FraudGuard: Unsupervised Anomaly Detection in Financial Transactions using VAEs and Conformal Prediction

Problem Statement: Credit card fraud represents a massive financial liability, yet detecting it is complex due to extreme class imbalance (<0.2% fraud). Standard supervised models often fail to generalize or bias toward the majority class. Furthermore, "black box" anomaly detectors lack statistical guarantees, making it hard for banks to control how many legitimate customers get blocked. This project builds an unsupervised system that not only detects outliers but uses statistical calibration to provide rigorous guarantees on the False Positive Rate, balancing risk control with user experience.

Dataset: We will utilize the Credit Card Fraud Detection dataset provided by the Machine Learning Group (ULB) on Kaggle.

- **Link:** https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
- **Details:** The dataset contains 284,807 transactions with 492 frauds. It features PCA-transformed input variables (V1-V28), allowing us to focus on architecture and calibration rather than feature engineering.

Methodology: We treat fraud detection as a reconstruction problem with a statistically calibrated threshold. We will use PyTorch to implement two distinct deep learning architectures :

1. **Supervised Baseline (Feedforward Network):** We will train a deep Fully Connected Network (FCN) with weighted loss functions (to handle imbalance) as a performance baseline (Referencing *Lecture 02: Backpropagation*)
2. **Unsupervised Anomaly Detector (VAE):** We will train a Variational Autoencoder exclusively on legitimate transactions. The model learns to compress and reconstruct valid data. Fraudulent transactions, being "out of distribution," will yield a high Reconstruction Error.
3. **Conformal Prediction for Thresholding:** Instead of setting an arbitrary error threshold, we will apply **Conformal Prediction** on a hold-out calibration set. We will compute non-conformity scores (based on reconstruction error) to mathematically determine a dynamic threshold. This allows us to strictly control the rate of false alarms.

Expected Outcome: We aim to move beyond simple accuracy and provide a reliable, "business-ready" detector.

- **Statistical Guarantee:** By leveraging Conformal Prediction, we expect to calibrate the model to **guarantee blocking at most Alpha in {1%, 2%, 5%} of legitimate transactions** (controlled False Positive Rate), while maximizing the percentage of fraud detected at those levels.
- **Metrics:** We target an **ROC-AUC > 0.90** and will present a "Calibration Curve" showing the trade-off between the guaranteed Alpha (blocked legitimate users) and the empirical fraud detection rate.
- **Visualization:** Visualizing the separation of classes in the VAE latent space and the distribution of non-conformity scores.