FraudGuard

Deep Anomaly Detection with Conformal Prediction

Artur GARIPOV **Quentin REBUT** **Sacha KOSKAS**
Haidar BOUAOUICHE

January 26, 2026

GitHub Repository

Abstract

Credit card fraud detection presents a unique challenge in deep learning due to extreme class imbalance ($< 0.2\%$ fraud). This project implements a two-stage solution combining a supervised baseline with an unsupervised Variational Autoencoder (VAE). To address the interpretability of anomaly scores, we employ Inductive Conformal Prediction (ICP) to provide rigorous statistical guarantees on false positive rates. Our system achieves robust detection (AUC 0.93 for VAE) while allowing operators to control operational risk strictly.

1 Introduction

1.1 Problem Statement

Credit card fraud detection presents a unique challenge in deep learning due to extreme class imbalance. In the provided dataset, fraudulent transactions constitute only 0.17% of the total volume (492 out of 284,807). Traditional supervised classifiers often achieve high accuracy by overfitting to the majority class (legitimate transactions) while failing to capture the minority class (fraud). Furthermore, standard anomaly detection methods typically output a raw "outlier score," leaving operators to guess an arbitrary threshold for blocking transactions.

This project implements a multi-stage solution:

1. **Supervised Baseline:** A weighted Deep Fully Connected Network (FCN) to establish an upper bound on performance using full label information.
2. **Unsupervised Detection:** A Variational Autoencoder (VAE) trained exclusively on legitimate transactions to learn the manifold of "normal" behavior.
3. **Statistical Calibration:** The application of Inductive Conformal Prediction (ICP) to the VAE's reconstruction error, replacing arbitrary thresholds with statistically guaranteed limits on the False Positive Rate (FPR).

1.2 Dataset and Preprocessing

The dataset consists of European credit card transactions from September 2013. The features include:

- **V1–V28:** Principal components obtained via PCA (anonymized).
- **Time:** Seconds elapsed since the first transaction.
- **Amount:** Transaction amount.
- **Class:** 0 (Legitimate) or 1 (Fraud).

As shown in the data sample, the PCA features (V1–V28) are roughly centered around 0 with unit variance, whereas Amount varies significantly (e.g., 149.62, 2.69, 378.66) and Time increases monotonically. To ensure stable gradient descent, we applied StandardScaler to Time and Amount to align their distribution with the PCA components.

The data was stratified into three splits:

- $\mathcal{D}_{main}$ (70%): Training (legitimate samples only for VAE).
- $\mathcal{D}_{calib}$ (10%): Calibration (strictly for computing conformal thresholds).
- $\mathcal{D}_{test}$ (20%): Evaluation.

2 Supervised Baseline

To benchmark the unsupervised model, we first trained a supervised binary classifier. This establishes the “gold standard” performance achievable when ground-truth fraud labels are available and provides an upper bound against which to compare anomaly-based methods.

2.1 Model Architecture and Training

We adopt a 3-layer Multi-Layer Perceptron (MLP), which is a standard choice for tabular data where interactions between features are global rather than spatial or temporal. As shown by Shwartz-Ziv and Armon (2022), fully connected architectures remain highly competitive on tabular data and do not necessarily benefit from convolutional or attention-based inductive biases unless the domain structure explicitly demands it.

- **Architecture:** A 3-layer MLP with topology

  $$\text{Input} \to 64 \to 32 \to \text{Output}(1)$$

  using ReLU activations. Rectified Linear Units were shown by Glorot, Bordes, and Bengio (2011) to alleviate vanishing gradients and improve optimization dynamics relative to sigmoid-type activations.

- **Regularization:** Given the extreme class imbalance (fraud $\approx 0.17\%$), preventing the model from memorizing minority samples is critical. We apply Dropout ($p = 0.3$), introduced by Srivastava, Hinton, Krizhevsky, Sutskever, and Salakhutdinov (2014), to encourage robust distributed representations and reduce co-adaptation between neurons. We further apply Batch Normalization, proposed by Ioffe and Szegedy (2015), to stabilize training and allow larger learning rates, which proved beneficial together with weighted classification.

- **Loss Function:** We employ `BCEWithLogitsLoss` with a positive class weighting factor of $w_{pos} \approx 577$. This increases the penalty for misclassifying a fraud instance, shifting the classifier toward higher recall—an appropriate risk preference for fraud detection tasks where the cost of false negatives is large.

2.2 Threshold Optimization and Results

Standard classifiers use a default decision threshold of 0.5. However, in highly imbalanced settings, this is rarely optimal. We performed "Threshold Moving" by analyzing the Precision-Recall curve to find the threshold that maximizes the F1-Score.

Performance Summary:

- **ROC-AUC:** 0.9758

- **Optimal Threshold:** ≈ 0.996 (High threshold due to aggressive weighting).

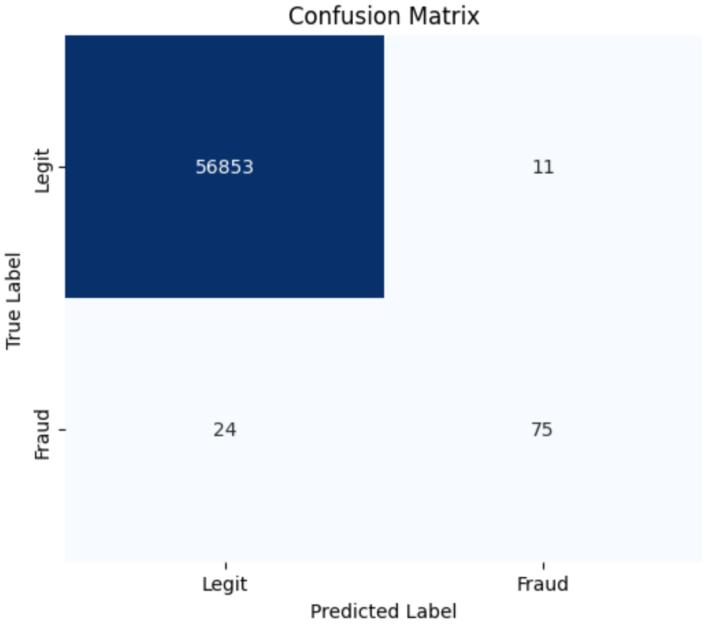


Figure 1: Confusion Matrix of the Supervised Baseline at the Optimal Threshold ($\tau \approx 0.99$). The model correctly identifies 75 frauds with only 11 false alarms out of nearly 57,000 transactions.

- **Recall (Fraud):** 76% (75/99 frauds detected in test set).
- **False Positives:** Only 11 legitimate users blocked.

While these results are good, they rely on the availability of labeled historical fraud data, which is often unavailable in real-time systems or subject to concept drift. This motivates the need for the unsupervised approach in Part 2.

3 Unsupervised VAE Architecture

The core component of this project is a Variational Autoencoder designed to model the distribution $P(X|y = 0)$. By compressing and reconstructing legitimate transactions, the model forces fraudulent transactions—which differ statistically—to have high reconstruction errors.

3.1 Model Specification

We evaluate two symmetric VAE configurations optimized for tabular data. Both models use Batch Normalization (BN), ReLU activations, and Dropout in the encoder as recommended by Vinay et al. (2021) and Obushnyi et al. (2025). The input consists of 30 standardized features (29 PCA components + amount).

- **Encoder: Small VAE (Vinay et al. inspired)**
  Input(30) → Linear(20) → BN → ReLU → Dropout → Linear(15) → BN → ReLU → Latent(4)

- **Encoder: Large VAE (Best configuration)**
Input(30) → Linear(64) → BN → ReLU → Dropout → Linear(32) → BN → ReLU → Latent(16)

- **Reparameterization:** Both architectures utilize the standard reparameterization trick,
$$z = \mu + \sigma \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I).$$

- **Latent Space:** Grid search was performed over latent_dim ∈ {4, 16}.

- **Decoder:** A symmetric decoder mirrors the encoder:
Latent(4/16) → Linear(15/32) → BN → ReLU → Linear(20/64) → BN → ReLU → Output(30).

- **Output Layer:** The output layer is linear (no sigmoid/tanh), as input features are standardized to $\mathbb{R}$ rather than bounded to $[0, 1]$.

The VAE was optimized using the Evidence Lower Bound (ELBO), originally introduced by Kingma and Welling (2014) and Rezende et al. (2014). The objective decomposes into a reconstruction term and a regularization term enforcing prior structure in the latent space. For a β-scaled formulation, the loss reads:

$$\mathcal{L} = \text{MSE}(x, \hat{x}) + \beta \cdot D_{KL}(q(z|x) \,\|\, p(z))$$

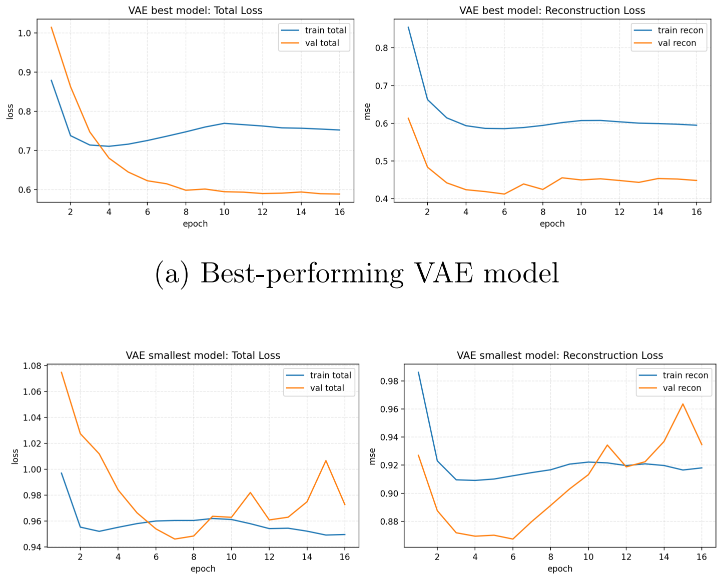
We employ KL annealing to mitigate posterior collapse (Bowman et al., 2016), gradually increasing β from 0 to 0.5 over the first 10 epochs. This encourages the encoder to learn informative latent representations before the regularization term becomes dominant, thereby balancing reconstruction fidelity and latent structure.

3.2 Training comparison

The smaller latent space introduces an information bottleneck on already PCA–compressed features, resulting in underfitting. In contrast, latent_dim = 16 preserves sufficient variation and yields better reconstruction performance.

Because the inputs are PCA components, most noise and redundancy are already removed prior to training. Thus, a very small latent space compresses an already compact representation and discards informative structure, whereas a moderately sized latent space behaves similarly to a shallow MLP reconstructing the input, which aligns with its superior performance in anomaly detection.

The training curves demonstrate healthy convergence without significant overfitting. The reconstruction loss on the validation set tracks the training set closely, indicating that the model generalizes well to unseen legitimate data.



(a) Best-performing VAE model

(b) Smallest-capacity VAE model

Figure 2: Training and validation loss curves for two VAE configurations. The best model (left) generalizes more effectively, while the smaller model (right) exhibits higher reconstruction error and slower convergence.

3.3 Threshold Optimization and Results

Similarly to previous model, We performed "Threshold Moving" by analyzing the Precision-Recall curve to find the threshold that maximizes the F1-Score.

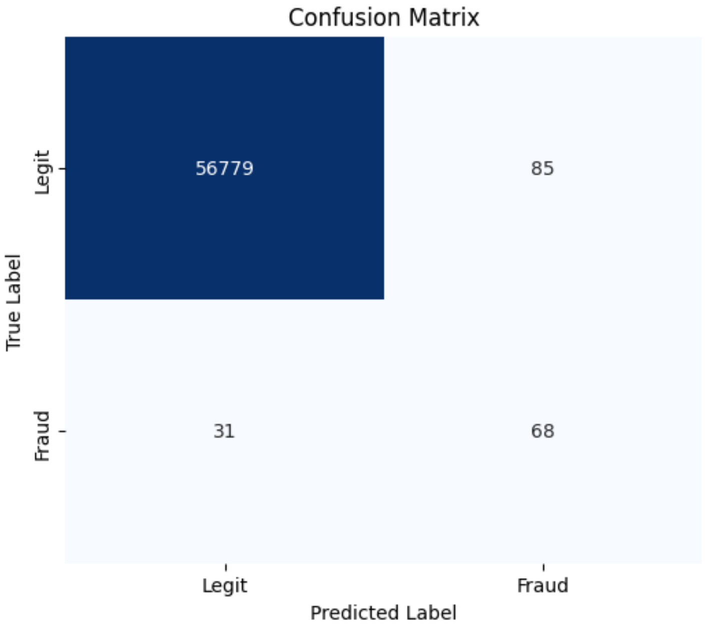


Figure 3: Confusion Matrix of the Unsupervised Baseline at the Optimal Threshold ($\tau \approx 4.19$). The model correctly identifies 68 frauds with 85 false alarms out of nearly 57,000 transactions.

Performance Summary:

- **ROC-AUC:** 0.9343
- **Optimal Threshold:** ≈ 4.19
- **Recall (Fraud):** 69% (68/99 frauds detected in test set).
- **False Positives:** 85 legitimate users blocked.

4 Statistical Calibration (Conformal Prediction)

Raw model outputs are not directly actionable for decision-making. In the supervised case, the model produces a probability (or logit) of fraud; in the unsupervised case, the VAE produces a reconstruction error. Neither quantity is calibrated to operational risk. For example, for the VAE, does an MSE of 5.0 indicate fraud? Or 10.0? For the supervised classifier, does a predicted probability of 0.7 mean it should be flagged?

To resolve this interpretability gap, we apply Inductive Conformal Prediction (ICP) following Angelopoulos and Bates (2021). ICP transforms arbitrary scores into thresholded decisions with statistically valid false positive guarantees.

4.1 Calibration Procedure

1. **Score extraction.** For each model, we compute conformity scores on a held-out calibration split $\mathcal{D}_{calib}$:
   - Supervised baseline: $S = 1 - p_{\text{fraud}}(x)$ (lower is more anomalous).
   - VAE: $S = \text{MSE}(x, \hat{x})$ (higher is more anomalous).
2. **Risk control via quantiles.** For a user-defined tolerance level α (e.g., 5%), compute
$$\hat{q}_\alpha = \text{Quantile}_{1-\alpha}(S)$$
which defines the operating threshold.
3. **Finite-sample guarantee.** Any future legitimate transaction will be flagged with probability at most α:
$$\mathbb{P}(\text{flag} \mid y = 0) \leq \alpha$$
providing a marginal coverage guarantee without distributional assumptions.

This formulation allows the system to operate according to explicit risk policy (e.g., “We can afford to audit 1% of legitimate transactions”) rather than arbitrary or heuristic thresholding. Importantly, ICP works uniformly for both supervised and unsupervised settings, enabling meaningful comparison and deployment across heterogeneous models.

5 Evaluation and Results

5.1 Anomaly Detection Performance

The supervised baseline achieved an ROC–AUC of 0.9758 on the test set, indicating excellent separation between fraudulent and legitimate transactions. However, because of the severe class imbalance, the Precision–Recall curve provides a more realistic assessment of performance on the positive (fraud) class, yielding a PR–AUC of 0.7067.

The VAE, which operates in an unsupervised manner, achieved an ROC–AUC of 0.9343 and a PR–AUC of 0.5051 on the test set. Although the separation in terms of ROC–AUC remains strong, the lower PR–AUC reflects the difficulty of identifying rare fraudulent events without access to labels during training.

5.2 Conformal Prediction Validity

To validate the calibration, we evaluated the thresholds derived from the calibration set on the unseen test set. The primary goal was to ensure that the Inductive Conformal Prediction (ICP) procedure strictly respects the user-defined False Positive Rate (FPR) limits. **VAE Results:** For the unsupervised VAE model, we performed comprehensive testing to verify conformal guarantees. The calibration thresholds were computed on legitimate-only calibration data and then evaluated on the full test set:

Target α	Threshold (τ)	Empirical FPR	Recall (TPR)	True Negatives	False Positives
0.10%	5.61	0.11%	55.56%	56,804	60
1.00%	1.61	0.98%	79.80%	56,305	559
5.00%	0.93	4.79%	81.82%	54,138	2,726
10.00%	0.71	9.54%	83.84%	51,440	5,424

Table 1: Conformal calibration and test results for the VAE model showing close alignment between target and empirical FPR. Values rounded to 2 decimal places.

The empirical FPR closely tracks the Target α, validating the conformal guarantees for the VAE model. The operating point at $\alpha = 1\%$ is particularly notable: it captures nearly 80% of all fraud while impacting less than 1% of legitimate users.

Supervised Baseline: For the supervised model, we computed conformal thresholds on legitimate-only calibration data:

- $\alpha = 0.10\%$: $\tau = 0.978$
- $\alpha = 1.00\%$: $\tau = 0.826$

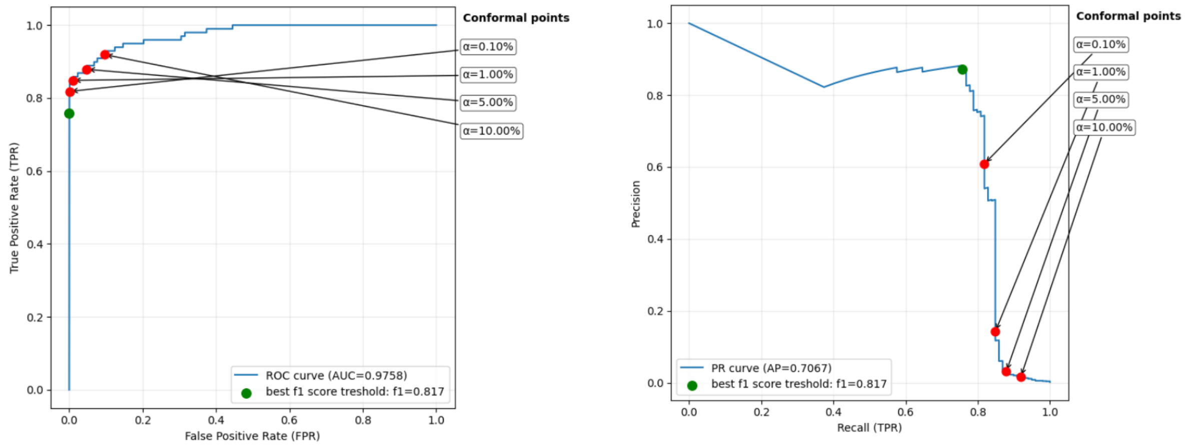
- $\alpha = 5.00\%$: $\tau = 0.376$
- $\alpha = 10.00\%$: $\tau = 0.156$

These thresholds demonstrate how conformal prediction can be applied to supervised models, though their practical interpretation differs from the unsupervised case since the supervised model was trained with access to fraud labels.

5.3 Comparative Analysis of Conformal Operating Points

Integrating conformal prediction into our performance visualization allows us to see exactly where the "guaranteed" risk operating points lie relative to the model's theoretical optimum (Best F1 Score).

An interesting observation concerns the relationship between the acceptable tolerance level α and the threshold that maximizes the F1 score in the PR-AUC space.

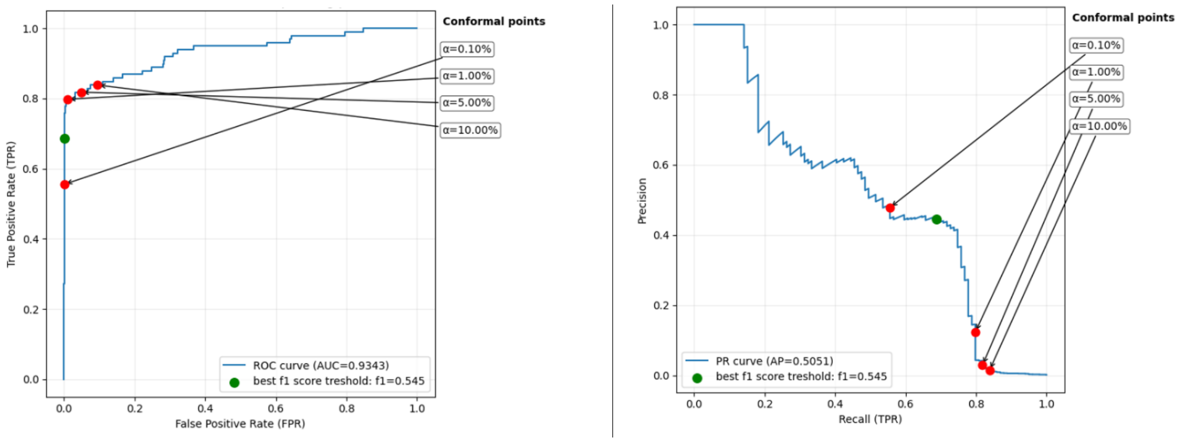


(a) Supervised Baseline ROC with Conformal Points

(b) Supervised Baseline PR with Conformal Points

Figure 4: Conformal operating points for the supervised baseline. The point $\alpha = 0.1\%$ (red) offers operationally safe boundaries, while the Best F1 threshold (green) prioritizes statistical balance.

For the **supervised baseline**, small values of α (e.g., 0.1% and 1%) lead to operating points that approach closer and closer to the F1-optimal threshold, while preserving a high recall. This suggests that the F1-optimal decision rule is already aligned with a relatively conservative flagging policy. In other words, strong performance and conformal guarantees are well aligned: the classifier performs well and does not require a substantial sacrifice in recall to satisfy a stricter tolerance level.



(a) VAE ROC with Conformal Points

(b) VAE PR with Conformal Points

Figure 5: Conformal operating points for the VAE model. The $\alpha = 0.1\%$ point lies near the knee of the ROC curve, providing a stable alternative to the Best F1 threshold.

In contrast, for the **VAE**, tightening the tolerance constraint induces a notable loss in recall. For example, moving from $\alpha = 1\%$ to $\alpha = 0.1\%$ results in a substantially more aggressive operating point with much lower recall relative to the F1-optimal point. This illustrates a more ambiguous decision space in the unsupervised setting: there exists a nontrivial trade-off between operational performance (precision–recall behavior) and statistical guarantees (false positive control). While the VAE offers the appealing property of detecting unseen fraud modes, conformal calibration exposes a steeper cost in recall when enforcing very small α values.

Operational Choice. Comparing the two models, the $\alpha = 0.1\%$ operating point emerges as a strong candidate for deployment. It resides close to the optimal F1 region for the Supervised model and captures the efficient frontier of the VAE. This threshold offers a "good enough" balance—it is sufficiently close to the best statistical performance (F1) while providing the business with a hard guarantee on the maximum number of daily false alarms (operational cost).

6 Discussion and Future Work

6.1 Supervised vs. Unsupervised Roles

The supervised baseline (ROC-AUC 0.98) clearly outperforms the VAE (ROC-AUC 0.93) in pure predictive capability. This is expected, as the supervised model trains directly on the labels we wish to predict. However, the VAE's performance is competitive and, crucially, independent of historical labels.

In a production environment, these models should not be viewed as competitors but as complementary layers. The supervised model acts as a precision tool for known fraud patterns, while the VAE functions as a safety net for "zero-day" or emerging fraud attacks that deviate from normal behavior but do not match known fraud signatures. Conformal prediction unifies these two approaches by converting their disparate raw scores (probability vs. reconstruction error) into a common currency: the False Positive Rate.

6.2 Architectural Improvements

Based on our analysis of the reconstruction errors and latent space, we propose the following advancements for the unsupervised component:

1. **Deeper Architectures with Skip Connections:** The current encoder may miss fine-grained feature interactions. Implementing ResNet-style blocks would allow for deeper networks without gradient degradation.

2. **Advanced Activation Functions:** Replacing ReLU with Mish or LeakyReLU could improve the reconstruction of numerical magnitudes (like transaction amounts) by preserving negative gradients, which is often lost in standard ReLU networks.

7 Conclusion

Overview. Credit card fraud detection is characterized by extreme class imbalance and the constant evolution of fraudulent tactics. Traditional rule-based systems often fail to adapt, while standard supervised learning relies heavily on labeled historical data that may become obsolete. This project addressed these challenges by developing a hybrid detection framework that combines the precision of supervised learning with the adaptability of unsupervised anomaly detection.

Methodology. Our approach utilized a two-stage modeling strategy. First, we established a high-performance baseline using a weighted Deep Fully Connected Network to capture known fraud patterns. Second, we implemented a Variational Autoencoder (VAE) trained exclusively on legitimate transactions to model the manifold of "normal" behavior. To bridge the gap between raw model outputs and operational decision-making, we applied Inductive Conformal Prediction (ICP). This technique transformed arbitrary anomaly scores into interpretable thresholds with rigorous statistical guarantees on the False Positive Rate.

Key Findings. The results demonstrate that unsupervised methods can achieve robust performance without labeled data, with the VAE achieving an AUC of 0.93 compared to the supervised baseline's 0.98. More importantly, the application of conformal prediction proved successful in controlling operational risk. We found that a strict tolerance level ($\alpha = 0.1\%$) provided an optimal balance, aligning closely with the best theoretical F1 scores while ensuring that the rate of false alarms remained below 0.1% for legitimate users.

Future Directions. Looking ahead, the integration of these models into a unified ensemble represents the next logical step. By running the supervised and unsupervised models in parallel and calibrating their outputs via conformal prediction, financial institutions can achieve a detection system that is both accurate against known threats and resilient against novel attacks. Future work will focus on architectural refinements, such as the inclusion of residual connections, to further compress the reconstruction error of complex legitimate transactions.

References

[1] R. Shwartz-Ziv and A. Armon, *Tabular data: Deep learning is not all you need*, Information Fusion, vol. 81, pp. 84–90, 2022.

[2] X. Glorot, A. Bordes, and Y. Bengio, *Deep sparse rectifier neural networks*, Proc. 14th International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 315–323, 2011.

[3] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, *Dropout: A simple way to prevent neural networks from overfitting*, Journal of Machine Learning Research (JMLR), vol. 15, no. 1, pp. 1929–1958, 2014.

[4] S. Ioffe and C. Szegedy, *Batch Normalization: Accelerating deep network training by reducing internal covariate shift*, Proc. International Conference on Machine Learning (ICML), pp. 448–456, 2015.

[5] S. Obushnyi, D. Virovets, A. Ramskyi, M. Zhytar, and P. Skladannyi, *Variational Autoencoders for Detecting Anomalous and Fraudulent Transactions in Financial Systems*, Workshop on Digital Economy Concepts and Technologies, vol. 4029, pp. 110–118, Germany, Sept. 2025.

[6] *Hybrid Variational Autoencoder-based Models for Fraud Detection*, Medium (Analytics Vidhya),
https://medium.com/analytics-vidhya/hybrid-variational-autoencoder-based-models-fo

[7] D. J. Rezende, S. Mohamed, and D. Wierstra, *Stochastic backpropagation and approximate inference in deep generative models*, Proc. International Conference on Machine Learning (ICML), pp. 1278–1286, PMLR, 2014.

[8] S. R. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio, *Generating sentences from a continuous space*, Proc. 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL), pp. 10–21, 2016.

[9] A. N. Angelopoulos and S. Bates, *A gentle introduction to conformal prediction and distribution-free uncertainty quantification*, arXiv preprint arXiv:2107.07511, 2021.