

A Generative and Cluster-Aware Multi-model Architecture for Detecting Heterogeneous Fraud in Imbalanced Transaction Data

Berat Can Karakaş[1(✉)], Hakan Hakvar[2], Oğulcan Söyler[2], Şevval Türkyılmaz[2], Ilgin Şafak[2,3], Tolga Durdu[2], Emin Anarım[4], and Mehmet S. Aktaş[1]

[1] Yıldız Technical University, Istanbul, Turkey
berat.karakas@std.yildiz.edu.tr, aktas@yildiz.edu.tr
[2] Fibabanka R&D Center, Istanbul, Turkey
{hakan.hakvar,ogulcan.soyler,sevval.turkyilmaz,ilgin.safak,tolga.durdu}@fibabanka.com.tr, ilgin.i.safak@jyu.fi
[3] University of Jyväskylä, Jyväskylä, Finland
[4] Boğaziçi University, Istanbul, Turkey
anarim@bogazici.edu.tr

Abstract. With the rapid development of digital banking, the issue of financial fraud detection has grown even more pressing. Real-world fraud systems have two major challenges: extreme class imbalance and behavioral variability of fraudulent activities, which are frequently not captured by monolithic global models. We propose a cluster-aware multi-model framework that simultaneously tackles both challenges. To handle the class imbalance efficiently, we first employ a Conditional Tabular Generative Adversarial Network (CTGAN) to generate high-fidelity synthetic fraud examples. Second, we cluster fraudulent transactions into behaviorally consistent subgroups using non-linear dimensionality reduction (UMAP) and density-based clustering (DBSCAN). This breakdown enables a Mixture-of-Experts (MoE) architecture in which specialized classifiers are trained for different fraud behaviors. We evaluate our approach on a popular benchmark for credit card fraud and show that it outperforms classical resampling methods and strong global baselines. Ultimately, the suggested technique detects infrequent and confined fraud patterns more effectively, resulting in a much higher F1-score and operational viability.

Keywords: Fraud Detection · Class Imbalance · CTGAN · UMAP · Mixture-of-Experts

1 Introduction

In today's digital financial systems, efficient credit card fraud detection is essential to prevent direct financial losses, defend reputations of institutions and

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16758, pp. 16–32, 2027.
https://doi.org/10.1007/978-3-032-30497-1_2

assure the compliance with regulations [1]. While machine learning has developed significantly, it is still quite hard to detect fraud transactions in real life. Fraud detection is challenging due to severe class imbalance, where genuine transactions are greatly in excess of fraudulent transactions [2], and the ongoing growth of adversarial methods. As a consequence, fraud shows up in very varied patterns, which classic global models are not able to model sufficiently [3].

In the past, resampling methods at the data level, such as SMOTE with linear interpolation [4], have been utilized to address the class imbalance. However, in high-dimensional tabular domains (e.g., with 31 complex variables), features commonly exhibit non-linear interactions and multi-modality. In these settings, interpolation-based methods often generate unrealistic samples from the majority class, which blurs the decision boundaries and hurts the classification performance.

To address these constraints, Generative Adversarial Networks (GANs) have gained popularity in directly learning and reproducing the underlying data distributions through adversarial training [5]. Specialized designs, e.g., CT-GAN, have been shown to be particularly effective for tabular data with mixed feature types [6]. However, producing minority samples solves just part of the problem. Most present strategies address class imbalance in isolation, ignoring the structural variety of fraudulent behavior. Fraud is not a single monolithic event but a collection of different and localized sub-patterns, which are usually not detected by single global classifiers [8].

To address this issue, we offer a dual-layered, cluster-aware multi-model framework, which is evaluated on a well-known public benchmark dataset. First, we perform oversampling with CTGAN to better capture the complex distribution of the minority class. Second, we breakdown fraudulent transactions into behaviorally consistent subtypes using non-linear dimensionality reduction (UMAP) and density-based clustering (DBSCAN). This behavior-aware decomposition results in an architecture like the Mixture-of-Experts (MoE) paradigm, where specialized classifiers are assigned to certain portions of the feature space for full fraud coverage [9].

The main contributions of this paper are as follows:

- A GAN-based oversampling pipeline that models the complex distributions of minority classes in tabular fraud data better than standard interpolation methods.
- A clustering-based multi-model detection system that addresses fraud heterogeneity and demonstrates that specialized "expert" models outperform monolithic global classifiers in detecting rare fraud sub-patterns.
- A detailed evaluation on the public European cardholders dataset, with the use of precision-recall-based metrics (F1-score) and cost-sensitive evaluation to ensure robust performance measurement and operational viability despite high class imbalance [10].

The rest of this paper is organized as follows. Section 2 discusses related work on fraud detection and unbalanced learning. Section 3 presents the proposed technique. Section 4 outlines the experimental setup and the characteristics of the

dataset. The experimental data are presented and discussed in Sect. 5. Section 6 provides the conclusions.

2 Related Work

Financial systems have been studied for many years in the context of detecting fraud and unusual activities. However, there are still major practical hurdles, especially in real-world scenarios with extreme class imbalance, dynamic attack techniques, and varied fraud behaviors. The existing literature on fraud detection can be broadly divided into three categories: (i) supervised learning paradigms, (ii) methods dealing with class imbalance, and (iii) approaches modeling explicitly the structural heterogeneity of fraud.

2.1 Supervised Approaches to Fraud Detection

In the typical setting, credit card fraud detection is seen as a supervised classification problem. In this area, several experiments have been carried out on conventional models, like decision trees, logistic regression and neural networks [12]. The dominant paradigm in these approaches implies that genuine consumer behavior follows generally consistent patterns, which makes it possible to detect fraudulent transactions by looking for variations from historical spending [11].

However, accuracy-driven models are typically not sufficient in operational contexts because of the substantial class imbalance in fraud datasets and the high cost of false negatives [2]. This constraint has driven the development of cost-sensitive learning methods and alternate evaluation measures [13]. Cost-sensitive learning encourages optimization but does not solve the inherent lack of minority class samples.

2.2 Imbalanced Learning and Resampling Methods

Imbalanced categorization is generally handled by resampling approaches at the data level. Baseline approaches such as SMOTE and its derivatives generate synthetic minority samples by interpolating existing minority samples [4,13]. Interpolation-based oversampling is commonly used; however, it might create unrealistic samples in high-dimensional transaction data (e.g., with 31 complex characteristics), particularly when fraud behavior is multi-modal. In such cases, synthetic samples are close to the majority class, leading to fuzzy decision boundaries. To avoid these difficulties, generative approaches that actively learn the underlying data distribution have gained popularity recently. GAN-based techniques enable the generation of complex minority class distributions [5]. Specialized tabular architectures such as CTGAN have been shown to generate high-fidelity synthetic data for datasets with mixed feature types [6].

2.3 Fraud Heterogeneity and Multi-Model Architectures

Fraudulent behavior is not a single phenomenon because it involves different attack vectors that evolve dynamically [3]. Hence, a single global detection model can be effective in detecting common fraud patterns but often fails to discover unusual yet important sub-patterns [14].

To solve this problem, subsequent studies have explored clustering-based and multi-model approaches. Density-based clustering methods like DBSCAN and HDBSCAN are useful for separating different fraud tendencies in labeled datasets. These methods are quite robust for high dimensional tabular data when combined with dimensionality reduction methods such as UMAP to avoid curse of dimensionality [7,15]. Likewise, Mixture-of-Experts (MoE) architectures provide a systematic framework where specialist models contribute collectively to boost performance across different parts of the feature space [9]. However, many existing approaches ignore class imbalance as a secondary concern or do not integrate synthetic data generation with behavior-based decomposition seamlessly.

2.4 Positioning of the Proposed Framework

The aforementioned research gaps are addressed in this work via a comprehensive dual-tier methodology. Existing studies usually consider class imbalance [6] and structural heterogeneity [7] as individual challenges, whereas our framework unifies them into a single pipeline. This study is ideally positioned to overcome the limits of monolithic global classifiers by integrating generative minority oversampling with behavior-aware expert modeling to address unusual, multi-modal fraud detection patterns.

3 Proposed Methodology

The suggested fraud detection approach can simultaneously deal with the severe class imbalance and the behavioral heterogeneity of credit card transaction data. The system combines GAN-based minority class oversampling with a clustering-based multi-model architecture. The technique is structured around four main stages: (i) Problem characterization and terminology, (ii) Generative minority oversampling, (iii) Dimensionality reduction with density-based clustering, and (iv) Cluster-aware expert routing for inference.

The architecture depends on a cluster-aware modeling pipeline to comprehensively address varied fraud patterns (see Fig. 1). After dataset isolation, the "Cluster-Aware Data Splitter" module conducts dimensionality reduction and DBSCAN only on fraudulent transactions to identify behaviorally distinct groupings. Then, each detected cluster is divided into dedicated training and testing sets. The legitimate transactions are proportionally stratified over these cluster-specific subsets to keep the original class distribution and to give a realistic evaluation.

During training, these separate subsets optimize different tree-based ensemble models, yielding a localized set of "expert" classifiers. In real-time inference

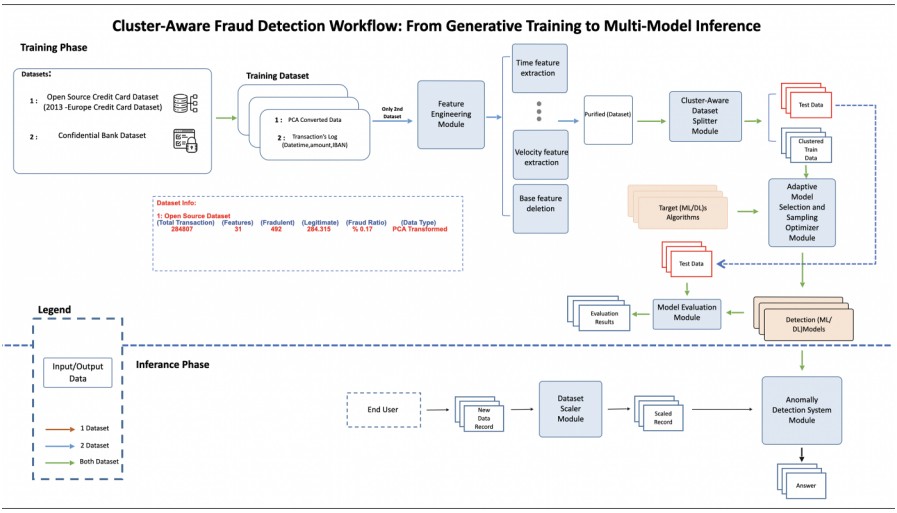

Fig. 1. The proposed clustering-based multi-model fraud detection workflow, illustrating the progression from data augmentation (CTGAN) and dimensionality reduction (UMAP) to cluster-aware expert routing.

the framework acts as an adaptive routing engine, projecting incoming transactions in the learned topological space, mapping them to the most relevant fraud cluster, and routing them to the related pre-trained expert. This mutually exclusive routing strongly precludes multiple counting of negative samples.

3.1 Problem Definition and Notation

Let $D = \{(x_i, y_i)\}_{i=1}^{N}$ be a set of credit card transactions, where $x_i \in \mathbb{R}^d$ is a d-dimensional feature vector ($d = 30$, comprising 28 PCA components, time, and amount) and $y_i \in \{0, 1\}$ is the class label ($y_i = 1$ for fraud, $y_i = 0$ for legitimate), yielding 31 total variables. In real-world fraud detection environments, the class distribution is severely imbalanced, as noted in $P(y = 1) \ll P(y = 0)$ [2].

The goal is to learn a classification function $f : \mathbb{R}^d \to \{0, 1\}$ to discriminate legal vs. fraudulent transactions while being robust to excessive class imbalance and various and developing fraud behaviors.

3.2 Framework Overview

The suggested framework adopts a two-level approach. At the data level, CTGAN-based oversampling extends the distribution of the minority class while maintaining structural fidelity and without deleting samples of the majority class. At the model level, UMAP and density-based clustering break fraudulent cases into behaviorally consistent subtypes. For these discovered clusters, dedicated expert classifiers (e.g., XGBoost, LightGBM) are trained separately. This

isolates localized fraud dynamics for a more fine-grained representation than monolithic architectures.

3.3 GAN-Based Minority Class Oversampling

To alleviate class imbalance, we employ a Conditional Generative Adversarial Network (CTGAN) built for tabular datasets [6]. Unlike classic interpolation techniques such as SMOTE [4], CTGAN learns the underlying distribution of the minority class directly, which allows it to successfully capture non-linear relationships and multi-modal feature dependencies. The generator network only trains on fraudulent instances, generating synthetic transactions based on learned feature distributions, while the discriminator assesses their validity [5]. Importantly, this augmentation is entirely confined to the training set; validation and test sets remain unaltered to categorically exclude data leaking.

3.4 Dimensionality Reduction and Density-Based Clustering

Fraudulent transactions are naturally diversified because they are generated by different adversarial methods [3]. Applying distance-based algorithms like DBSCAN [7] directly in high-dimensional continuous space (i.e., over all 30 input features) would always suffer from the curse of dimensionality and the loss of discriminative ability of distance metrics.

To address this, we first perform Uniform Manifold Approximation and Projection (UMAP) on only the fraud-labeled training samples. UMAP lowers the feature space while keeping non-linear topological linkages, greatly boosting cluster separation while decreasing the growth of unclassified noise. After projecting, DBSCAN clusters the data into coherent behavioral clusters.

To ensure statistical reliability and thorough evaluation (e.g., permitting robust cross-validation), we add a subsequent *Cluster Merging* protocol. Clusters with high centroid proximity in the UMAP space are iteratively merged until each resulting cluster includes a statistically feasible minimum threshold of instances (e.g., at least 7 fraudulent transactions in each corresponding test set). Unclustered or extremely anomalous points are assigned to a dominant noise cluster (Cluster -1), which provides a generic baseline representation of fraud.

3.5 Cluster-Specific Expert Modeling

Following the grouping and merging step, each detected fraud subtype is considered as a separate modeling domain. We train one expert classifier per cluster on the particular fraudulent instances of that cluster and a stratified sample of the legal transactions. Tree based ensemble models are used, as they are shown to work well on tabular financial data [12]. Inspired by the principles of Mixture-of-Experts [9], this specialization forces separate algorithms to attend only to different portions of the feature space, thereby finding localized abnormalities that a global classifier might miss.

3.6 Inference Routing and System Integration

We propose an explicit inference routing system to process incoming, unlabeled transactions in a live or testing environment without risk of data leaking. First, the new transaction is projected onto the already trained UMAP embedding space, as described in Algorithm 1. The system estimates the Euclidean distance between the new instance and the centroids of the discovered fraud clusters, and the transaction is deterministically routed to the expert classifier of the nearest cluster.

If the minimum distance is more than a predetermined dynamic threshold (τ), indicating a completely novel pattern, the transaction is directed to the dominant/noise expert (Cluster -1). This deterministic mutually exclusive routing ensures that each transaction is reviewed by exactly one expert. This, in turn, structurally removes the double-counting of negative samples and makes the aggregated global metrics (Precision, Recall, F1-score) statistically sound and strictly comparable to a single-model global baseline.

Algorithm 1 Inference Routing and Multi-Model Classification Strategy

Input: A new transaction x_{new}, learned UMAP projection mapping U, set of fraud cluster centroids $C = \{c_1, c_2, ..., c_k\}$, dominant (noise) cluster centroid c_{noise}, set of trained expert models $E = \{E_1, E_2, ..., E_k, E_{noise}\}$, distance threshold τ.
Output: Fraud probability score p.

$x_{proj} \leftarrow U(x_{new})$ {Project transaction into the reduced UMAP space}
$d_{min} \leftarrow \infty$
$c_{target} \leftarrow c_{noise}$
for each $c_i \in C$ **do**
  $d \leftarrow$ EuclideanDistance(x_{proj}, c_i)
  if $d < d_{min}$ **then**
    $d_{min} \leftarrow d$
    $c_{target} \leftarrow c_i$
  end if
end for
if $d_{min} > \tau$ **then**
  $c_{target} \leftarrow c_{noise}$ {Route to generalized expert if it is an outlier}
end if
$E_{target} \leftarrow$ GetExpert(c_{target})
$p \leftarrow E_{target}(x_{new})$ {Obtain prediction from the assigned expert}
return p

4 Experimental Setup

This section describes the experimental setting used to evaluate the proposed fraud detection framework. We discuss the benchmark dataset, the data partitioning procedures, the baseline models, the resampling methods, and the overall assessment protocols used to evaluate the results.

4.1 Dataset

The proposed approach is assessed using the famous European cardholder dataset, released in 2013, which is widely considered as the benchmark standard for credit card fraud detection research [2]. We use this public dataset to assure reproducibility and to allow for fair benchmarking with the current state-of-the-art investigations.

The data set contains 284,807 records of transactions performed by European cardholders throughout two days. It has a strong class imbalance, which is indicative of actual world financial environments [13]. In particular, it has 284,315 legitimate transactions and just 492 fraudulent cases, i.e., fraud represents only 0.17% of the data set.

There are a total of 31 variables in the dataset (30 input features and 1 target class label). To preserve the privacy of the customers, the original transaction details are transformed via Principal Component Analysis (PCA) [3], and 28 continuous main components (V1 to V28) are obtained. The only input features that are not changed by PCA are *Time* (seconds elapsed from the first transaction) and *Amount* (amount of the transaction).

The PCA components are already standardized, thus only *Time* and *Amount* were scaled using a RobustScaler. The scaling technique is based on interquartile ranges and as such is quite resilient to extreme values and monetary outliers. Thus, this dataset is a demanding test bed for our system to work on extreme class imbalance and high dimensional continuous tabular data.

4.2 Data Partitioning Strategy

Experiments are carried out in a supervised learning setting with stratified data split to preserve the initial extreme class imbalance for all subsets. The two suggested solutions (GAN-based oversampling and the cluster-based multi-model framework) demand various partitioning strategies arising from their different architectural objectives.

Method 1: GAN-Based Oversampling For the generative baseline experiments the dataset is divided into three mutually exclusive subsets: 70% for training, 15% for validation and 15% for testing. The training set is utilized just for the optimization of classifiers and for the creation of synthetic minority samples. The validation set is used for selecting optimal decision thresholds, whereas the test set is strictly segregated for the final performance evaluation. This strict separation architecturally prevents information leakage.

Method 2: Cluster-Based Multi-model Framework The multi-model architecture partitioning technique consists of two parts. Firstly, dimensionality reduction and density based clustering are only used to the fraud-labeled instances in the 70% training sample to detect coherent behavioral sub-patterns.

Legitimate transactions are totally excluded to protect boundary integrity during this clustering phase.

Then, for each fraud cluster, a cluster-specific dataset is built by merging the assigned fraudulent samples with a proportional stratified sample of valid transactions. Then each cluster-specific dataset is divided into training, validation and testing sets (70/15/15). *It is important to emphasize that this multi-model routing architecture is naturally modular; it does not strictly require CTGAN to function and can operate effectively with baseline data or normal resampling approaches (e.g. SMOTE) inside each cluster. Our fully integrated ideal workflow, however, is to run CTGAN locally on the clusters.*

4.3 Baseline Classification Models

In order to do a complete evaluation and to confirm the model-agnostic nature of the presented framework, we have included a variety of classification algorithms that are well-known in the literature on tabular data: Random Forest, AdaBoost, CatBoost, XGBoost and LightGBM. All classifiers are trained with the same feature representations and pre-processing procedures.

4.4 Data Resampling and Augmentation Strategies

To benchmark the efficacy of the generative augmentation against the extreme class imbalance, the following strategies were evaluated:

- No resampling (Baseline original class distribution)
- Random Under-Sampling (RUS)
- SMOTE and SMOTEENN
- CTGAN-based generative oversampling (Proposed)

To rigorously prevent data leakage, all synthetic samples are generated exclusively from the training folds. The CTGAN model is trained solely on the minority-class samples of the training data.

4.5 Evaluation Metrics and Cost-Sensitive Assessment

Accuracy-based measures are not enough in principle because of the high class imbalance. We use three complementing measures to evaluate both statistical performance and practical business value comprehensively:

F1-Score: The main objective function is to optimize the F1-score to balance the precision and recall. Optimal decision criteria are selected by maximizing the F1-score on the validation set and are then applied to the test set for fair comparisons.

Recall at Fixed Precision (Recall@90%): Reducing false positives in operational banking environments is crucial to reduce client frustration. We present the recall attained when the threshold of precision is kept firmly at 90% and show the ability of the model to detect fraud under a tight false alarm budget.

Cost-Sensitive Assessment (Cost Savings %): A cost-sensitive evaluation is performed to measure the direct financial impact, simulating actual operations. The cost of a False Negative is the unrecoverable amount of the transaction, while a false positive incurs a fixed administrative investigation cost (estimated at €10). We report the *Cost Savings* percentage, which is the amount of financial loss prevented with respect to a baseline scenario without detection system.

4.6 Implementation Details and Hyperparameter Configuration

To ensure full reproducibility, hyperparameters were systematically configured and maintained identically across global and multi-model experiments; the complete implementation—including the CTGAN training pipeline, UMAP projection, DBSCAN clustering, centroid-based merging, and expert routing—is publicly available at [16].

Generative Model (CTGAN): Implemented via the SDV library, the model trained exclusively on the minority class using fully connected layers (256, 256). Training ran for 200 epochs (batch size 1,000) using the Adam optimizer (learning rate 2×10^{-4}). We generated three synthetic samples per real fraud instance. Quality was validated feature-wise using Kolmogorov-Smirnov (KS) statistics and Wasserstein distances, yielding an average Wasserstein score of ≈ 0.3, indicating moderate distributional similarity; despite this limitation, CTGAN consistently outperformed all interpolation-based baselines in downstream classification.

Clustering Pipeline (UMAP & DBSCAN): UMAP projected the fraud-only feature space into 3 components (`n_neighbors` = 15, `min_dist` = 0.0). DBSCAN was subsequently applied to this 3D embedding with `min_samples` = 5, and `eps` was determined analytically via the k-distance elbow method.

Expert Classifiers: Decision thresholds were optimized on the validation set. Key configurations included:

- **LightGBM:** `learning_rate`=0.025, `num_leaves`=7, `max_depth`=4.
- **XGBoost:** `learning_rate`=0.03, `max_depth`=2, `tree_method`="hist".
- **CatBoost:** `learning_rate`=0.02, `depth`=12.
- **AdaBoost:** `learning_rate`=0.1, 100 estimators.
- **Random Forest:** 100 estimators, `class_weight`="balanced".

5 Experimental Results

This section provides the experimental results for evaluating the performance of the proposed fraud detection framework. The evaluation is performed rigorously according to the methods defined in Sect. 4, and aims at assessing both statistical performance and operational feasibility.

First, we isolate and examine the effect of CTGAN-based minority class augmentation in a vanilla supervised learning setting. Next, we assess the clustering-driven multi-model architecture to test its effectiveness in detecting heterogeneous fraud patterns. All experiments are conducted on isolated test sets with the mutually exclusive routing system to provide rigorous validation and strictly avoid data leaking.

The final multi-model system is benchmarked with cost-sensitive operational metrics and compared with present state-of-the-art research to reflect real world financial needs. The computing environment overview is finally presented to show actual deployment scalability.

5.1 Results of Generative Oversampling (Method 1)

In this subsection, we examine the performance of CTGAN based minority class augmentation in a traditional supervised context. The generative approach is benchmarked against a number of data level resampling baselines: No Resampling (Normal), Random Under-Sampling (RUS), SMOTE and SMOTEENN. While SMOTE is a popular interpolation method [4], CTGAN uses adversarial learning to model complex tabular distributions better [5,6].

Table 1 shows the test F1-scores of five baseline classifiers under different resampling procedures.

Table 1. Test F1-scores under different resampling strategies on the public dataset (European 2013).

Model	Normal	RUS	SMOTE	SMOTEENN	CTGAN
Random Forest	0.8449	0.7937	0.8380	0.8000	**0.8649**
AdaBoost	0.7634	0.7917	0.8042	**0.8105**	0.8000
CatBoost	0.8427	0.8040	0.7845	0.8290	**0.8649**
XGBoost	0.8541	0.7817	0.8105	0.8148	**0.8817**
LightGBM	0.8492	0.7650	0.8021	0.7935	**0.8817**

The best average performance across the classifiers is achieved by CTGAN with the mean F1-score of 0.8586, which is an absolute +0.0277 improvement over the Normal baseline (mean F1 = 0.8309). On the other hand, RUS always drops overall performance (mean F1 = 0.7872), which shows that a significant exclusion of majority-class data has a damaging effect on border generalization in highly imbalanced domains.

Model-specific analysis demonstrates that CTGAN significantly improves advanced gradient boosting topologies. In particular, F1-score of LightGBM goes up to 0.8817 (+0.0325), XGBoost to 0.8817 (+0.0276), CatBoost to 0.8649 (+0.0222). AdaBoost has its best results with SMOTEENN (0.8105) but still improves significantly above its baseline (0.7634) with CTGAN (0.8000). These results validate that CTGAN for the generation of high-fidelity synthetic data better addresses the problem of excessive class imbalance than traditional interpolation.

5.2 Fraud Cluster Analysis and Merging Strategy

We analyze the clustering-based multi-model architecture in this part. The aim is to explore whether an explicit modeling of fraud subtypes with localized expertise leads to better performance than a monolithic global classifier.

UMAP Projection and Centroid-Based Merging. Direct application of density-based clustering on high-dimensional continuous spaces typically triggers curse of dimensionality. In preliminary tests, the typical DBSCAN algorithm categorized over 91% of fraudulent transactions as uninformative noise. To this end, we employed UMAP for non-linear dimensionality reduction.

The UMAP-enhanced DBSCAN removed all unclassified noise, decomposing the 492 fraudulent transactions into 25 behavioral micro-clusters. But the statistical reliability of testing expert models with micro-clusters in quite a restricted sample size is compromised. To ensure robust validation conditions, we proposed a dynamic *Centroid-Based Cluster Merging* protocol. We set a statistical criterion of at least 7 fraudulent cases in the test set of each cluster. Micro-clusters below this level were iteratively merged with their closest neighbors based on Euclidean centroid proximity. This aggregated the data into 6 extremely robust, behaviorally consistent macro-clusters (see Table 2).

Table 2. Final fraud macro-cluster distribution obtained via UMAP dimensionality reduction and centroid-based merging, ensuring statistical reliability.

Cluster ID	Total Fraud	Training Set	Validation Set	Test Set
Cluster 0	114 (23.2%)	79	18	17
Cluster 1	62 (12.6%)	43	10	9
Cluster 2	125 (25.4%)	87	19	19
Cluster 3	86 (17.5%)	60	13	13
Cluster 4	58 (11.8%)	40	9	9
Cluster 5	47 (9.5%)	32	8	7
Total	**492 (100.0%)**	**341**	**77**	**74**

5.3 Multi-model Inference and Bootstrap Validation

The framework was tested after separate training of localized classifiers exploiting the routing technique described in Sect. 3.6. A key drawback of traditional imbalanced ensemble approaches is the unintended duplicate counting of majority class samples. Our deterministic routing is designed to avoid such leaks intentionally. For each test transaction, we embed it into the UMAP space and only send it to the nearest cluster expert.

We directly address potential concerns about metric unreliability on localized test sets by doing a thorough bootstrap evaluation (N = 1,000 iterations) for each cluster-specific expert. The bootstrap analysis produced consistently low standard deviations across macro-clusters (e.g., $\sigma = 0.03$ for Cluster 0 and $\sigma = 0.14$ for the most difficult Cluster 3), experimentally demonstrating that the reported F1-scores are statistically credible and not products of small sample numbers.

The aggregated performance is computed by adding the mutually exclusive confusion matrices of these validated experts as shown in Table 3. The overall number of evaluated genuine occurrences (TN: 42,645 + FP: 2 = 42,647) is exactly equal to the original isolated test set, which essentially proves the absence of double-counting.

Table 3. Cluster-wise selected expert configurations, bootstrap validation metrics ($\mu \pm \sigma$), and mutually exclusive confusion matrices. Data leakage is categorically prevented.

Cluster	Expert Configuration	Bootstrap F1	TP	FP	FN	TN
0	Random Forest + RUS	0.97 ± 0.03	16	0	1	9882
1	CatBoost + Normal	1.00 ± 0.00	9	0	0	5374
2	Random Forest + SMOTE	1.00 ± 0.00	19	0	0	10835
3	CatBoost + Normal	0.42 ± 0.14	4	2	9	7453
4	CatBoost + RUS	1.00 ± 0.00	9	0	0	5027
5	AdaBoost + Normal	0.92 ± 0.09	6	0	1	4074
Aggregated	**Multi-Model System**	**-**	**63**	**2**	**11**	**42645**

5.4 Cost-Sensitive Evaluation and State-of-the-Art Comparison

To adapt the suggested framework to real-world financial needs we compare it to the best-performing worldwide single-model baseline (Random Forest + CTGAN). For a fair comparison we evaluated the two architectures on the same recombined test set.

As illustrated in Table 4, the multi-model framework exceeds the best global baseline on all key criteria. The global baseline achieves an F1-score of 0.8905, and the multi-model architecture improves this to 0.9065. Importantly, with the

strong operational constraint of 90% precision, the multi-model design yields a Recall@90% of 0.8514, compared to 0.8243 for the global baseline. This elimination of false negatives directly leads into a greater Cost Savings of 85.10%. This illustrates that the explicit separation of diverse sub-patterns results in improved operational viability than monolithic designs (see Table 4)

Table 4. Optimal global single-model baseline vs. clustering-based multi-model performance, incorporating operational metrics.

Method	Precision	Recall	Recall@90%	F1-Score	Cost Savings (%)
Global Baseline (RF + CTGAN)	0.9683	0.8243	0.8514	0.8905	84.43
Multi-Model (Experts)	**0.9692**	**0.8514**	**0.8514**	**0.9065**	**85.10**

We also compare our performance with the state-of-the-art architectures on the same European 2013 benchmark in Table 5. The suggested combination of generative augmentation with cluster-aware routing results in a highly competitive architecture (see Table 5).

Table 5. Performance comparison against contemporary studies utilizing the European 2013 dataset.

Reference	Methodology	Reported F1-Score
Dal Pozzolo et al. (2015) [2]	RF + Under-Sampling	0.812
Yang et al. (2019) [8]	Federated Learning	0.835
Ghalwash et al. (2025) [15]	DBSCAN-augmented Ensemble	0.867
Proposed [16]	**CTGAN + UMAP + Multi-Expert**	**0.906**

5.5 Computing Environment and Computational Costs

The experimental pipeline was conducted on an NVIDIA RTX 3060 GPU for the generative deep learning components and on an Apple Silicon M-series architecture for the tree-based ensemble training.

The data pre-processing step, including UMAP dimensionality reduction, DBSCAN clustering and dynamic cluster merging took only 3.6 s. The full grid of 25 global combinations of baselines took 770.8 s to evaluate. On the other hand, the training and evaluation of all expert combinations strictly inside their localized macro-clusters was finished in 337.9 s. Thus, the total end-to-end execution time of the suggested multi-model pipeline was around 5.7 min. Fast execution shows strong computing efficiency and scalability to real-world systems.

5.6 Discussion

Empirical findings support that robust credit card fraud detection should take into account both class imbalance and behavioral heterogeneity at the same time. The application of CTGAN for oversampling the minority class led to stable improvements across gradient boosting frameworks by creating high-quality tabular synthetic samples that preserve complicated relations of features.

Moreover, the clustering study revealed that fraudulent activity is not a homogeneous issue. The architecture was able to remove noise and segregate the minority class into six viable macro-clusters using UMAP and centroid-based merging. By routing the transactions to the experts of each cluster, we removed the structural constraints of the monolithic global model, achieving an aggregate F1-score of 0.9065, which is highly competitive.

Threats to Validity Internal Validity: The same hyperparameter configurations were used for global baselines and multi-model studies to provide a controlled comparison. This, however, isolates the performance benefits directly related to our system, and an exhaustive independent hyperparameter optimization for each localized expert could potentially change the final performance margins.

External Validity and Data Dependency: The main weakness of the system is its dependence on static historical data. Finally, the exact cluster formations and generating benefits are inherently related to the structural features of the European 2013 dataset. In real-world contexts, the fraud methods are under continual concept drift [3]. The analytical paradigm is structurally robust and transportable. The static cluster definitions we provide here will vary over time.

6 Conclusion and Future Work

In this paper, we present a unified approach for jointly tackling two critical difficulties in credit card fraud detection, namely high class imbalance and behavioral heterogeneity. The proposed strategy is based on oversampling the minority class using CTGAN and a clustering-driven multi-model architecture. Experiments have shown that the use of CTGAN to generate high fidelity synthetic data improves the prediction capabilities of some tree based classifiers compared to traditional interpolation procedures. However, the different levels of improvement also suggest that generative augmentation should be combined with appropriate high-capacity downstream classifiers. Furthermore, the suggested multi-model framework achieves consistently better performance than a well-tuned global single-model baseline. UMAP dimensionality reduction and explicit centroid-based routing to cluster-specific expert models allow the system to detect both common and rare localized fraud trends. We demonstrate through an extensive empirical investigation that concurrent treatment of data-level imbalance and model-level heterogeneity confers substantial improvements

in operational viability, assessed by better F1-scores and cost-sensitive assessment measures. Further work will investigate the application of adaptive and incremental clustering approaches to deal with continual concept drift and evaluate the resilience of the framework for online, real-time streaming fraud detection scenarios.

Acknowledgments. The authors would like to thank Fibabanka management and its Research and Development (R&D) department for their support, domain insights, and resource provisions during the course of this work. The authors would like to thank Prof. Dr. Fatih Alagoz for his valuable and constructive feedback.

Disclosure of Generative AI and AI-assisted Technologies in the Writing Process. In the process of preparing this manuscript, the authors used generative AI technologies solely for language improvement, proofreading, and maintaining an academic tone. The scientific content, data analysis, and results of the research were not generated by AI techniques. After utilizing these tools, the authors carefully edited and evaluated the text and assume full responsibility for the correctness, originality, and integrity of the final publication.

References

1. Duman, E., Sahin, Y.: A comparison of classification models on credit card fraud detection with respect to cost-based performance metrics. In: Duman, E., Atiya, A. (eds.) Use of Risk Analysis in Computer-Aided Persuasion, NATO Science for Peace and Security Series E: Human and Societal Dynamics, vol. 88, pp. 88–99. IOS Press (2011)
2. Dal Pozzolo, A., Caelen, O., Johnson, R.A., Bontempi, G.: Calibrating probability with undersampling for unbalanced classification. In: 2015 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 159–166. IEEE (2015). https://doi.org/10.1109/SSCI.2015.33
3. Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., Bontempi, G.: Credit card fraud detection and concept-drift adaptation with delayed supervised information. In: 2015 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2015). https://doi.org/10.1109/IJCNN.2015.7280527
4. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intel. Res. **16**, 321–357 (2002). https://doi.org/10.1613/jair.953
5. Goodfellow, I.J., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems 27 (NIPS 2014), pp. 2672–2680 (2014)
6. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional GAN. In: Advances in Neural Information Processing Systems 32 (NeurIPS 2019), pp. 7333–7343 (2019)
7. Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96), pp. 226–231. AAAI Press (1996)

8. Yang, W., Zhang, Y., Ye, K., Li, L., Xu, C.-Z.: FFD: a federated learning based method for credit card fraud detection. In: Chen, K., Seshadri, S., Zhang, L.-J. (eds.) BIGDATA 2019. LNCS, vol. 11514, pp. 18–32. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23551-2_2
9. Shazeer, N., et al.: Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. In: Proceedings of the ICLR 2017 (2017). https://openreview.net/forum?id=B1ckMDqlg
10. Davis, J., Goadrich, M.: The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning, ICML 2006, pp. 233–240. ACM (2006)
11. Bolton, R.J., Hand, D.J.: Unsupervised profiling methods for fraud detection. In: Credit Scoring and Credit Control VII (2001)
12. Sahin, Y., Duman, E.: Detecting credit card fraud by decision trees and support vector machines. In: Proceedings of IMECS 2011 (2011)
13. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **21**(9), 1263–1284 (2009)
14. Phua, C., Lee, V., Smith, K., Gayler, R.: A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119 (2010)
15. Ghalwash, M.A., Abdelrazek, S.M., Eladawi, N.H., Ghalwash, H.A.: Enhancing credit card fraud detection using DBSCAN-augmented disjunctive voting ensemble. Sci. Rep. **15**(1), 39754 (2025). https://doi.org/10.1038/s41598-025-22960-w
16. Karakas, B.C.: ICCSA2026-hybrid-fraud-detection–244-. GitHub Repository (2026). https://github.com/BcKBA/ICCSA2026-Hybrid-Fraud-Detection--244-