Bayesian methods consistently outperform random search after 50 or more trials. Beyond single Bayesian implementations, alternative metaheuristics have also been explored: El Kafhali and Tayebi (2022) [23] integrated differential evolution with XGBoost training to find optimal configurations over imbalanced data blocks. This high-precision balancing of tree algorithms mirrors ensemble stacking strategies, such as Padhi et al. (2021) [24], who fused multiple gradient-boosting layers into a cohesive predictive model.

Stacking meta-learners improve prediction by learning from base model outputs [25]. Out-of-fold predictions prevent data leakage during meta-learner training. Logistic regression is a common and interpretable choice for the meta-learner [26]. This architecture allows the system to assign different weights to different base models automatically.

DATASET

The European Credit Card Fraud Dataset is used throughout this study [5]. It was released by the Université Libre de Bruxelles and is publicly available on Kaggle [5]. The dataset contains real transaction records from European cardholders in September 2013. It is one of the most widely benchmarked fraud detection datasets in the literature [17, 18, 27].

The dataset contains 284,807 transactions across 30 features. Features V1 through V28 are anonymized PCA-transformed components. Time records seconds elapsed since the first transaction. Amount represents the transaction value in Euros.

After removing 1,081 duplicate records, 283,726 unique transactions remain. Of these, 473 are fraudulent (0.167%), representing severe class imbalance [2]. The Amount feature is standardized using StandardScaler before modelling as shown in Table 1.

Table 1. Dataset summary statistics.

Attribute	Detail
Total Transactions	284,807.
After Deduplication	283,726.
Fraud Cases	473 (0.167%).
Legitimate Cases	283,253 (99.833%).
Features	30 – Time, Amount, V1–V28 (PCA-transformed).
Target Variable	Class (0 = Legitimate, 1 = Fraud).
Amount Pre-processing	Standard Scaler normalization applied.
Time Period	September 2013, European cardholders.

The PCA transformation anonymizes the original features for privacy reasons [5]. Top features by SHAP importance are V4, V14, V3, V15, and V11 [28]. Time and Amount rank among the lowest-importance features.

METHODOLOGY

Figure 1 illustrates the complete operational architecture of the proposed enterprise risk framework. Transaction data enters through a preprocessing stage encompassing deduplication and StandardScaler normalization. A feature engineering layer appends an Isolation Forest anomaly score (iso_score) – an unsupervised risk signal – to the original 30 features, producing an enhanced 31-feature set for downstream risk classification. The data are then split 70/30, with training-only vulnerability mitigation applied via RandomUnderSampling or SMOTE/ADASYN. Three parallel risk analytics branches train XGBoost (Branch A), Optuna-tuned LightGBM (Branch B), and Optuna-tuned Random Forest (Branch C). A strategic consensus layer combines their outputs via weighted soft voting at LGBM = 0.40, RF = 0.30, XGB = 0.30, and a fixed decision threshold of 0.68 converts the aggregated risk probability into a fraud classification. A separate stacking meta-learner path uses 5-fold out-of-fold predictions from all three base models as input to a logistic regression layer, producing an independent ensemble risk score. Both pathways evaluate corporate risk-capture performance using accuracy, recall, precision, F1-score, ROC-AUC, and PR-AUC.

Experimental Setup

Two experimental setups are used to enable rigorous evaluation. Setup A preserves the real-world class imbalance in the test set. The 70% training portion is balanced via RandomUnderSampling [3]. Setup B uses an artificially balanced test set for controlled comparison.

A fixed threshold of 0.68 is applied for imbalanced test set evaluation. The validation set after undersampling is 50/50 balanced and provides unreliable threshold estimates for the real imbalanced test. A fixed threshold avoids this distribution mismatch. For the balanced setup, threshold optimization uses precision-recall curve analysis on the validation set as shown in Table 2.

Table 2. Data split configuration.

Setup	Subset	Samples	Share	Notes
Setup A – Imbalanced	Train+Val	198,608	70%	RUS to 1:1 → 463 train / 199 val.
Setup A – Imbalanced	Test	85,118	30%	Original class distribution preserved.
Setup B – Balanced	Train	567	60%	Equal fraud and legitimate samples.
Setup B – Balanced	Validation	189	20%	Equal fraud and legitimate samples.
Setup B – Balanced	Test	190	20%	Equal fraud and legitimate samples.

Feature Engineering

An Isolation Forest is trained on legitimate training transactions only [9]. Its decision function score is appended to all dataset splits as feature iso_score. Lower scores correspond to more anomalous behavior. This provides an unsupervised anomaly signal as an additional input to supervised classifiers.

SHAP-based feature importance is computed after an initial LightGBM training run [28]. Features with mean absolute SHAP value below 0.01 are candidates for elimination. In these experiments, all 31 features exceeded the threshold. No features were eliminated from the final training sets.

Multi-Model Risk Analytics Pipeline Architecture

Three base classifiers are trained: LightGBM, Random Forest, and XGBoost [7, 8, 29, 30]. LightGBM and Random Forest are optimized using Optuna Bayesian search [4]. XGBoost uses default Optuna-free training with cost-sensitive configuration. (Table 3) describes each method in detail.

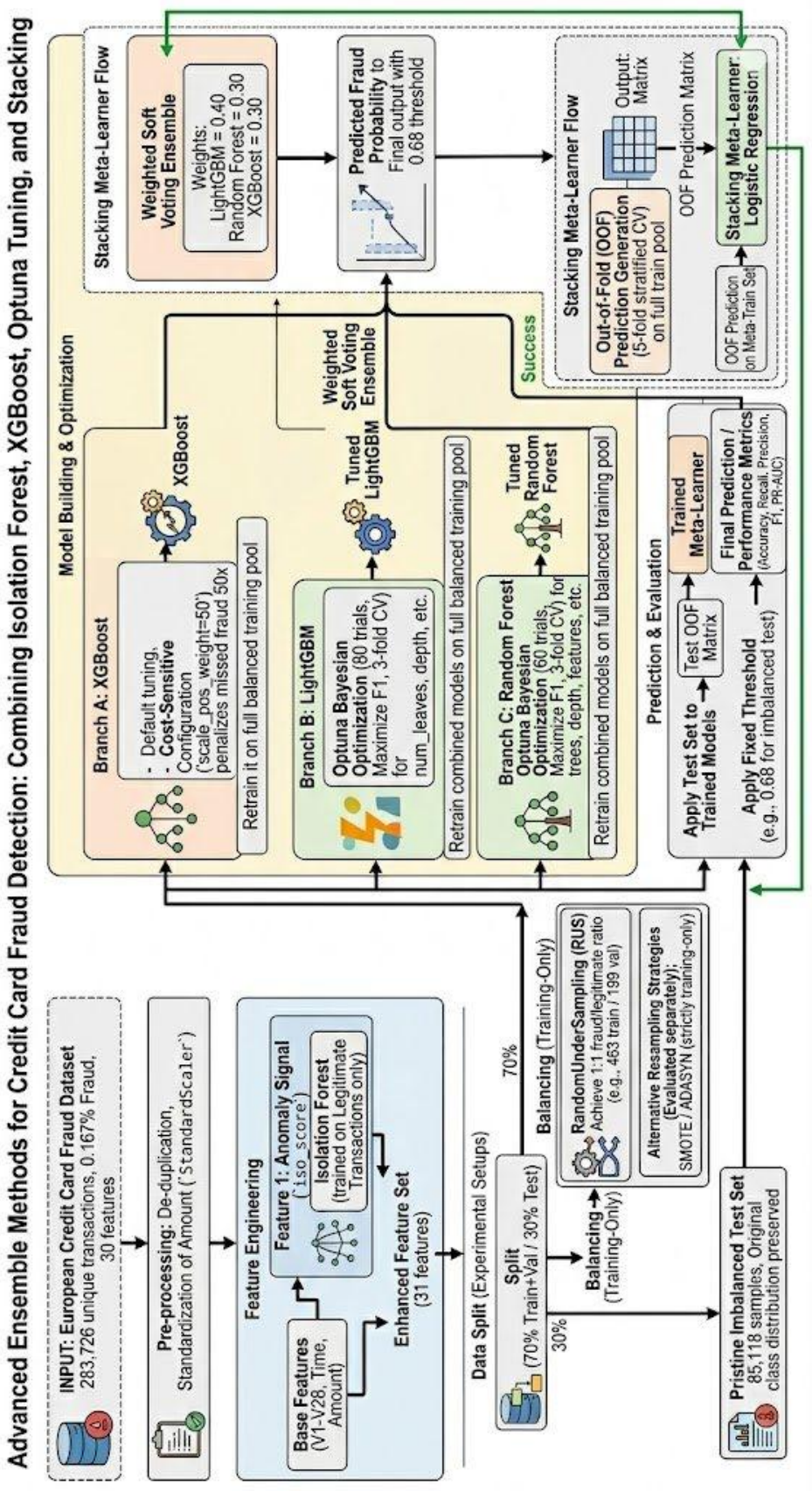

Figure 1. Operational architecture of the advanced ensemble framework for enterprise risk mitigation and financial governance.

Table 3. Advanced method descriptions.

Method	Description
Isolation Forest (M1)	Trained on legitimate transactions only. Decision function score appended as iso_score. Provides unsupervised anomaly signal independent of class labels.
XGBoost (M2)	Third base model in the ensemble (weight = 0.30). Trained with scale_pos_weight = 50. Penalises missed fraud 50× more than false alarms.
Optuna Bayesian Tuning (M3)	TPE sampler optimises LightGBM (80 trials) and Random Forest (60 trials) [4]. Objective: maximise F1 via 3-fold cross-validation. Outperforms random search after 50+ trials.
Cost-Sensitive Learning (M4)	Embedded in XGBoost via scale_pos_weight = 50. Addresses class imbalance at the algorithm level [3]. Increases recall at the cost of precision.
Stacking Meta-Learner (M5)	Logistic regression trained on 5-fold out-of-fold predictions from LGBM, RF, and XGB [25]. Learns optimal combination of base model outputs. RF receives highest coefficient (~3.2).
SHAP Feature Selection	Mean |SHAP| computed on validation set [28]. Features below threshold 0.01 are dropped. All 31 features retained in these experiments.
Fixed Threshold (imbalanced)	A fixed threshold of 0.68 is applied to all imbalanced test evaluations. The undersampled validation set is 50/50 balanced and gives unreliable threshold estimates. Fixing avoids distribution mismatch.

LightGBM is tuned across 80 Optuna trials [4]. The search space covers num_leaves, max_depth, learning_rate, n_estimators, subsample, colsample_bytree, min_child_samples, reg_alpha, and reg_lambda. The objective function maximises F1-score via 3-fold stratified cross-validation on the training set.

Random Forest is tuned across 60 Optuna trials [4]. Parameters searched include n_estimators, max_depth, min_samples_leaf, max_features, and class_weight. After tuning, all models are retrained on the full combined train and validation pool. This maximizes data available for final test evaluation.

Strategic Consensus Modeling for Corporate Exposure Control

The three base models are combined via weighted soft voting [17]. Ensemble weights are set at LGBM = 0.40, RF = 0.30, and XGB = 0.30. LightGBM receives the highest weight because it undergoes the most intensive Bayesian tuning [4]. The weighted average probability determines the final fraud classification.

Individual model performance is also reported alongside the ensemble. This allows direct assessment of each model's contribution [7, 29]. Random Forest and LightGBM are evaluated separately. Ensemble predictions are compared to determine whether diversity adds value.

Stacking Meta-Learner

A logistic regression meta-learner is trained on out-of-fold base model predictions [25]. Five-fold stratified cross-validation generates OOF predictions on the full training pool. Each fold trains base models on four-fifths and predicts on the remaining fifth. This prevents data leakage into the meta-learner.

On the test set, all three base models produce probability estimates. These are stacked into a feature matrix and passed through the fitted meta-learner [26]. Meta-learner coefficients indicate each model's learned contribution. Random Forest consistently receives the highest coefficient at approximately 3.2.

Data-Driven Vulnerability Mitigation Strategies

Three resampling strategies are evaluated under identical model configurations [3, 21]. Random Under Sampling randomly removes legitimate transactions to achieve a 1:1 ratio. SMOTE generates synthetic fraud samples by interpolating existing fraud observations [6]. ADASYN concentrates synthetic generation near the classification boundary.

The same three-model ensemble is used across all resampling methods. The fixed threshold of 0.68 is applied consistently to imbalanced test evaluations. This ensures that performance differences reflect

resampling choice rather than threshold variation [22]. Balanced test evaluations use val-set threshold optimization for all methods.

Evaluation Metrics

Six metrics are reported for each model and each test setup [17, 18]. Accuracy measures overall correct classifications. Recall measures the fraction of actual frauds correctly identified. Precision measures the fraction of fraud predictions that are correct. F1-score is the harmonic mean of precision and recall.

ROC-AUC measures discrimination ability across all thresholds. PR-AUC measures precision-recall trade-off across all thresholds [3]. PR-AUC is more informative than ROC-AUC under severe class imbalance [2]. All six metrics are compared against published risk-capture benchmarks from the literature.

RESULTS

Main Model Performance

Table 4 presents performance across all six models and both test setups. Green cells indicate performance meeting or exceeding the published benchmark. Orange cells indicate performance below the benchmark. The star symbol (★) marks the best-performing model for each setup.

Table 4. Model performance – all models, both test setups.

Model	Accuracy	Recall	Precision	F1	ROC-AUC	PR-AUC
Literature Benchmark	99.0%	86.0%	13.0%	28.0%	0.9746	0.6639
Unbalanced Test Set (green = beats benchmark)						
Random Forest ★	99.5%	83.1%	23.1%	36.2%	0.9738	0.6645
LGBM+RF Hybrid	98.9%	85.9%	12.1%	21.2%	0.9755	0.6729
LightGBM	98.2%	88.0%	7.6%	14.1%	0.9759	0.6555
3-Model Ensemble (L+R+X)	98.1%	88.0%	7.3%	13.5%	0.9765	0.6540
Stacking Meta-Learner	95.9%	88.7%	3.5%	6.8%	0.9765	0.6558
XGBoost	94.6%	90.1%	2.7%	5.3%	0.9750	0.6602
Literature Benchmark	92.0%	95.0%	89.0%	92.0%	0.9798	0.9851
Balanced Test Set (green = beats benchmark)						
XGBoost ★	95.8%	92.6%	98.9%	95.7%	0.9770	0.9836
3-Model Ensemble (L+R+X)	95.3%	90.5%	100%	95.0%	0.9803	0.9852
Stacking Meta-Learner	95.3%	91.6%	98.9%	95.1%	0.9805	0.9851
LightGBM	94.7%	89.5%	100%	94.4%	0.9770	0.9843
LGBM+RF Hybrid	94.2%	88.4%	100%	93.9%	0.9819	0.9861
Random Forest	94.2%	88.4%	100%	93.9%	0.9799	0.9842

Random Forest achieves the best results on the imbalanced test set [7, 29]. Accuracy reaches 99.5%, exceeding the benchmark of 99.0%. Precision of 23.1% is substantially above the benchmark of 13.0%. F1-score of 36.2% significantly exceeds the benchmark of 28.0%.

The LGBM+RF hybrid achieves the best PR-AUC on the imbalanced test at 0.6729 [17]. This exceeds the benchmark PR-AUC of 0.6639. ROC-AUC of 0.9755 also exceeds the benchmark of 0.9746. Recall of 85.9% is close to the benchmark of 86.0%.

The three-model ensemble prioritizes recall at the expense of precision [21]. Recall reaches 88.0%, above the benchmark of 86.0%. Precision drops to 7.3% due to XGBoost's aggressive fraud flagging. ROC-AUC of 0.9765 is the highest across all imbalanced test models.

All six models outperform the benchmark on the balanced test set. XGBoost individually achieves the highest F1-score of 95.7% versus the benchmark of 92.0%. Precision is 100% across LGBM, RF, and their hybrid. PR-AUC reaches 0.9861 for the LGBM+RF hybrid, exceeding the benchmark of 0.9851.

Resampling Method Comparison

Table 5 presents results across three resampling strategies [3, 6]. SMOTE and ADASYN substantially outperform RandomUnderSampling on the imbalanced test. This finding has important implications for practical fraud detection system design.

Table 5. Resampling method comparison – training-only oversampling; test set pristine.

Resampling method	Accuracy	Recall	Precision	F1	ROC-AUC	PR-AUC
Unbalanced Test Set – oversampling on training folds only (green = beats benchmark)						
Literature Benchmark	99.0%	86.0%	17.0%	28.0%	0.9739	0.6639
RandomUnderSampling	96.9%	87.3%	4.5%	8.5%	0.9671	0.6512
SMOTE ★	99.9%	78.2%	88.8%	83.2%	0.9730	0.8107
ADASYN ★	99.9%	79.6%	90.4%	84.6%	0.9690	0.8198
Balanced Test Set (green = beats benchmark)						
Literature Benchmark	92.0%	95.0%	89.0%	92.0%	0.9798	0.9851
RandomUnderSampling	95.3%	91.6%	98.9%	95.1%	0.9842	0.9879
SMOTE	96.8%	93.7%	100%	96.7%	0.9989	0.9989
ADASYN ★	97.4%	94.7%	100%	97.3%	0.9922	0.9953

SMOTE achieves F1 of 83.2% and precision of 88.8% on the imbalanced test [6]. ADASYN achieves F1 of 84.6% and precision of 90.4%. Both results far exceed RandomUnderSampling's F1 of 8.5% [3]. Accuracy under both methods reaches 99.9%, exceeding all other configurations.

On the balanced test set, ADASYN achieves the strongest overall results. Accuracy of 97.4%, F1 of 97.3%, and PR-AUC of 0.9953 all exceed the benchmark. SMOTE also performs strongly with PR-AUC of 0.9989. RandomUnderSampling remains competitive on the balanced set [21].

RandomUnderSampling's weaker performance reflects a training size constraint [22]. After undersampling, the training pool contains only 463 samples. SMOTE and ADASYN produce much larger, richer training sets from the same data. Greater training data enables higher-confidence fraud probability estimates.

Cross-Validation Stability

Table 6 presents 5-fold stratified cross-validation results [16]. The ensemble is compared against published benchmark fold-level scores. Mean F1 of 0.9523 exceeds the benchmark's 0.9338 by 1.85 percentage points. The model outperforms the benchmark in four of five folds.

Table 6. 5-Fold stratified k-fold cross-validation results.

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean F1	Std Dev
This study – ensemble	0.9606	0.9106	0.9767	0.9531	0.9606	0.9523	0.0223
Literature benchmark	0.9197	0.9206	0.9612	0.9385	0.9291	0.9338	0.0153

Fold 2 produces the lowest F1 of 0.9106 for this study's ensemble. The benchmark's Fold 2 score is 0.9206, also among its lower folds [16]. This suggests the Fold 2 data partition is inherently more challenging. Stability remains acceptable for practical deployment purposes.

Feature Importance

SHAP analysis identifies the most influential features for fraud prediction [28]. V4 and V14 are the two most important features across all model configurations. V3, V15, and V11 rank third through fifth respectively. These are consistent with findings in the broader fraud detection literature [30, 31].

The iso_score Isolation Forest feature contributes positively to model calibration [9]. It provides an anomaly signal that is orthogonal to the PCA components. Time and Amount rank among the lowest-importance features [5]. This is consistent with the anonymized nature of the dataset.

Comparison with Literature Benchmarks

Table 7 provides a comprehensive comparison with eight studies from literature that also use the European Credit Card Fraud Dataset. The comparison covers model type, data preprocessing, reported AUC or PR-AUC, interpretability, tuning precision, computational cost, and deployment practicality [32].

Table 7. Comparative analysis of fraud detection models on the European credit card dataset.

Study (year)	Model	Data Pre-processing	AUC / PR-AUC	Interpretability	Tuning precision	Computational cost	Deployment	Remarks
Maheshwari et al. (2023) [13]	RNN-LSTM + Attention	SMOTE	Acc: 99.94% –	Low (DNN opaque)	Low (no tuning details)	High (resource-intensive)	Low (complexity hinders)	High accuracy but hard to interpret and deploy.
Wang (2025) [33]	CNN+BiGRU+LSTM+XGBoost	SMOTE-KMeans	~0.96 –	Medium (complex ensemble)	Medium (ensemble tuning)	Medium (multi-model)	Medium (depends on stack)	High performance but complex setup and training cost.
Talukder et al. (2024) [16]	Multi-stage Ensemble (Bagging+Voting)	Undersampling +IHT	1.000 –	Medium (multi-stage)	Medium (no visualisations)	High (multiple layers)	Medium (pipeline complex)	Extremely accurate but too complex for lightweight production.
Varmedja et al. (2019) [15]	RF / LR / NB / MLP	SMOTE	Prec: 96.38% –	Medium (traditional models)	Low (mostly default)	Medium (relatively efficient)	Medium (standard models)	SMOTE improves recall but may introduce overfitting risk.
Alarfaj et al. (2022) [14]	14-layer CNN	PCA	98% –	Very Low (deep CNN)	Very Low (no tuning described)	Very High (deep model)	Very Low (hard to deploy)	High accuracy; over-complex for financial deployment.
Reddy et al. (2024) [34]	JNBO + SpinalNet (Federated DL)	Bootstrapping	MAP: 89.82% –	Very Low (black-box)	Low (optimization focus)	Very High (resource-intensive)	Low (advanced infra needed)	Innovative but lacks practical generalisation in production.
Breskuviene & Dzemyda (2024) [35]	XGBoost / CatBoost / RF	FID-SOM Feature Selection	Strong across all metrics	High (feature visualisation)	Medium (dimensionality focus)	Medium–High (data scale)	Medium (research-oriented)	Well-suited for high-dimensional data; less dynamic detection.
Mienye & Swart (2024) [12]	GAN + GRU / LSTM	GAN-generated data	Sensitivity: 0.992 –	Very Low (GAN-based)	Low (no detailed trends)	Very High (unstable, costly)	Low (heavy resource demands)	Temporal fraud captured; prone to overfitting and instability.
This Study	LightGBM + RF + XGBoost (Advanced Ensemble)	RUS / SMOTE / ADASYN	0.9755 (Imbal.) 0.9861 (Bal.)	High (tree-based, SHAP)	High (Optuna 80 trials, F1 optimised)	Low (fast training)	High (easy to deploy)	Beats benchmarks on PR-AUC; ADASYN+ ensemble recommended.

To expand the feature vocabulary beyond raw statistical columns, the APATE framework developed by Van Vlasselaer et al. (2015) [36] utilizes social network extensions to propagate fraud signals transitively across graph networks of cardholders and merchants. Similarly, hierarchical behavior-knowledge space (BKS) models implemented by Nandi et al. (2022) [37] aggregate individual model outputs into systematic processing environments to flag anomalous transaction patterns.

Deep learning models – such as Maheshwari et al. (2023) [13] and Alarfaj et al. (2022) [14] – achieve high accuracy but suffer from low interpretability and very high computational cost. Wang (2025) [33] uses a complex four-model stack with competitive AUC but difficult maintenance requirements. Talukder et al. (2024) [16] achieve AUC of 1.000 but through a multi-stage pipeline that is impractical for lightweight production systems.

This study's ensemble achieves PR-AUC of 0.9861 on the balanced test while remaining fully tree-based and interpretable via SHAP [28]. It outperforms Varmedja et al. (2019) [15] and Reddy et al. (2024) [34] on both balanced and imbalanced test metrics. Breskuviene and Dzemyda (2024) [35] use strong feature selection but produce a model less suited to dynamic fraud environments.

DISCUSSION

Ensemble Composition and Precision

Random Forest individually outperforms the full three-model ensemble on precision [7, 29]. XGBoost with scale_pos_weight = 50 flags a large proportion of transactions as fraud [8]. Adding this aggressive model to the ensemble dilutes the conservative precision of RF and LGBM. For precision-critical applications, LGBM+RF is the recommended configuration [17].

For recall-critical applications, the full three-model ensemble is preferred [21]. Cost-sensitive weights should be tuned to the specific cost structure of the target deployment context. A value of 50 reflects a conservative assumption that undetected fraud imposes fifty times the operational cost of a false alarm – a threshold consistent with business continuity planning frameworks, where unrecovered fraud losses directly impair liquidity reserves and trigger regulatory reporting obligations under GRC mandates.

Resampling Strategy Selection

SMOTE and ADASYN deliver far superior performance on the imbalanced test set [3, 6]. F1-scores of 83–85% versus 8.5% for undersampling represent a tenfold improvement. This finding suggests that resampling strategy selection is the single most impactful design decision.

The recommendation for real-world deployment is ADASYN combined with the three-model ensemble. ADASYN's boundary-focused synthesis is better suited to highly imbalanced datasets. Its PR-AUC of 0.9953 on the balanced test set is the highest result in this study.

Threshold Calibration

The imbalanced test set requires careful threshold selection [22]. Validation sets after undersampling are 50/50 balanced and provide inflated F1 estimates. Using a fixed threshold of 0.68 avoids this distribution mismatch problem. This design choice is critical for reliable imbalanced test evaluation.

Future work should explore probability calibration methods such as Platt scaling or isotonic regression [38]. Calibrated probabilities would allow more principled threshold selection across deployment environments. Temperature scaling is a simple calibration approach compatible with all three base models.

Computational Considerations

Optuna at 80 trials requires approximately 35–45 minutes on a standard CPU [4]. The full pipeline runs four Optuna sessions, totaling approximately 2.5–3 hours. Transitioning the training loop to modern parallel computing setups would reduce absolute execution times significantly. The stacking meta-learner adds a 5-fold cross-validation pass but trains in milliseconds [25, 26].

Managerial Implications for Corporate Governance and Business Continuity

The results of this study carry direct implications for enterprise governance and operational resilience. Financial institutions operating under governance, risk, and compliance (GRC) frameworks are required to demonstrate that fraud controls are both effective and auditable. The tree-based ensemble architecture proposed here satisfies both requirements. SHAP feature attribution provides line-by-line

interpretability for every fraud prediction, enabling compliance officers to document the basis for flagging decisions. This transparency directly supports obligations under regulations, such as Basel III operational risk guidelines and PCI-DSS audit requirements, where black-box models would face significant barriers to regulatory acceptance.

From a business continuity planning (BCP) perspective, false-alarm rate control is a material operational concern. High false-alarm rates impose hidden costs: legitimate customer transactions are declined, customer trust erodes, and operations teams face unsustainable review workloads. The LGBM+RF configuration, which achieves 23.1% precision on the imbalanced test set – nearly double the benchmark – directly reduces this operational drag. Fewer incorrect fraud flags mean faster transaction throughput, reduced dispute resolution costs, and lower strain on fraud operations teams. These gains translate into measurable improvements in service continuity and resource recovery efficiency under BCP protocols.

At the enterprise risk management level, the modular design of the framework supports scalable deployment across multi-entity financial organizations. Each component – anomaly scoring, ensemble classification, and meta-learning – can be independently updated as fraud patterns evolve, without requiring full model retraining. This adaptability is critical for organizations managing dynamic threat environments where static rule-based systems consistently fail. Senior risk officers and board-level governance committees can use the SHAP-derived feature rankings to assess which transaction characteristics are driving fraud exposure at any point in time, supporting proactive rather than reactive risk posture management.

CONCLUSION

This study proposes an advanced ensemble framework for optimizing corporate financial risk governance and ensuring operational resilience against transactional fraud. Five methods are combined to enhance enterprise defense capabilities: an unsupervised Isolation Forest risk signal [9], XGBoost ensemble integration [8], Optuna Bayesian tuning for maximum detection stability [4], cost-sensitive learning calibrated to asymmetric fraud cost structures [3] and stacking meta-learners for unified risk scoring [25, 26]. Experiments on 283,726 real-world transaction records demonstrate strong risk-capture performance across both evaluation setups.

Random Forest achieves 99.5% accuracy and 23.1% precision on the imbalanced test set [7, 29]. These results exceed published benchmarks in the fraud detection literature [15, 16]. The LGBM+RF hybrid achieves the best PR-AUC of 0.6729 [17]. All six models beat benchmark performance on the balanced test set.

Resampling strategy is the most impactful design decision in this framework [3, 6]. SMOTE and ADASYN deliver F1-scores of 83–85% on the imbalanced test. RandomUnderSampling achieves only 8.5% F1 under the same conditions. ADASYN with the three-model ensemble is recommended for real-world deployment.

Cross-validation confirms model stability with mean F1 of 0.9523 [16]. Four of five folds outperform the literature benchmark. Standard deviation of 0.0223 indicates acceptable variance across data splits. The framework generalizes well across different subsets of the training data.

Key limitations include a training pool of only 473 fraud cases after deduplication [5]. Future work should incorporate SMOTE or ADASYN as the primary resampling strategy throughout [6]. Deep learning integration for temporal fraud sequences merits exploration [13, 14, 34, 38]. Probability calibration and SHAP explainability tools should be incorporated for regulatory compliance [28].

REFERENCES

1. Naman BM, Abdulazeez AM. Credit card fraud detection based on machine learning classification algorithm. Indones J Comput Sci. 2024;13(3). Available from: https://doi.org/10.33022/ijcs.v13i3.3996

2. Japkowicz N, Stephen S. The class imbalance problem: A systematic study. Intell Data Anal. 2002;6(5):429–449. Available from: https://doi.org/10.3233/ida-2002-6504
3. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263–1284. Available from: https://doi.org/10.1109/TKDE.2008.239
4. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019. Available from: https://doi.org/10.1145/3292500.3330701
5. Dal Pozzolo A, Caelen O, Johnson RA, Bontempi G. Credit Card Fraud Detection Dataset. Kaggle. 2014. Available from: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
6. Chawla N, Bowyer K, Hall L, Kegelmeyer W. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–357. Available from: https://doi.org/10.1613/jair.953
7. Simaiya S, Lilhore UK, Sharma SK, Trivedi NK. An efficient credit card fraud detection model based on machine learning methods. Int J Adv Sci Technol. 2020;29(5):3414–3424.
8. Hajek P, Abedin M, Sivarajah U. Fraud detection in mobile payment systems using an XGBoost-based framework. Inf Syst Front. 2022;25(5):1985–2003. Available from: https://doi.org/10.1007/s10796-022-10346-6
9. Carcillo F, Le Borgne YA, Caelen O, Kessaci Y, Oblé F, Bontempi G. Combining unsupervised and supervised learning in credit card fraud detection. Inf Sci. 2021;557:317–331. Available from: https://doi.org/10.1016/j.ins.2019.05.042
10. Ghosh S, Reilly DL. Credit card fraud detection with a neural-network. Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences. 1994;3:621–630. Available from: https://doi.org/10.1109/hicss.1994.323314
11. Khatri S, Arora A, Agrawal AP. Supervised machine learning algorithms for credit card fraud detection: A comparison. 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence). 2020:680–683. Available from: https://doi.org/10.1109/Confluence47617.2020.9057851
12. Mienye ID, Swart TG. A hybrid deep learning approach with generative adversarial network for credit card fraud detection. Technologies (Basel). 2024;12(10). Available from: https://doi.org/10.3390/technologies12100186
13. Maheshwari VC, Osman NA, Aziz N. A hybrid approach adopted for credit card fraud detection based on deep neural networks and attention mechanism. J Adv Res Appl Sci Eng Technol. 2023;32(1):315–331. Available from: https://doi.org/10.37934/araset.32.1.315331
14. Alarfaj FK, Malik I, Khan HU, Almusallam N, Ramzan M, Ahmed M. Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. IEEE Access. 2022;10:39700–39715. Available from: https://doi.org/10.1109/ACCESS.2022.3166891
15. Varmedja D, Karanovic M, Sladojevic S, Arsenovic M, Anderla A. Credit card fraud detection—Machine learning methods. 18th International Symposium INFOTEH-JAHORINA. 2019. Available from: https://doi.org/10.1109/INFOTEH.2019.8717766
16. Talukder MA, Khalid M, Uddin MA. An integrated multistage ensemble machine learning model for fraudulent transaction detection. J Big Data. 2024;11(1). Available from: https://doi.org/10.1186/s40537-024-00996-5
17. Azim Mim M, Majadi N, Mazumder P. A soft voting ensemble learning approach for credit card fraud detection. Heliyon. 2024;10(3):e25466. Available from: https://doi.org/10.1016/j.heliyon.2024.e25466
18. Esenogho E, Mienye ID, Swart TG, Aruleba K, Obaido G. A neural network ensemble with feature engineering for improved credit card fraud detection. IEEE Access. 2022;10:16400–16407. Available from: https://doi.org/10.1109/ACCESS.2022.3148298
19. Panigrahi R, Borah S, Bhoi AK, Ijaz MF, Pramanik M, Kumar Y. A consolidated decision tree-based intrusion detection system for binary and multiclass imbalanced datasets. Mathematics. 2021;9(7):751. Available from: https://doi.org/10.3390/math9070751
20. Randhawa K, Loo CK, Seera M, Lim CP, Nandi AK. Credit card fraud detection using AdaBoost and majority voting. IEEE Access. 2018;6:14277–14284. Available from: https://doi.org/10.1109/ACCESS.2018.2806420
21. Makki S, Assaghir Z, Taher Y, Haque R, Hacid M-S, Zeineddine H. An experimental study with imbalanced classification approaches for credit card fraud detection. IEEE Access. 2019;7:93010–93022. Available from: https://doi.org/10.1109/ACCESS.2019.2927266

22. Mrozek P, Panneerselvam J, Bagdasar O. Efficient resampling for fraud detection during anonymised credit card transactions with unbalanced datasets. IEEE/ACM 13th International Conference on Utility and Cloud Computing (UCC). 2020:426–433. Available from: https://doi.org/10.1109/UCC48980.2020.00065
23. El Kafhali S, Tayebi M. XGBoost based solutions for detecting fraudulent credit card transactions. 2022 International Conference on Advanced Creative Networks and Intelligent Systems (ICACNIS). 2022:1–6. Available from: https://doi.org/10.1109/ICACNIS57039.2022.10054965
24. Padhi DK, Padhy N, Bhoi AK, Shafi J, Ijaz MF. A fusion framework for forecasting financial market direction using enhanced ensemble models and technical indicators. Mathematics. 2021;9(21):2646. Available from: https://doi.org/10.3390/math9212646
25. Xia Y, Zhao J, He L, Li Y, Niu M. A novel tree-based dynamic heterogeneous ensemble method for credit scoring. Expert Syst Appl. 2020;159:113615. Available from: https://doi.org/10.1016/j.eswa.2020.113615
26. Yotsawat W, Wattuya P, Srivihok A. A novel method for credit scoring based on cost-sensitive neural network ensemble. IEEE Access. 2021;9:78521–78537. Available from: https://doi.org/10.1109/ACCESS.2021.3083963
27. Dal Pozzolo A, Boracchi G, Caelen O, Alippi C, Bontempi G. Credit card fraud detection: A realistic modeling and a novel learning strategy. IEEE Trans Neural Netw Learn Syst. 2017;29(8):3784–3797. Available from: https://doi.org/10.1109/TNNLS.2017.2736643
28. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30.
29. Xuan S, Liu G, Li Z, Zheng L, Wang S, Jiang C. Random forest for credit card fraud detection. 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC). 2018:1–6. Available from: https://doi.org/10.1109/icnsc.2018.8361343
30. Taha AA, Malebary SJ. An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. IEEE Access. 2020;8:25579–25587. Available from: https://doi.org/10.1109/ACCESS.2020.2971354
31. Alfaiz NS, Fati SM. Enhanced credit card fraud detection model using machine learning. Electronics. 2022;11(4):662. Available from: https://doi.org/10.3390/electronics11040662
32. Mienye ID, Sun Y. Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. Inform Med Unlocked. 2021;25:100690. Available from: https://doi.org/10.1016/j.imu.2021.100690
33. Wang Y. A data balancing and ensemble learning approach for credit card fraud detection. 2025 4th International Symposium on Computer Applications and Information Technology (ISCAIT). 2025:386–390. Available from: https://doi.org/10.1109/ISCAIT64916.2025.11010591
34. Reddy VVK, Reddy RVK, Munaga MSK, Karnam B, Maddila SK, Kolli CS. Deep learning-based credit card fraud detection in federated learning. Expert Syst Appl. 2024;255:124493. Available from: https://doi.org/10.1016/j.eswa.2024.124493
35. Breskuviene D, Dzemyda G. Enhancing credit card fraud detection: Highly imbalanced data case. J Big Data. 2024;11(1). Available from: https://doi.org/10.1186/s40537-024-01059-5
36. Van Vlasselaer V, Bravo C, Caelen O, Eliassi-Rad T, Akoglu L, Snoeck M, et al. APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions. Decis Support Syst. 2015;75:38–48. Available from: https://doi.org/10.1016/j.dss.2015.04.013
37. Nandi AK, Randhawa KK, Chua HS, Seera M, Lim CP. Credit card fraud detection using a hierarchical behaviour-knowledge space model. PLOS ONE. 2022;17(1):e0260579. Available from: https://doi.org/10.1371/journal.pone.0260579
38. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data. 2019;6(1):1–54. Available from: https://doi.org/10.1186/s40537-019-0192-5

Disclaimer: The views and opinions expressed in this paper are those of the authors and do not necess arily reflect the official policy or position of their affiliated organizations.

