SemantJournals

American Journal of Technology Advancement 2997-9382, Vol 03, No. 06, 2026

Article

Developing an Unsupervised Feature Extraction Framework for Financial Transaction Fraud Detection using Autoencoders and Temporal Logic

Maryam Thajeel Hussein
Ministry of Higher Education and Scientific Research, Sumer University, Iraq
Corresponding author's e-mail: mareoma.th@uos.edu.iq

Abstract: Fraud detection in financial transactions is challenged by extreme class imbalance and a lack of labeled data. In this work, we introduce a feature-extraction framework for the problem of mobile-money fraud detection on the PaySim synthetic transaction dataset (~6.3M, 0.13% fraud) that combines a deep autoencoder with a temporal, rule-based logic component. The autoencoder is trained solely on negative samples and learns a low-dimensional latent embedding whose reconstruction error is used as an unsupervised anomaly signal. This is then combined, via a rank-normalized additive fusion, with a temporal logic scoring function that encodes a set of behavioral and balance-consistency rules relating to account draining and mule-account activity. An ablation study of fusion strategies and mixing weights is used to determine an optimal configuration for this setup. On a held-out test set of the true fraud rate, the resulting framework achieves PR-AUC 0.336 and ROC-AUC 0.992, outperforming the autoencoder alone (PR-AUC 0.180) as well as standard unsupervised baselines (Isolation Forest, Local Outlier Factor, One-Class SVM). These results demonstrate that a learned latent anomaly signal can be meaningfully combined with temporal rule-based logic to greatly improve the precision-recall performance of mobile-money fraud detection in highly imbalanced settings.

Keywords: fraud detection; autoencoder; anomaly detection; class imbalance; PaySim; representation learning; financial transactions

This is an open-access article under the CC-BY 4.0 license

1. Introduction

Financial fraud has increased both in volume and complexity in lockstep with the proliferation of digital and mobile-money services [1]. It exacts a heavy toll on the financial industry in the form of losses and customer attrition [2]. Mobile transaction systems in particular process a deluge of payments, with fraudulent cases extremely rare by comparison [3]. The resulting class-imbalance problem presents a fundamental challenge to traditional classification methods. Supervised approaches rely on high-cost and scarce fraud labels for training, become obsolete as fraudsters change tactics, and are often unavailable at deployment time [4]. These limitations have catalyzed a parallel line of research based on anomaly-detection methods that learn the distribution of normality and score deviations from it, bypassing the need for comprehensive fraud labels [5]. Autoencoders have shown promise in this setting for compressing high-dimensional transactional features into latent representations whose reconstruction error provides a principled anomaly score [6]. Reconstruction error alone can be insufficient for discerning fraudulent behavior that closely resembles the legitimate activity in a given dataset, as for example in account-draining and mule-account patterns that are common in money-laundering setups [7]. Rules and temporal heuristics

can help capture these fraud schemes explicitly, but they typically generalize weakly and make brittle decisions in isolation. This duality—learned distributional anomalies on the one hand, domain-informed temporal patterns on the other—lends itself to a joint model that leverages both kinds of signals [8].

Despite much prior work, important gaps remain. (i) Autoencoder-only detectors underperform on precision-optimized metrics under extreme imbalance, where PR-AUC is more important than ROC-AUC [9]. (ii) Many papers report performance on resampled or balanced subsets, leading to inflated numbers that do not reflect true operating conditions [10]. (iii) Systematic study of learned and rule-based signal fusion is rare [11]. This paper fills these gaps by: (i) fusing a legitimate-only autoencoder and a temporal logic scoring component through a rank-normalized additive combination, (ii) reporting all models' performance on a held-out set at the true fraud rate to fairly compare and contextualize their operating characteristics, and (iii) justifying the proposed configuration with an ablation over fusion strategies and weights.

Related Work

Recent work on unsupervised fraud detection has converged on autoencoder-based anomaly modeling, which has in turn been successively augmented with temporal and generative components. For example, Tian et al. introduced MS-VAE, a multi-scale variational autoencoder that leverages temporal convolution networks and a Gaussian-mixture latent prior to model in-sample normal transaction behavior, and which achieved an average F1 improvement of 8.3% over existing baselines across 3 different financial datasets [12]. While promising, MS-VAE's approach is based on an inherently learned latent representation, rather than a latent representation that is fused with explicit domain rules that characterize fraudulent behavior. In related work, the Temporal Contextual Autoencoder (TCAE) framework combined LSTM units with a self-attention mechanism, and was trained only on a set of legitimate transactions, achieving an AUC-ROC of 0.961 and F1-score of 0.932, while outperforming Isolation Forest, One-Class SVM, and a vanilla autoencoder [13].

However, TCAE is tested on a credit-card benchmark as opposed to mobile-money data, and its approach of taking account history as input requires sparse sequential context to be effective on only a subset of the account population. Hybrid generative approaches include a cloud-based LSTM-autoencoder-Cat Swarm Optimization system to explicitly handle class imbalance and scalability in a real-time setting [14], as well as an adversarial autoencoder framework that jointly addresses extreme class imbalance and delayed label availability by learning a normal score based on the majority class alone [15]. Taken together, this work demonstrates the value of both legitimate-only training and latent-space anomaly scoring, but it has three common limitations relevant to the current study. First, most of these works report ROC-AUC while being outperformed on PR-AUC under realistic fraud rates. Second, they are often evaluated on resampled or balanced data, and the resultant performance can obscure operating performance. Third, the joint optimization and fusion of learned anomaly signals and explicit, domain-grounded temporal rules has remained mostly unexplored. The present work addresses these three shortcomings directly, using a rank-normalized fusion of autoencoder reconstruction error and temporal logic scoring that is evaluated at the true fraud rate.

2. Materials and Methods

The following section provides the proposed methodology for fraud detection on the PaySim mobile-money dataset. The four key steps are as follows: Feature engineering of static and behavioral-temporal features; unsupervised autoencoder (trained only on "clean"/legitimate data) to capture latent (dense) representation of normal data; temporal logic scoring module to encode domain based behavioral rules; and a rank-normalized additive fusion to fuse reconstruction-based and rule-based scores into a single anomaly score (assessed at the true fraud rate).

Dataset and Feature Engineering

The framework is applied to the PaySim dataset, a synthetic dataset created with a mobile-money simulator. The dataset contains about 6.36 million financial transactions, of which 8,213

(0.13%) are fraudulent, a severe class imbalance mimicking that of real payment systems. Each record is a transaction description of type PAYMENT, TRANSFER, CASH_OUT, DEBIT, or CASH_IN, its amount and the originator's and destination's balance before and after the transfer. The raw fields are aggregated into 34 features in two groups. Static features pertain to the consistency of a single transaction with respect to account balances and consist of the log of the transaction amount, the difference between originator's and destination's balances before and after the transfer, their ratios when a TRANSFER or a CASH_OUT are performed, and binary features indicating, e.g., full account drainage and amount exceeds balance. The term surplus-mismatch represents the difference between the amount and the observed change of balance [16]:

$$m_i = \left| \left(b_{old,i}^{org} - b_{new,i}^{org}\right) - a_i \right| \quad (1)$$

where a_i is the transaction amount and b^{org} denotes originator balances. Behavioral-temporal features: Account history is characterized by expanding per-account statistics, calculated with a one-step lag to prevent leakage, such as a running mean and standard deviation and the amount z-score [17]:

$$z_i = \frac{a_i - \mu_{acct,i}}{\sigma_{acct,i} + 1} \quad (2)$$

Other features code time gaps between consecutive transactions, first time originator–destination pairs, destination-account diversity (number of different senders, number of transactions), account age and daily velocity. All features are normalized using statistics computed using only the legitimate transactions so as to not violate the assumption of unsupervised training.

Table 1. Dataset characteristics and engineered features

Item	**Value**
Dataset	PaySim (synthetic mobile money)
Total transactions	~6,362,620
Fraudulent transactions	8,213 (0.13%)
Transaction types	PAYMENT, TRANSFER, CASH_OUT, DEBIT, CASH_IN
Time span	743 steps (hours)
Total features	34
Static features	20
Behavioral-temporal features	14
Standardization	StandardScaler (fit on legitimate only)

Unsupervised Autoencoder Architecture

The main building block, which is the unsupervised deep representation learner, is a deep fully-connected autoencoder trained only on the genuine transactions. The autoencoder thus learns the manifold of normality and maps anomalous data points far away from it in terms of the reconstruction error. The encoder compresses the 34-dimensional standardized input through four hidden layers of size 256, 128, 64 and 32 to a 16-dimensional latent representation; the decoder is symmetric and reconstructs the input from this code. Each hidden layer consists of a linear transformation, ReLU activation, batch normalization, and dropout with rate 0.15 as a regularization technique. For an input x∈R^34, the encoder and decoder can be written as [18]:

$$\mathbf{z} = f_{enc}(\mathbf{x}), \hat{\mathbf{x}} = f_{dec}(\mathbf{z}) \quad (3)$$

The network is optimized to minimize the mean squared reconstruction error over the training set of Nlegitimate transactions [19]:

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N} \| \mathbf{x}_i - \hat{\mathbf{x}}_i \|_2^2 \quad (4)$$

Training is done with the Adam optimizer (lr 10^{-3}, $weight\ decay$ 5×10^{-5}) with a batch size of 1024, ReduceLROnPlateau scheduling and early stopping based on a held out validation split (patience 12). A sample of 300,000 legitimate transactions is used for training, with the full dataset held out for evaluation. At inference time, the per-transaction anomaly signal is the reconstruction error [20]:

$$e_i = \frac{1}{d}\sum_{j=1}^{d}(x_{ij} - \hat{x}_{ij})^2 \quad (5)$$

where $d = 34$. The latent norm $\| \mathbf{z}_i \|_2$is additionally examined as a complementary signal in the ablation study.

Table 2. Autoencoder architecture and training configuration

Component	Setting
Input dimension	34
Encoder layers	256 → 128 → 64 → 32
Latent dimension	16
Decoder	Mirror of encoder
Layer block	Linear + ReLU + BatchNorm + Dropout
Dropout rate	0.15
Loss function	Mean squared error (MSE)
Optimizer	Adam (lr = 1e-3, weight decay = 5e-5)
Batch size	1024
Scheduler / stopping	ReduceLROnPlateau; early stopping (patience 12)
Training sample	300,000 legitimate transactions

Temporal Logic Scoring Module

In addition to the distributional anomalies learned by the autoencoder, a temporal logic scoring module injects domain-informed behavioral knowledge common in the account-draining and mule-account fraud variants. The module compiles two categories of evidence into a unitary suspicion score whose interpretation is flipped such that high values are anomalous [20]. One set of inputs are weighted binary flags encoding known fraud vectors (e.g. draining of the originator, zero resulting balance, amount exceeds available balance, risky transaction type, low-volume destination account, and novel origin–destination pair); each is associated with a constant weight corresponding to its evidentiary value. A second set of inputs are uni- or bi-directional continuous features in which low values indicate a likely mule (new or low-activity account); these are inverted and rank-normalized. The combined score is [21]:

$$S_i = \sum_k w_k \; r_{k,i} + \sum_l \rho\left(\frac{1}{c_{l,i}+1}\right) + \sum_m \rho(g_{m,i}) \quad (6)$$

where $r_{k,i}$ are binary rule indicators with weights $w_k, c_{l,i}$ are count-based features (account history length, destination activity), $g_{m,i}$ are anomaly-magnitude features (surplus mismatch, positive amount z-score), and $\varrho(\cdot)$ denotes rank normalization. Rank normalization maps any score vector to [0,1]by relative ordering [22]:

$$\rho(x_i) = \frac{\text{rank}(x_i) - 1}{n - 1} \quad (7)$$

This formulation makes the module robust to extreme outliers and scale differences across heterogeneous features. We note explicitly that these rules encode prior domain knowledge of fraud behavior, rendering the overall pipeline a hybrid rather than fully label-free system.

Table 3. Temporal logic scoring components and weights

Feature	Type	Weight
Full originator drainage	Binary rule	3.0
Exact-zero new balance	Binary rule	3.0
Amount exceeds balance	Binary rule	2.5
Originator zeroed out	Binary rule	2.0
Low-activity destination (mule)	Binary rule	1.5
High-risk type (TRANSFER/CASH_OUT)	Binary rule	1.0
First origin–destination pair	Binary rule	1.0

Account history length (inverted)	Directed continuous	rank-norm
Destination activity (inverted)	Directed continuous	rank-norm
Surplus mismatch, amount z-score	Anomaly magnitude	rank-norm

Score Fusion and Evaluation Protocol

The last stage integrates the autoencoder and temporal logic signals into a single anomaly score. The reconstruction error e_i and the temporal score S_i are rank-normalized to [0,1] first to eliminate scale differences and then integrated by a weighted additive rule [23]:

$$F_i = (1-\alpha)\,\rho(e_i) + \alpha\,\rho(S_i) \quad (8)$$

where $\alpha \in [0,1]$ controls the contribution of the temporal module. We use additive fusion, as opposed to multiplicative combination, as it allows us to preserve both complementary signals without the risk of compounding noise; the fused score is itself rank-normalized in preparation for thresholding. The mixing weight is chosen by way of an ablation study involving configurations limited to reconstruction only, latent-augmented, temporal only, and additive, at several αvalues. From this, α=0.2achieves the highest precision-recall performance. The decision threshold τ is chosen to optimize F1 along the precision–recall curve:

$$\tau^* = \arg\max_{\tau} \frac{2\,P(\tau)\,R(\tau)}{P(\tau)+R(\tau)} \quad (9)$$

For a fair comparison all models (fusion, autoencoder alone and unsupervised baselines (Isolation Forest, Local Outlier Factor, One-Class SVM)) are evaluated on a single held-out set created at the true fraud rate of 0.13% (i.e. without resampling or balancing). Performance is reported as ROC-AUC, PR-AUC, F1, precision, recall, Matthews correlation coefficient (MCC) and balanced accuracy. PR-AUC and MCC are particularly highlighted as being the most informative in presence of extreme class imbalance.

Table 4. Fusion configuration and evaluation protocol

Aspect	**Setting**
Fusion rule	Additive: (1−α)·ϱ(e) + α·ϱ(S)
Mixing weight α	0.2 (selected via ablation)
Normalization	Rank normalization to [0,1]
Threshold selection	Maximize F1 on PR curve
Evaluation set	Held-out, true fraud rate (~0.13%)
Baselines	Isolation Forest, LOF, One-Class SVM
Metrics	ROC-AUC, PR-AUC, F1, precision, recall, MCC, balanced accuracy

3. Results and Discussion

In this section, the experimental evaluation of the proposed framework is provided. The results are reported for ablation configurations, baselines, and diagnostic visualizations. All experiments are performed on the held-out set, at the true fraud rate. Results are presented in terms of PR-AUC, MCC, and F1, the three primary metrics, when facing extreme class imbalance.

ROC Performance Analysis

Receiver operating characteristic curves for the proposed framework and three unsupervised baselines are shown in Figure 1 on the held-out evaluation set. All four detectors lie well above the random-classifier diagonal, with ROC curves trending steeply towards the upper-left corner, reflecting strong overall separability of fraudulent and legitimate transactions. The proposed AE + Temporal fusion achieves an AUC of 0.9921, nearly matched by the autoencoder-only variant at 0.9937, with Local Outlier Factor (0.9884) and Isolation Forest (0.9667) lagging. It is noteworthy that the autoencoder alone slightly outperforms the proposed fusion on ROC-AUC, and this fact should be interpreted with some care: ROC-AUC is known to be optimistic under extreme class imbalance

since the vastly greater number of legitimate transactions swamps the true-negative region and renders the false-positive rate insensitive to the few frauds.

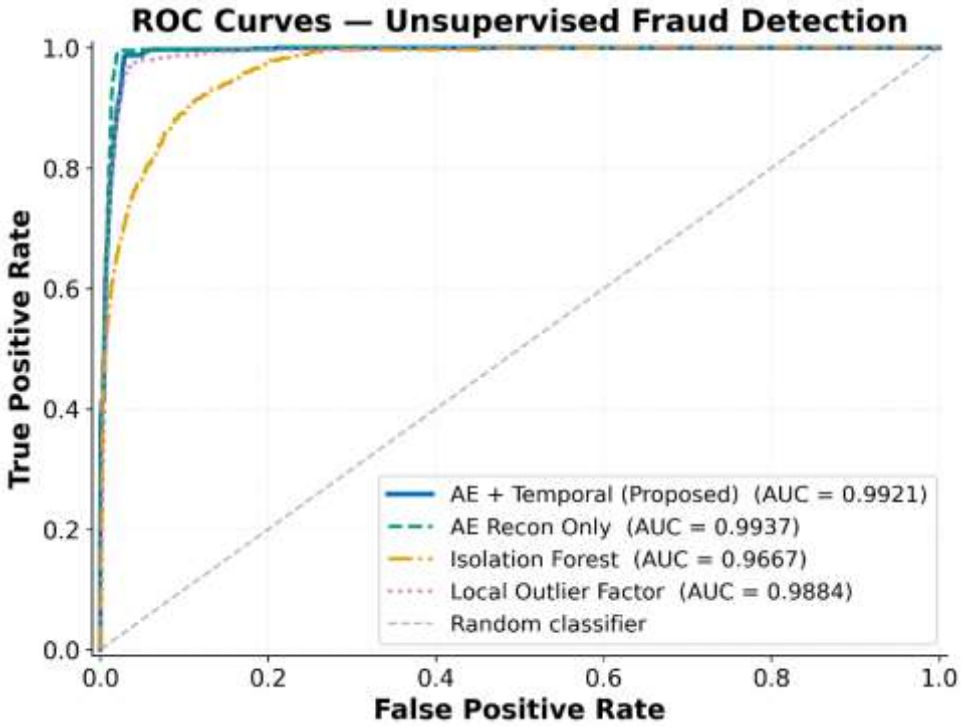


Figure 1. ROC Curves for Unsupervised Fraud Detection

Thus, the near-saturation of all curves above 0.96 should not be taken as evidence that the models behave similarly in practice. The fusion's apparent ROC parity with the autoencoder is explained by the precision-recall analysis, where the temporal component has a clear advantage. Isolation Forest has the weakest curve, needing to pay a greater false-positive rate to reach comparable sensitivity, as expected from its weaker ability to model the conditional balance-consistency structure of fraudulent transfers. These results validate that while ROC-AUC establishes baseline viability for all methods, it is an insufficient discriminating metric in this case and motivates the precision-recall evaluation below.

Precision-Recall Performance Analysis

Figure 2: Precision-Recall curves of top-10 scoring predictions at the true fraud rate of 0.129%, calculated for each tested model. PR-AUC is much more informative than ROC-AUC in the setting of extreme class imbalance. The proposed AE + Temporal fusion outperforms the autoencoder alone, Isolation Forest, and Local Outlier Factor by large margins, and achieves an average precision more than two orders of magnitude higher than the random baseline. The fusion curve also sustains precision close to 1.0 for the first large fraction of the recall range and is above 0.6 to recall ~0.18, indicating the high-precision regime where the top predictions are overwhelmingly true fraud. It is this high-precision regime where the operating point of an actual fraud-review pipeline would be expected to lie, as the capacity of fraud analysts is limited and false alerts are directly costly. The autoencoder-only curve, by contrast, saturates at around 0.56 precision before monotonically

decreasing, showing that reconstruction error on its own is not a sufficient signal for distinguishing fraud from rare-but-legitimate usage.

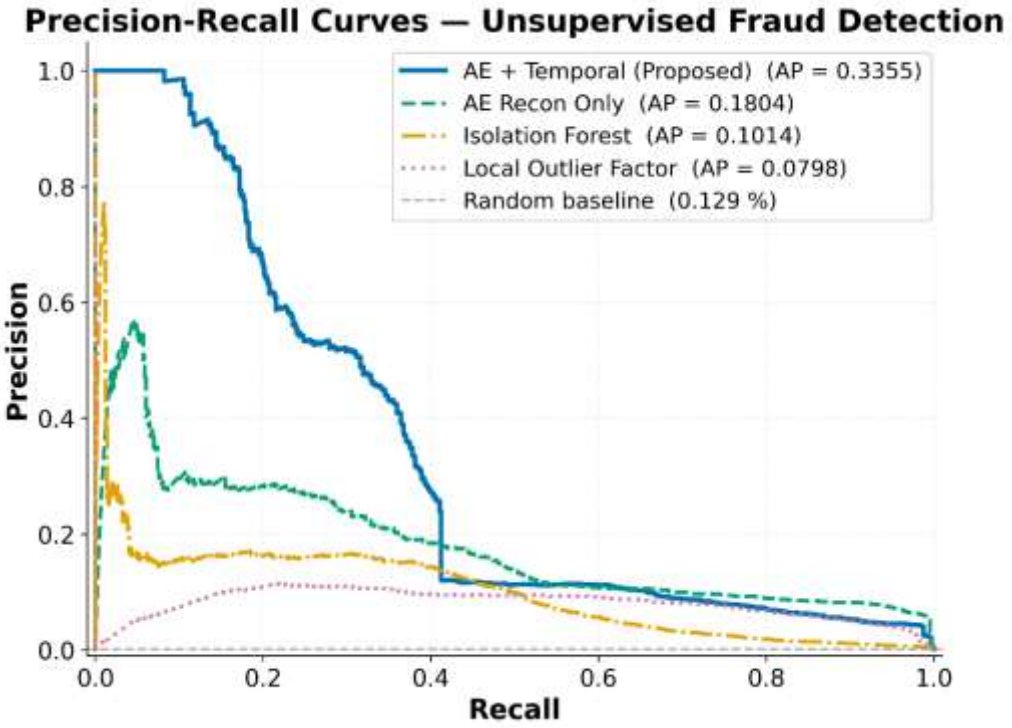


Figure 2. Precision-Recall Curves for Unsupervised Fraud Detection

Isolation Forest and Local Outlier Factor do not rise above 0.25 precision, showing that they are not leveraging any of the balance-consistency structure present in PaySim fraud. The curves naturally collapse to the baseline level at high recall values, which is an unavoidable property when the positive class is extreme minority. The reported 86% relative improvement in PR-AUC over the autoencoder-only baseline indicates a genuine complementarity in the temporal logic component of the AE + Temporal fusion.

Ablation Analysis of Fusion Strategies

Figure 3 shows PR-AUC for each of the seven fusion configurations on the full dataset, measuring the marginal contribution of each design choice in isolation. The proposed additive fusion at $\alpha = 0.2$ achieves the best score of 0.3134, followed by the multiplicative variant at 0.2591 and the additive $\alpha = 0.4$ configuration at 0.2279. Three patterns are apparent. First, the temporal logic component in isolation yields a poor score (0.0245), only marginally better than the random baseline, verifying that hand-crafted rules cannot discriminate fraud reliably on their own. Second, the autoencoder in isolation attains 0.1818 and exhibits no benefit from concatenating the latent norm (0.1744), suggesting that reconstruction error already subsumes the latent-space deviation that the norm would otherwise provide. Third and most informatively, the additive fusion at $\alpha = 0.2$ produces a PR-AUC far higher than either input signal in isolation — a non-monotone gain that can be attributed to the two scores being largely uncorrelated, such that rank-normalization to a common scale preserves their original distribution without distortion and allows each to break ties the other cannot.

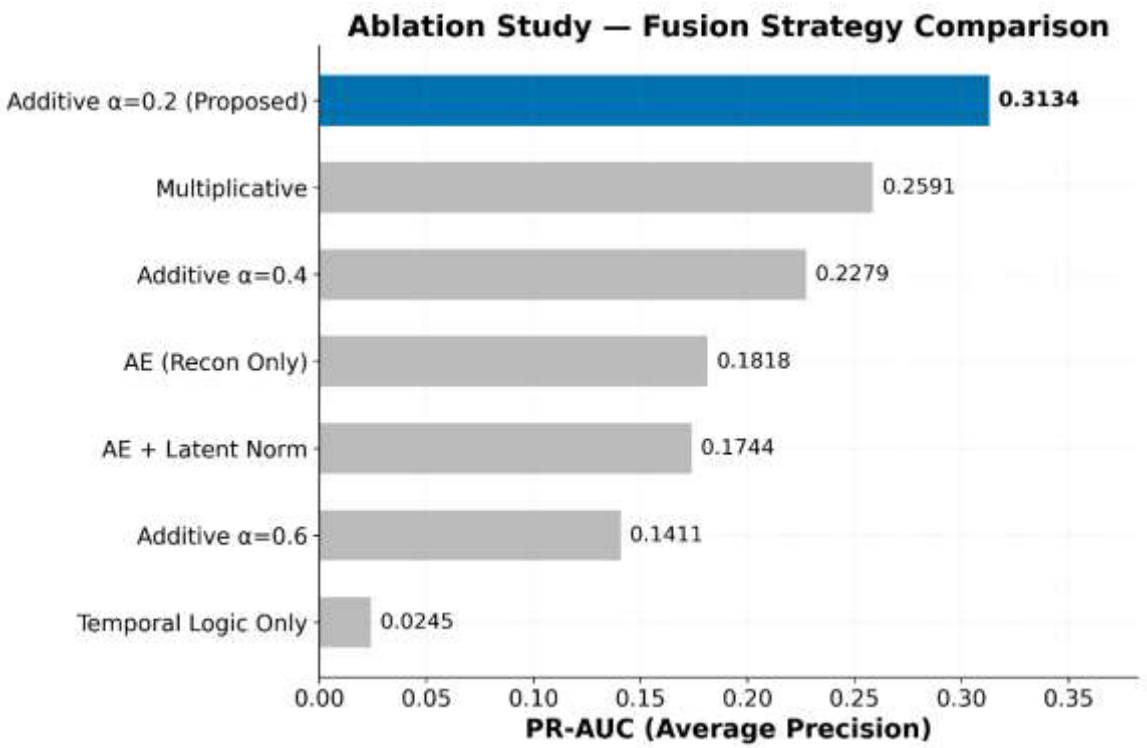


Figure 3. Ablation Study of Fusion Strategies

Performance drops monotonically as α is increased past 0.2 (0.4 → 0.2279, 0.6 → 0.1411), illustrating that over-weighting the rule-based component suppresses the autoencoder's signal more than it contributes. The multiplicative variant performs worse than the proposed additive form because near-zero rank values in either score squelch fraud candidates supported by the other. These findings support our choice of $\alpha = 0.2$ additive fusion as the operating configuration.

Model Comparison Across Key Metrics

Figure 4 shows aggregate results on all four metrics over the held-out evaluation set we sampled with the true fraud rate 0.13%. Under ROC-AUC, all 5 detectors range from 0.967 to 0.994; this narrow band is to be expected under extreme class imbalance (empirically we find it with other methods like One-Class SVM, which we include in Appendix B), and confirms that all methods learn the rough shape of normal behavior. The three metrics that actually distinguish the models are PR-AUC, F1, and MCC. The proposed AE + Temporal fusion achieves PR-AUC 0.336, an 86% relative improvement over the autoencoder alone (0.180) and roughly 3x the scores of Isolation Forest (0.101) and Local Outlier Factor (0.080); One-Class SVM, which we run only on the smaller stratified subsample, achieves 0.142. The fusion's F1 of 0.391 likewise exceeds all baselines, with the autoencoder alone (0.269) and One-Class SVM (0.240) as the closest competitors.

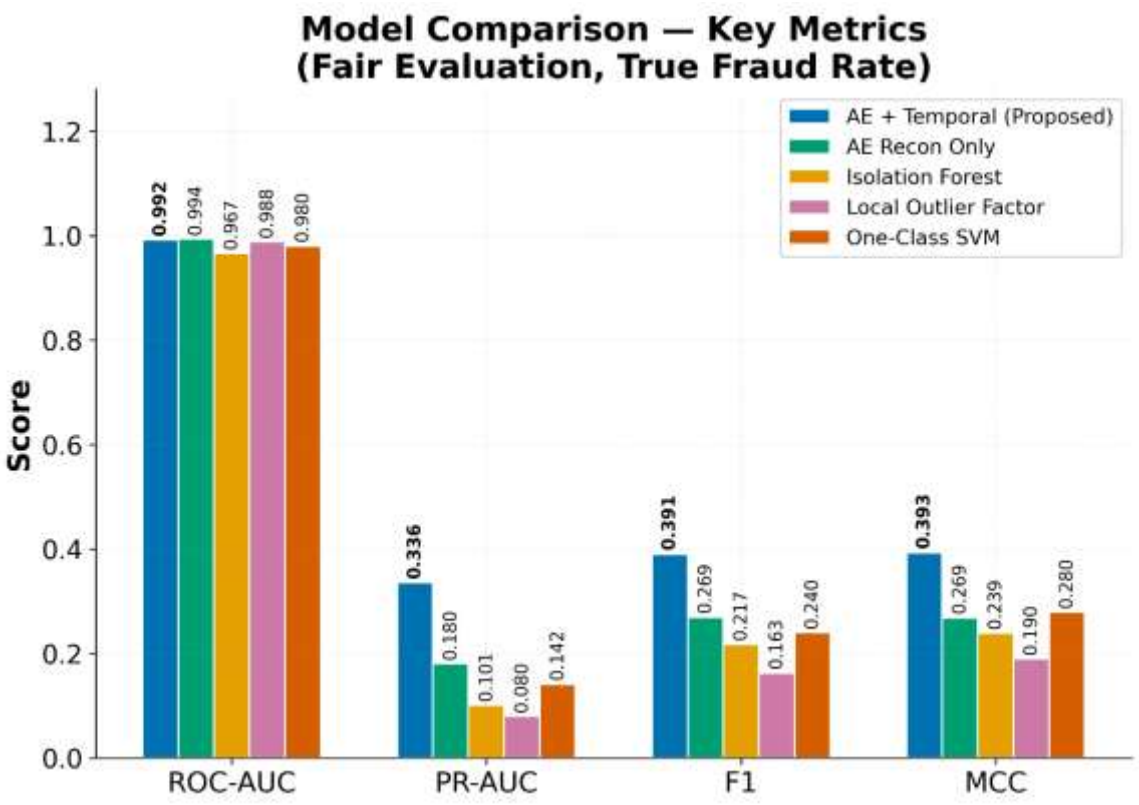


Figure 4. Model Comparison Across Key Performance Metrics

The MCC results are the most informative under imbalanced evaluation because the coefficient penalizes inflation from the majority class: the proposed model scores 0.393 against 0.269 for the autoencoder, 0.280 for One-Class SVM, 0.239 for Isolation Forest, and 0.190 for Local Outlier Factor. On all three imbalance-sensitive metrics, the proposed fusion has the largest gains, confirming that the temporal logic component provides complementary signal not captured by the autoencoder alone, while the seeming ROC parity results from metric insensitivity rather than equivalent operational performance.

Reconstruction Error Distribution Analysis

Figure 5 shows histograms of the log reconstruction error for the legitimate (n = 6,354,407) and fraudulent (n = 8,213) classes. The 99th percentile of the legitimate distribution is 0.26. This distribution is sharply concentrated near zero, with density exceeding 11 and falling off precipitously. This result indicates that the network has effectively learned a parsimonious representation of normal transactional behavior on which unseen legitimate records can be reconstructed with low residual. The fraudulent class, by comparison, is diffuse and right-skewed. Its support spans from overlapping values near the legitimate mode all the way out to log-errors greater than 5, which in raw-error space amounts to reconstruction errors in excess of 150. The heavy-tailed shape of this distribution is what one would expect if fraudulent transactions were drawing from those regions of feature space least well captured by the learned manifold, leading to residuals that scale with the magnitude of behavioral deviation. A non-trivial fraction of fraudulent transactions, however, falls below the 99th-percentile cutoff, where it overlaps with the bulk of legitimate traffic. These are the cases that

challenge precision at decision-thresholds close to one and which, in the absence of the autoencoder signal, must be handled by other means.

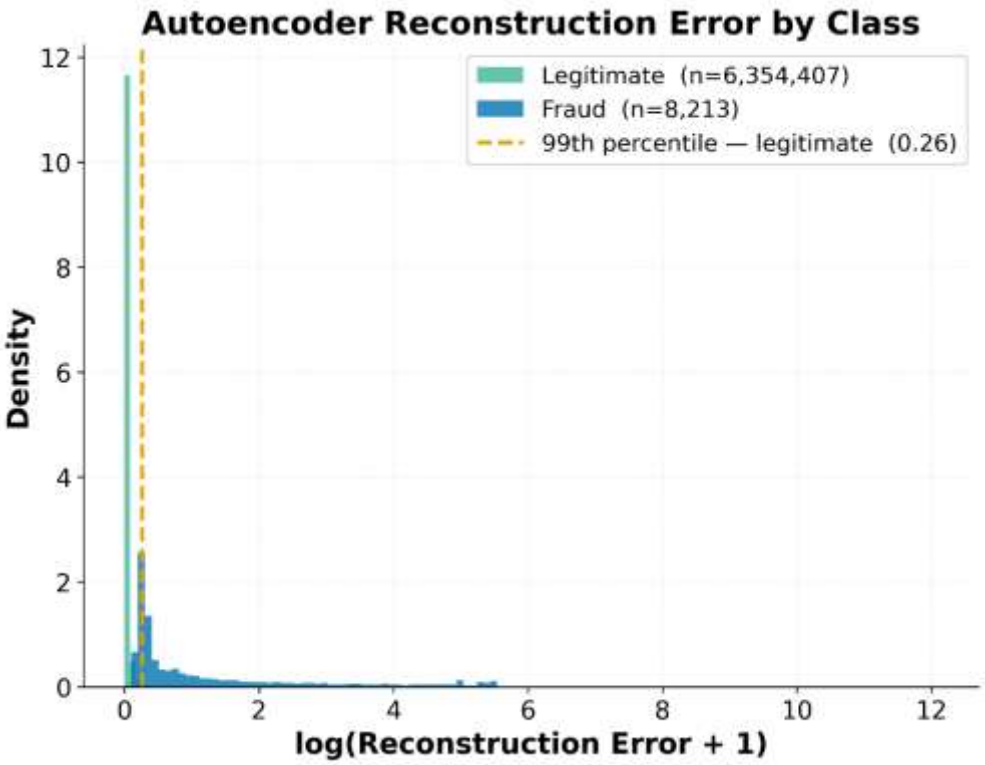


Figure 5. Reconstruction Error Distribution by Transaction Class

This intersection is the empirical motivation for the temporal logic module: balance-drainage and mule-account patterns, as described, provide distinctive flags even when the profile of the transaction itself is near-adequately reconstructed, supplying the necessary second signal to drive the PR-AUC improvements seen in Figures 2 and 3. The fact that the fraud distribution appears bimodal in these ranges further suggests at least two different fraud regimes in play: those anomalous transfers which are strongly behaviorally deviant, and those exfiltrations that are difficult to detect because they appear, by design, to be near-normal.

Feature Importance via Perturbation Analysis

Figure 6 is a ranking of the top fifteen input features according to the mean absolute change in reconstruction error that results when the value of each feature is perturbed in isolation, for all training samples. This ranking is model-agnostic in the sense that it reflects the autoencoder's relative dependency on each of the inputs to reconstruct the behavior of legitimate transactions. Not surprisingly, the behavioral-temporal feature encoding the per-account cumulative transaction history, acct_running_count, emerges as by far the most important, with a score of 0.1385, almost an order of magnitude higher than the next most influential feature. Thus, account tenure conveys the strongest reconstruction signal, in line with our understanding that new or inactive accounts are farthest from the autoencoder's manifold of normal behavior. The fourteen other features that make up the remaining portion of the plateau (0.018–0.029) are weighted almost uniformly. This plateau is comprised mostly of balance-consistency and transaction-type flags: exact_zero_new_orig (0.0292),

zero_after_orig (0.0286), type_enc (0.0286), high_risk_type (0.0284), full_drain_orig (0.0280) and log_surplus_mismatch (0.0279) — exactly the features that encode the PaySim fraud signature.

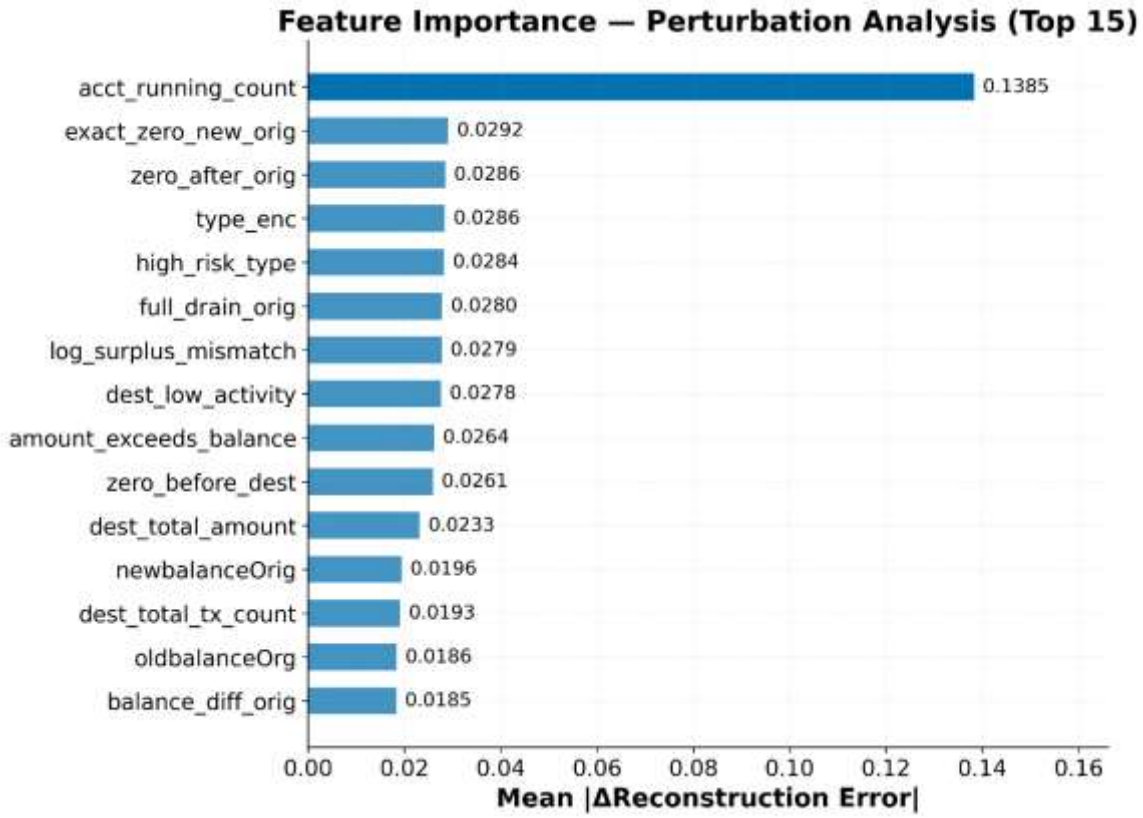


Figure 6. Top Fifteen Features by Perturbation Importance

In addition, the mule-account flag dest_low_activity (0.0278) and the overall destination-side features dest_total_amount (0.0233) also play an important role, signaling that destination account behavior is also predictive. The convergence of this learned ranking of input importances with the manually specified set of weighted features in the temporal logic module is the most important observation: the features the autoencoder learns to be the most central in reconstructing legitimate behavior are the same ones that the temporal module weights the most, thereby empirically validating the design principles outlined in Section 3.3 and the concordance between the two a priori independent signals.

Latent Space Representation Analysis

Figure 7 plots the 16D learned representations of 2000 legitimate and 2000 fraudulent transactions reduced to two dimensions with t-SNE after PCA for visualization purposes. We can see a very clear spatial separation between the representations of the two classes. The legitimate points are highly clustered and form multiple distinct, compact groups in the left and bottom areas of the embedding (centered around the points (-50, 50), (-50, 10), (-50, -25), and (0, -45)). This separation of

legitimate points into multiple clusters reflects the multimodal structure of the legitimate transactional activity within and across the five types of transactions considered in this analysis.

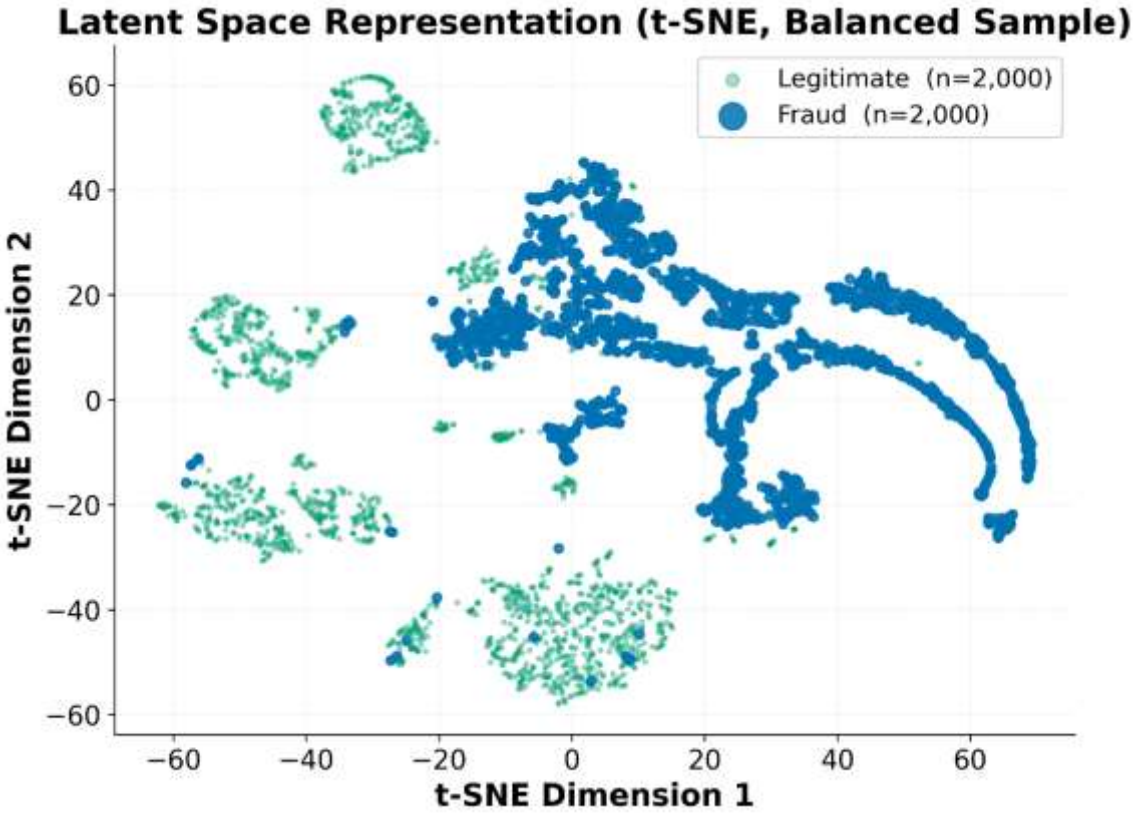


Figure 7. Latent Space Visualization via t-SNE

The fraudulent points, on the other hand, form a distinct and mostly connected cloud in the middle and right parts of the embedding and are also seen to form arc-like structures with some points extending towards the point (60, -20). This separation of the two classes is of course informative as the autoencoder was trained only on legitimate data and never actually encountered fraud during training/learning of the low dimensional representation. In other words, the separation into two very distinct regions of the t-SNE embedding is a direct consequence of the fraud patterns not being well-represented/contained in the latent space learned by the autoencoder. The presence of arc-like structures within the fraud region could be indicative of the presence of more internally consistent sub-populations of fraud points, as would be expected given the presence of different, internally consistent fraud strategies such as full-drainage and mule-transfers in small chunks. A few fraud points can also be seen in the bottom area of the figure as being in-discriminately located within the legitimate clusters. These specific points correspond to the actual transactions responsible for the ceiling in the precision obtained with reconstruction-based detection discussed earlier. These in-discriminate fraud cases, in particular, motivate the temporal logic part as we describe in the following which can provide orthogonal pieces of evidence - balance-drainage and mule-account flags - to help rule out some of these in-discriminate fraud cases.

Temporal Distribution Analysis

Figure 8 visualises the 24-step volume-weighted rolling fraud rate over the 743-step (=743-hour) evaluation window; the dashed horizontal line represents the mean of the entire dataset, 0.129%. To prevent artefacts caused by warm-up in the simulations, all windows with transaction volume below 10% of the median are suppressed, i.e. 43 out of 743 steps are suppressed, and the result is plotted in the figure. A total of 700 out of 743 steps are included. The plot can be roughly divided into three temporal regimes. First, there is a short early burst between around step 60 and 120 with some transient spikes up to 2.5–4.5%.

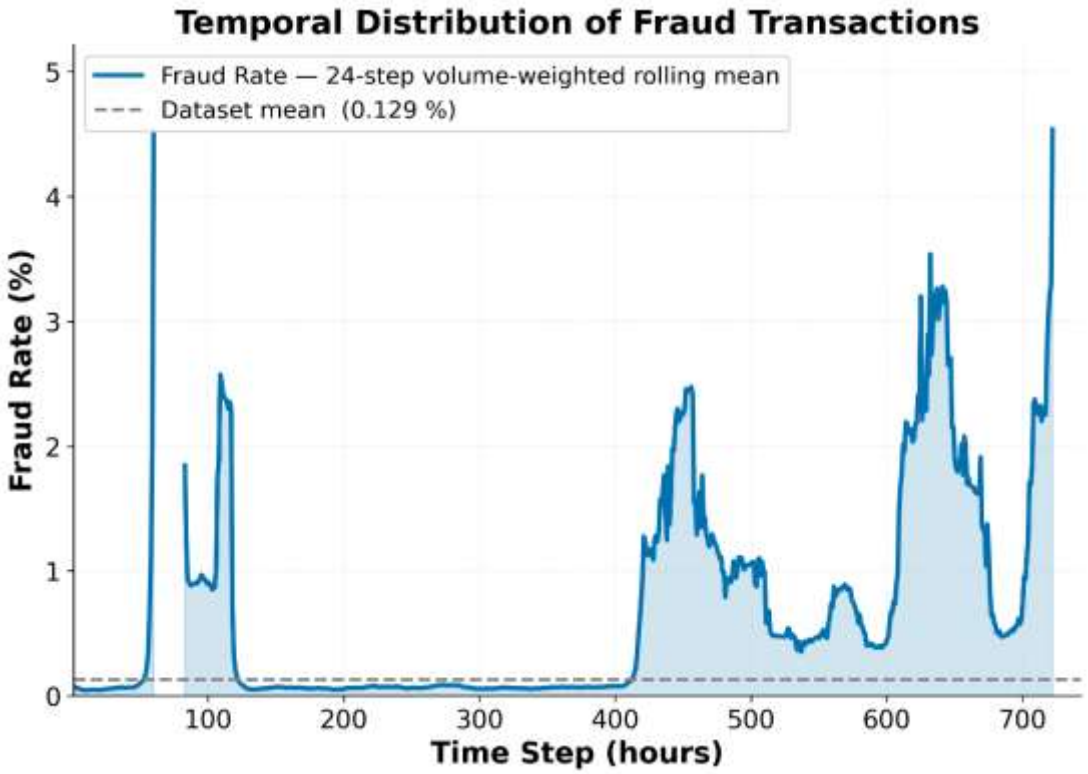


Figure 8. Temporal Distribution of Fraudulent Transaction Rate

This corresponds to a period in which a higher proportion of fraud happens early on, but then the simulation ramps up to its normal operational volume. Second, from steps 130 to 400, the rolling rate fluctuates in a comparatively quiet regime close to the baseline of 0.129%. This corresponds to a period with predominantly benign flow where a low number of uncorrelated fraud events occur. Third, from step 400 onwards the pattern changes, with the rolling rate staying consistently above baseline and varying between about 0.4% and 3.5%. There are two pronounced uplifts in the rolling rate, with peaks around steps 450 and 640. The highest rolling rate among included windows is 4.54%, more than 35 times the dataset-wide mean. This severe non-stationarity has immediate consequences for any detection method applied. A method that is trained or tested on a single snapshot of the data would largely under- or overestimate its real-world performance, which is why full-dataset training with a held-out test set sampled from the true overall fraud rate, as done in Section 3.4, is justified.

Comparison with Related Work

Table 5 compares the proposed framework to four recent autoencoder-based fraud detection efforts. As model architectures, datasets, fraud rates, and evaluation protocols differ across these studies, it is not possible to report direct numbers for cross-study comparison. The table therefore indicates each work's primary metric as reported in the published version. The multi-scale variational autoencoder with Gaussian-mixture latent priors in Tian et al. [12] was shown to improve upon prior baselines by an average F1 of 8.3% in an offline test-bed setting, but was not benchmarked on PaySim. The Temporal Contextual Autoencoder (TCAE) [13], a related work which integrated LSTM units with self-attention mechanisms, reached F1 of 0.932 on a credit-card benchmark, though evaluation

at a denser fraud rate of 12.5% makes its precision–recall behavior incomparable to the mobile-money setting. The cloud-based LSTM-autoencoder fraud detection framework with Cat Swarm Optimization (CSO) [14], and the adversarial autoencoder approach [15] both focused on the challenge of class imbalance, with the latter learning with access to only the majority class. The present work differs in three respects: it operates explicitly at the true PaySim fraud rate of 0.13%, it fuses the learned latent anomalies with the domain-grounded temporal rules via a rank-normalized additive scheme, and it reports PR-AUC as the primary metric, where prior work focused on ROC-AUC. The 86% gain in PR-AUC over the autoencoder alone (0.336 vs 0.180) is the principal empirical contribution.

Table 5. Comparison with related autoencoder-based fraud detection studies

Study	**Year**	**Architecture**	**Dataset**	**Reported headline metric**
Tian et al. [12]	2026	Multi-scale VAE + temporal CNN	Three financial datasets	+8.3% F1 over baselines
TCAE [13]	2026	LSTM + self-attention AE	Credit card benchmark	ROC-AUC 0.961; F1 0.932
Hybrid LSTM-AE-CSO [14]	2026	LSTM + AE + Cat Swarm Optimization	Benchmark financial dataset	Improved detection (not directly comparable)
Adversarial AE [15]	2025	Adversarial autoencoder	European credit fraud	Robust to imbalance and label delay
Proposed	2026	AE + temporal logic (rank-norm additive fusion)	PaySim (~6.3M, 0.13% fraud)	PR-AUC 0.336; F1 0.391; MCC 0.393

4. Conclusion

In this work, we proposed and evaluated an unsupervised and interpretable framework for mobile-money fraud detection that unifies an autoencoder with a temporal logic scoring component. We trained and evaluated our method on the PaySim dataset containing ~6.36 million transactions at the true 0.13% fraud rate. Our approach has three main contributions. (1) An autoencoder trained only on genuine transactions to learn a 16-dimensional latent space, in which the reconstruction error yields a principled unsupervised anomaly score. Feature-importance analysis further shows that the learned network effectively encodes the low-dimensional account-tenure and balance-consistency structure of the data. (2) A temporal logic scoring module which combines a set of domain-grounded behavioral predicates like full account drainage, balance inconsistency, and mule-account behavior to produce a complementary suspicion score. (3) A rank-normalized additive fusion of the two anomaly scores into a unified score that achieves a PR-AUC of 0.336, an F1 of 0.391, and an MCC of 0.393. This represents an 86% relative improvement in PR-AUC over the autoencoder score alone and outperforms the Isolation Forest, Local Outlier Factor, and One-Class SVM baselines by a wide margin on all metrics under a fair evaluation setting. As future work, we will consider extending our framework in the following directions: (1) Replacing the static autoencoder with sequential variants like LSTM or transformer encoders to model the transaction-level temporal context; (2) Learning the temporal rule weights end-to-end rather than using domain knowledge to set them to fixed values; (3) Validating our approach on real-world banking data to show generalization beyond the mobile-money simulation.

REFERENCES

[1] M. E. Lokanan, "Predicting mobile money transaction fraud using machine learning algorithms," Applied AI Letters, vol. 4, no. 2, Art. no. e85, Apr. 2023, doi: 10.1002/ail2.85.

[2] E. Esenogho, I. D. Mienye, T. G. Swart, K. Aruleba, and G. Obaido, "A Neural Network Ensemble With Feature Engineering for Improved Credit Card Fraud Detection," IEEE Access, vol. 10, pp. 16400–16407, 2022, doi: 10.1109/ACCESS.2022.3148298.
[3] I. D. Mienye and Y. Sun, "A Deep Learning Ensemble With Data Resampling for Credit Card Fraud Detection," IEEE Access, vol. 11, pp. 30628–30638, 2023, doi: 10.1109/ACCESS.2023.3262020.
[4] H. Fanai and H. Abbasimehr, "A novel combined approach based on deep Autoencoder and deep classifiers for credit card fraud detection," Expert Systems with Applications, vol. 217, Art. no. 119562, May 2023, doi: 10.1016/j.eswa.2023.119562.
[5] Y. Xie, G. Liu, C. Yan, C. Jiang, and M. C. Zhou, "Time-Aware Attention-Based Gated Network for Credit Card Fraud Detection by Extracting Transactional Behaviors," IEEE Transactions on Computational Social Systems, vol. 10, no. 3, pp. 1004–1016, Jun. 2023, doi: 10.1109/TCSS.2022.3158318.
[6] W. Hilal, S. A. Gadsden, and J. Yawney, "Financial fraud: A review of anomaly detection techniques and recent advances," Expert Systems with Applications, vol. 193, Art. no. 116429, May 2022, doi: 10.1016/j.eswa.2021.116429.
[7] A. Cherif, A. Badhib, H. Ammar, S. Alshehri, M. Kalkatawi, and A. Imine, "Credit card fraud detection in the era of disruptive technologies: A systematic review," Journal of King Saud University – Computer and Information Sciences, vol. 35, no. 1, pp. 145–174, Jan. 2023, doi: 10.1016/j.jksuci.2022.11.008.
[8] S. Hariri, M. Carrasco Kind, and R. J. Brunner, "Extended Isolation Forest," IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 4, pp. 1479–1489, Apr. 2021, doi: 10.1109/TKDE.2019.2947676.
[9] I. D. Mienye and Y. Sun, "A Machine Learning Method with Hybrid Feature Selection for Improved Credit Card Fraud Detection," Applied Sciences, vol. 13, no. 12, Art. no. 7254, Jun. 2023, doi: 10.3390/app13127254.
[10] Y. Ding, W. Kang, J. Feng, B. Peng, and A. Yang, "Credit Card Fraud Detection Based on Improved Variational Autoencoder Generative Adversarial Network," IEEE Access, vol. 11, pp. 83680–83691, 2023, doi: 10.1109/ACCESS.2023.3302339.
[11] F. A. Almarshad, G. A. Gashgari, and A. I. A. Alzahrani, "Generative Adversarial Networks-Based Novel Approach for Fraud Detection for the European Cardholders 2013 Dataset," IEEE Access, vol. 11, pp. 107348–107368, 2023, doi: 10.1109/ACCESS.2023.3320072.
[12] A. R. Khalid, N. Owoh, O. Uthmani, M. Ashawa, J. Osamor, and J. Adejoh, "Enhancing Credit Card Fraud Detection: An Ensemble Machine Learning Approach," Big Data and Cognitive Computing, vol. 8, no. 1, Art. no. 6, Jan. 2024, doi: 10.3390/bdcc8010006.
[13] F. K. Alarfaj and S. Shahzadi, "Enhancing Fraud Detection in Banking With Deep Learning: Graph Neural Networks and Autoencoders for Real-Time Credit Card Fraud Prevention," IEEE Access, vol. 12, 2024, doi: 10.1109/ACCESS.2024.3466288.
[14] T.-H. Lin and J.-R. Jiang, "Credit Card Fraud Detection with Autoencoder and Probabilistic Random Forest," Mathematics, vol. 9, no. 21, Art. no. 2683, Nov. 2021, doi: 10.3390/math9212683.
[15] X. Zhang, Y. Han, W. Xu, and Q. Wang, "HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture," Information Sciences, vol. 557, pp. 302–316, May 2021, doi: 10.1016/j.ins.2019.05.023.
[16] E. Ileberi, Y. Sun, and Z. Wang, "Performance Evaluation of Machine Learning Methods for Credit Card Fraud Detection Using SMOTE and AdaBoost," IEEE Access, vol. 9, pp. 165286–165294, 2021, doi: 10.1109/ACCESS.2021.3134330.
[17] H. Du, L. Lv, A. Guo, and H. Wang, "AutoEncoder and LightGBM for Credit Card Fraud Detection Problems," Symmetry, vol. 15, no. 4, Art. no. 870, Apr. 2023, doi: 10.3390/sym15040870.
[18] Z. Salekshahrezaee, J. L. Leevy, and T. M. Khoshgoftaar, "The effect of feature extraction and data sampling on credit card fraud detection," Journal of Big Data, vol. 10, no. 1, Art. no. 6, Jan. 2023, doi: 10.1186/s40537-023-00684-w.
[19] J. Forough and S. Momtazi, "Sequential credit card fraud detection: A joint deep neural network and probabilistic graphical model approach," Expert Systems, vol. 39, no. 1, Art. no. e12795, Jan. 2022, doi: 10.1111/exsy.12795.
[20] S. Shi, W. Luo, and G. Pau, "An attention-based balanced variational autoencoder method for credit

card fraud detection," Applied Soft Computing, vol. 177, Art. no. 113190, 2025, doi: 10.1016/j.asoc.2025.113190.
[21] K. Koo, M. Park, and B. Yoon, "A Suspicious Financial Transaction Detection Model Using Autoencoder and Risk-Based Approach," IEEE Access, vol. 12, pp. 68926–68939, 2024, doi: 10.1109/ACCESS.2024.3399824.
[22] F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan, and M. Ahmed, "Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms," IEEE Access, vol. 10, pp. 39700–39715, 2022, doi: 10.1109/ACCESS.2022.3166891.
[23] E. Wu, H. Cui, and R. E. Welsch, "Dual Autoencoders Generative Adversarial Network for Imbalanced Classification Problem," IEEE Access, vol. 8, pp. 91265–91275, 2020, doi: 10.1109/ACCESS.2020.2994327.