FraudFusion: Enhanced Diffusion Models for Synthetic Fraud Data Generation

Anish Rao, Raghu Ram Sattanapalle

April 17, 2025

Abstract

Building on the FraudDiffuse model, this work introduces FraudFusion, a diffusion-based generative model designed to synthesize realistic fraudulent transaction data using the Sparkov dataset, which exhibits different fraud patterns than the IEEE-CIS dataset used in the original FraudDiffuse paper. In addition to the enhancements proposed in the original paper, we developed several novel enhancements: an amount distribution loss to better capture bimodal transaction patterns, engineered feature range loss, cyclic encoding for temporal features, and post-processing distribution matching techniques. Our latest model demonstrates superior synthetic data quality through improved statistical distribution matching, particularly for transaction amounts. When used to augment training data for XGBoost classifiers, our synthetic samples improved fraud detection recall by 5.75 percentage points (from 82.75% to 88.50%), representing a significant operational advantage for financial institutions where the cost of missing fraudulent transactions far exceeds that of investigating false positives. We also introduce a controlled validation methodology that enables more reliable model selection when working with synthetic data, contributing a valuable framework for future research in this domain.

1 Introduction

Credit card fraud detection represents a critical challenge for financial institutions, where the ability to accurately identify fraudulent transactions can save billions of dollars annually while protecting consumers. This report presents FraudFusion, an enhanced diffusion-based approach for generating high-quality synthetic fraud samples to address the extreme class imbalance inherent in fraud detection.

Our experimentation phase builds upon the foundation of diffusion models for tabular data, with specific enhancements designed to capture the unique statistical properties of fraudulent transactions. This phase is critical to the overall project as it addresses two fundamental challenges that limit the effectiveness of current fraud detection systems: Extreme Class Imbalance and Complex Statistical Relationships. Fraudulent transactions typically represent less than 0.1% of all credit card activity, causing standard classification models to be biased towards the majority class and leading to unacceptably high false negative rates. Additionally, fraudulent transactions exhibit distinctive patterns across multiple dimensions (amount, timing, location) that cannot be adequately captured by simple resampling techniques or generic synthetic data generation approaches.

Our methodology employs a specialized diffusion model architecture with novel loss functions specifically designed to address these challenges. By generating synthetic fraud samples that preserve the statistical properties of real fraud while introducing meaningful variations, we aim to improve the downstream classifier's ability to detect fraud without overfitting to known patterns.

1.1 Problem Definition

Financial fraud detection represents a challenging **binary classification** task characterized by extreme class imbalance and domain-specific complexities: With fraud cases representing approximately 1 in 1,000 transactions, standard classification approaches tend to bias toward the majority class, resulting in poor fraud detection performance. Moreover, transaction data contains a mixture of continuous variables (e.g., transaction amount, location coordinates) and categorical variables (e.g., merchant category, transaction type) that exhibit complex interdependencies. Fraudulent activity often shows distinctive patterns in transaction timing, location, and amount that must be properly captured for effective detection. Additionally, The cost of missing a fraudulent transaction (false negative) typically far exceeds the cost of falsely flagging a legitimate transaction (false positive).

This experimentation phase explores how enhanced diffusion models can address these challenges by generating synthetic fraud samples that maintain the complex statistical relationships present in real fraud data while providing sufficient variation to improve model generalization.

1.2 Practical Significance

Effective fraud detection systems have substantial real-world impacts across multiple dimensions. Financially, the industry loses billions of dollars annually to fraud. Even a modest improvement in detection rates can translate to significant cost savings. For instance, a 6% increase in fraud detection sensitivity (as achieved in our approach) could potentially save hundreds of millions of dollars at the scale of major financial institutions. From the consumer perspective, undetected fraud not only affects institutions but also creates substantial distress for people whose accounts are compromised. Improved fraud detection directly benefits consumer financial security. Operationally, reducing false positives while maintaining high sensitivity decreases the burden on fraud investigation teams, allowing more efficient allocation of human resources.

Beyond these practical benefits, our approach addresses a persistent challenge in machine learning—how to effectively generate synthetic data that preserves the complex statistical relationships of the original data while introducing useful variations to improve model generalization.

2 Related Work

Our research builds upon recent advances in diffusion models for synthetic data generation, with a specific focus on applications in financial fraud detection. The field has seen significant innovation in recent years, with several approaches that inform our methodology:

FraudDiffuse: Roy et al. [5] introduced "FraudDiffuse," a diffusion-aided approach for synthetic fraud augmentation using the IEEE-CIS fraud dataset. Their work demonstrated the effectiveness of diffusion models in generating high-quality synthetic fraud samples that improve detection performance. Our experimentation expands upon FraudDiffuse by applying similar techniques to the Sparkov dataset, enabling us to test the generalizability of this approach to different fraud patterns while introducing novel enhancements for distribution matching.
TabDDPM: Kotelnikov et al. [1] presented "TabDDPM," which applies diffusion models to generic tabular datasets with mixed numerical and categorical data. Our experimentation differentiates itself by specifically targeting financial fraud detection, focusing on identifying complex transactional behaviors unique to credit card fraud. This requires tailored preprocessing and feature engineering techniques different from those used in generic tabular datasets.

FinDiff: "FinDiff: Diffusion Models for Financial Tabular Data Generation" [6] established a diffusion-based generative approach designed broadly for financial tabular datasets. While FinDiff is a general-purpose model applicable to tasks such as economic scenario modeling and stress testing, our experimentation uses a diffusion model specifically tailored for fraud detection, with optimizations that address the unique challenges of extreme class imbalance in fraud datasets.

Imb-FinDiff: "Imb-FinDiff: Conditional Diffusion Models for Class Imbalance Synthesis of Financial Tabular Data" [7] introduced a denoising diffusion framework specifically designed to address class imbalance in financial tabular datasets. While similar in objective to our work, Imb-FinDiff focuses on general financial data, whereas our experimentation incorporates fraud-specific optimizations, such as contrastive learning loss functions, to better capture patterns near the decision boundary.

Dual-Track Diffusion Approach: Pushkarenko and Zaslavskyi [3] explored a dual-track approach using two separate diffusion models on benchmark financial data and the IEEE-CIS dataset. While their methodology demonstrated the effectiveness of diffusion models for synthetic data generation, our experimentation focuses specifically on a single, optimized diffusion model with architectural modifications designed for the Sparkov dataset.

2.1 Innovations in Our Approach

Our experimentation builds upon these foundations while making several key contributions:

- We develop a specialized diffusion model architecture with enhanced handling of mixed data types (continuous and categorical features) specific to fraud transaction data.
- We incorporate novel loss functions targeting specific fraud-related features, particularly transaction amount distributions, which exhibit distinctive bimodal patterns in fraud cases.
- We implement a dual validation strategy that enables more reliable model selection when working with synthetic data.
- We provide comprehensive performance metrics beyond standard benchmarks, specifically focusing on the sensitivity-precision trade-off that is critical in fraud detection applications.

The broader implications of our methodology extend beyond fraud detection to other domains characterized by extreme class imbalance, such as disease diagnosis, network intrusion detection, and rare event prediction, where traditional resampling techniques prove inadequate.

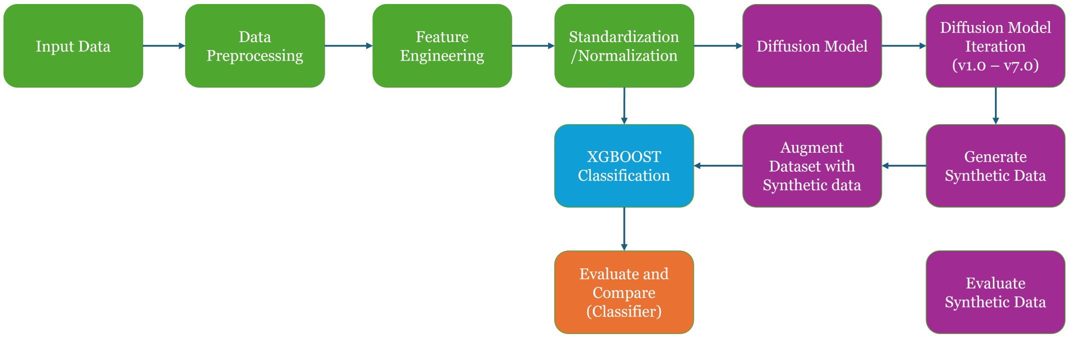
Input Data
Data Preprocessing
Feature Engineering
Standardization /Normalization
Diffusion Model
Diffusion Model Iteration (v1.0 – v7.0)
XGBOOST Classification
Augment Dataset with Synthetic data
Generate Synthetic Data
Evaluate and Compare (Classifier)
Evaluate Synthetic Data

Figure 1: Experimental Workflow: From Data Preprocessing to Model Evaluation

Figure 1 outlines our complete experimental workflow, from data preprocessing and feature engineering to synthetic data generation and model evaluation. This framework allowed us to systematically assess the impact of different model configurations, loss functions, and architectural decisions on the quality of synthetic data and downstream classifier performance.

3 Research Objectives and Hypotheses

This experimentation phase aims to address fundamental challenges in credit card fraud detection through enhanced synthetic data generation techniques. Our research is guided by specific objectives and testable hypotheses that directly align with the problem definition established in the previous sections.

3.1 Primary Research Objectives

Our experimentation is designed to achieve the following key objectives: develop an enhanced diffusion model architecture that generates high-quality synthetic fraud samples while preserving the complex statistical properties of real fraud transactions, particularly the bimodal patterns observed in transaction amounts; improve fraud detection performance by using synthetic data augmentation to address the extreme class imbalance problem, with a specific focus on increasing recall without substantially compromising precision; design and validate novel loss functions that specifically target the unique characteristics of fraud data, including temporal patterns and bimodal transaction amount distributions; establish a robust validation methodology for synthetic data that reliably assesses both distributional accuracy and downstream classification performance; and quantify the relationship between synthetic data quantity and classifier performance to determine optimal augmentation strategies.

3.2 Research Hypotheses

To systematically evaluate our approach, we formulated the following testable hypotheses:

1. **Distribution Matching Hypothesis**
   Synthetic fraud samples generated by our enhanced diffusion model will demonstrate statistically equivalent distributions to real fraud transactions across key features, as measured by Wasserstein distance, energy distance, and statistical tests (KS test, Anderson-Darling test).

2. **Bimodal Amount Distribution Hypothesis**
   Our specialized amount distribution loss will produce synthetic samples that more accurately capture the bimodal distribution of transaction amounts in real fraud data compared to the standard diffusion approach, particularly in preserving both low-value and high-value fraud patterns.

3. **Classification Performance Hypothesis**
   XGBoost models trained with synthetic fraud augmentation will achieve significantly higher recall than models trained only on imbalanced real data, with minimal degradation in overall discrimination capacity (ROC-AUC).

These hypotheses provide a clear framework for evaluating both the quality of our synthetic data generation technique and its practical utility in improving fraud detection performance.

4 Experimental Setup

This section details the hardware and software infrastructure used in our experimentation phase, along with the rationale for our model selection and evolution process based on previous iterations.

4.1 Hardware and Software Specifications

All experiments were conducted using the following computational resources:

Hardware:

- **GPU:** NVIDIA GeForce RTX 4060 with 8GB VRAM for diffusion model training
- **CPU:** Intel Core i7 14700F (20 cores: 8P + 12E) for data preprocessing and classifier training
- **RAM:** 64GB DDR5 for handling large dataset operations and efficient parallel processing
- **Storage:** 1 TB NVMe SSD for high-speed dataset access and model checkpoints

Software Environment:

- **Operating System:** Windows 11 with WSL2 for Linux compatibility
- **Programming Language:** Python 3.10
- **Deep Learning Framework:** PyTorch 2.6.0+cu118 with CUDA support
- **Machine Learning Libraries:**
  - XGBoost 2.1.3 for classification models
  - Scikit-learn 1.0.2 for preprocessing and evaluation
  - Pandas 2.2.3 and NumPy 1.26.4 for data manipulation
  - Joblib 1.4.2 for parallel processing
- **Visualization and Statistical Testing:**
  - Matplotlib 3.10.0 and Seaborn 0.13.2 for visualization
  - SciPy 1.13.1 for statistical tests (Kolmogorov-Smirnov, Anderson-Darling)
  - TQDM 4.67.1 for progress tracking during lengthy model training and generation

4.2 Model Selection and Refinement

This experimentation phase builds upon our previous iteration's model evaluation while implementing targeted enhancements based on comprehensive performance analysis. We maintained the core models from our previous work while systematically addressing identified limitations.

4.2.1 Models Evaluated in Previous Iterations

Our previous iteration involved extensive evaluation of multiple modeling approaches. Traditional resampling techniques such as SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling) were explored. SMOTE failed to capture the complex feature distributions and relationships common in fraud data since it is limited by linear interpolation between neighboring samples. ADASYN, while focusing on difficult examples, similarly proved inadequate for capturing complex dependencies between fraud features.

We also evaluated several deep generative models. Among these, GANs (Generative Adversarial Networks)—including WGAN, WCGAN, and WCGAN-GP variants—showed potential but suffered from training instability and mode collapse, issues particularly problematic when dealing with the mixed numerical and categorical features characteristic of fraud datasets. VAEs (Variational Autoencoders) offered more stable training but often produced samples with blurred feature distributions, potentially losing critical fraud patterns. Diffusion models, by contrast, demonstrated superior performance in generating high-quality tabular data while preserving important statistical properties.

For the final classification tasks, we experimented with tree-based models such as Random Forest and gradient boosting frameworks like XGBoost and LightGBM, which typically perform well on tabular data and are capable of handling class imbalance effectively. Additionally, we employed neural networks, including MLPs and more specialized architectures tailored for tabular data.

4.2.2 Finalized Classifier Model: XGBoost

Based on our previous evaluations, we retained XGBoost as our classification model due to its demonstrated advantages for fraud detection. XGBoost implements a regularized form of gradient boosting that minimizes the following objective:

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}i) + \sum k = 1^K \Omega(f_k) \tag{1}$$

where l is a differentiable convex loss function (typically logistic loss for binary classification), $\hat{y}_i$ is the prediction for the i-th instance, and $\Omega(f) = \gamma T + \frac{1}{2}\lambda|w|^2$ is a regularization term that penalizes model complexity. This regularization helps prevent overfitting, which is particularly important when modeling the complex patterns observed in fraud data.

XGBoost also offers native support for imbalanced datasets through the `scale_pos_weight` parameter, which scales the gradient for the minority class as,

$$\text{scale\_pos\_weight} = \frac{n_{\text{negative}}}{n_{\text{positive}}} \tag{2}$$

This adjustment helps address the severe class imbalance inherent in fraud detection tasks, even before applying our synthetic data augmentation strategy.

Another advantage of XGBoost is its ability to calculate feature importance scores based on their contribution to performance improvement. This provides valuable insights into which transaction attributes most strongly indicate fraudulent behavior:

$$\text{Importance}(X_j) = \sum_{k=1}^{K} \sum_{i=1}^{n} 1(x_{ij} \text{ is split on}) \times \text{Gain}_i \tag{3}$$

where Gain_i represents the improvement in accuracy brought by a split on feature X_j.

The model's efficient training process includes optimizations such as a sparsity-aware split finding algorithm and a distributed weighted quantile sketch for handling sparse data. These features allow effective utilization of computational resources, especially when working with large-scale financial datasets.

XGBoost has also been established as a benchmark model in the literature, having been used in the original FraudDiffuse paper (Roy et al., 2023 [5]) and other recent fraud detection studies. This consistency supports its continued use in evaluating the quality of synthetic data generation techniques.

Consistent with our previous experimental design, we maintained three distinct XGBoost configurations in this iteration. The Baseline XGBoost model was trained only on the original imbalanced data to establish performance benchmarks. The Augmented XGBoost model was trained on a combination of original data and synthetic fraud samples generated by our enhanced FraudDiffuse model. Finally, the Controlled XGBoost model was trained on a balanced dataset containing original non-fraud samples and a mix of original and synthetic fraud samples with controlled proportions.

This comparative approach allows us to systematically evaluate how the quality and quantity of synthetic data affect classifier performance, providing empirical validation for our enhanced generative approach.

5 Dataset and Preprocessing

5.1 Data Sources

Our project utilizes the Sparkov Credit Card Fraud Detection Dataset (also known as the "Credit Card Transactions Fraud Detection Dataset"), obtained from Kaggle [2], with no access restrictions.

Unlike the IEEE-CIS dataset used in the original FraudDiffuse paper, the Sparkov dataset simulates realistic credit card transaction patterns with different fraud distributions, allowing us to test the generalizability of diffusion-based synthetic data generation approaches.

5.2 Data Description

The dataset comprises 1,296,675 transactions after initial processing. It includes 23 original features, categorized into transaction details (amount, date/time, merchant information), cardholder demographics (age, gender, job with 494 unique categories), geospatial information (customer and merchant latitude/longitude, city population), and categorical features (merchant category with 14 categories, gender with 2 categories, and state with 51 categories).

As shown in Figure 2, only 0.52% of transactions are fraudulent, resulting in an extreme imbalance ratio of approximately 1:192.

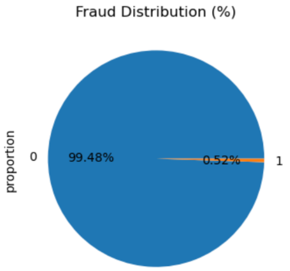

Figure 2: Class Imbalance

Our exploratory data analysis revealed several important patterns: fraud transactions display distinctive patterns in transaction amounts; fraud occurs more frequently during certain hours (especially early morning) and during the first six months of the year; and, as shown in Figure 3, fraud cases tend to cluster in specific geographic locations.

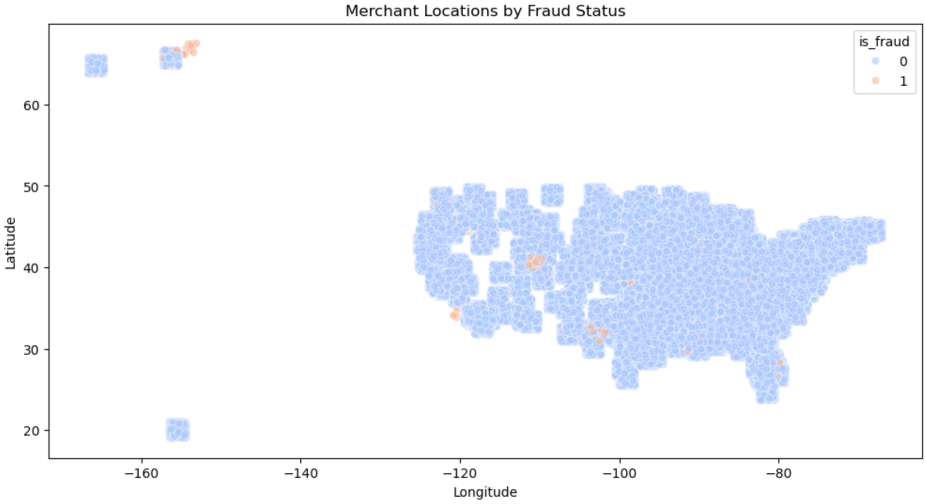

Figure 3: Fraud Locations

We also observed higher fraud rates in the 76–85 (0.92%) and 86–95 (0.87%) age groups.

5.3 Preprocessing Steps

During data cleaning, we found no missing values. We removed identifier columns (`cc_num`, `first`, `last`, `trans_num`) and redundant timestamps (`unix_time`) to prevent data leakage.

For feature transformation, we applied a log transformation (using `np.log1p`) to skewed numerical features, particularly transaction amount (`amt`) and city population (`city_pop`), and then standardized all numerical features using scikit-learn's `StandardScaler` to achieve zero mean and unit variance.

In temporal feature engineering, we extracted the hour of day, day of the week, and month from timestamps. Sine-cosine transformations were applied to these features to preserve their cyclical nature:

$$t_{sin} = \sin\left(\frac{2\pi \cdot t}{P}\right), \quad t_{cos} = \cos\left(\frac{2\pi \cdot t}{P}\right) \tag{4}$$

where t is the temporal feature (e.g., hour, day) and P is the period (e.g., 24 for hours, 7 for days of week)

We evaluated the predictive power of features using Mutual Information. Transaction amount (`amt`) showed the highest predictive power (MI: 0.0158), followed by geographic coordinates (`lat`, `long`), and then city population (`city_pop`) with lower signal (MI: 0.00302).

For categorical encoding, we applied a tiered strategy: one-hot encoding for low-cardinality features (e.g., `category`, `gender`), target encoding for medium-cardinality features (`state`), and frequency encoding for high-cardinality features (`job`, `merchant`). We also calculated the geographical distance between customer and merchant locations and included it as an additional feature for fraud detection.

Specialized handling was also used to capture fraud patterns: transaction amounts were modeled to reflect their distinctive bimodal distribution, with peaks in both low and high-value ranges. For temporal features, we computed observed min/max values in the standardized space and imposed constraints to ensure generated values remained within realistic bounds. We also developed augmentation techniques specifically tailored to preserve the bimodal transaction amount distribution and temporal fraud patterns.

5.4 Synthetic Data Strategy

Our preprocessing pipeline was designed to support synthetic data generation, with a particular focus on the distributional characteristics of fraud.

A detailed analysis of distributions was conducted for transaction amounts, time-of-day patterns, and geographic clustering of fraud. To prevent data leakage, we identified and excluded sensitive fields such as transaction identifiers (`trans_num`), credit card numbers (`cc_num`), and exact timestamps (`trans_date_trans_time`). Additionally, a dual validation strategy was implemented using a "pure" validation set containing only real data, and a "synthetic" validation set incorporating generated samples.

To ensure the quality of synthetic data, we developed a robust evaluation framework that includes statistical distribution matching tests, feature correlation preservation metrics, and validation of temporal and spatial pattern fidelity. Finally, our approach to feature engineering was guided by the specific requirements of synthetic data generation. We implemented specialized handling for the bimodal distribution of transaction amounts observed in fraudulent cases. This involved using distribution-aware initialization during the generation process, followed by quantile-based distribution matching in post-processing to better preserve this pattern.

For temporal features, we enforced range constraints based on patterns observed in real-world fraud data, ensuring that generated samples reflected realistic time distributions. Additionally, we developed feature-specific transformation techniques aimed at preserving the intricate relationships between variables that distinguish fraudulent transactions from legitimate ones.

Our preprocessing approach enables the diffusion model to learn the complex statistical relationships present in fraud transactions while providing a robust framework for synthetic sample evaluation and integration.

6 Methodology

6.1 Enhanced Diffusion Model Architecture

Our enhanced FraudDiffuse model represents a significant evolution from the baseline architecture described by Roy et al. [5]. The core of our implementation is the `CombinedNoisePredictor` neural network, which processes multi-modal input data. This architecture integrates three distinct data types: normalized numerical features (11 dimensions), embedded categorical features (8 categories with varying cardinality), and specialized cyclic encodings for temporal features (8 dimensions).

Each categorical feature is processed through dedicated embedding layers according to the equation $e_{i,j} = \text{Embedding}_j(x_{cat,i,j})$, where $x_{cat,i,j}$ represents the j-th categorical feature for the i-th sample. These embeddings are learned during training with dimension 4 per feature. The model employs a four-layer feed-forward structure with a combined input dimension of 52 (including timestep), hidden layers with dimension 256 (reduced from 320 in earlier versions for stability), gentle residual connections with scaling factor 0.1 between hidden layers, layer normalization after each hidden layer, ReLU activation functions, and dropout with rate 0.1 for regularization. All layers utilize Xavier uniform initialization to ensure stable gradient flow during training.

The forward pass of our model can be expressed as:

$$\begin{aligned} x_{input} &= \text{Concat}[x_{num}, x_{cat\_embedded}, x_{cyclic}, t_{norm}] \\ h_1 &= \text{Dropout}(\text{LayerNorm}(\text{ReLU}(W_1 x_{input} + b_1))) \\ h_2 &= \text{Dropout}(\text{LayerNorm}(\text{ReLU}(W_2 h_1 + b_2))) + 0.1 \times h_1 \\ h_3 &= \text{Dropout}(\text{LayerNorm}(\text{ReLU}(W_3 h_2 + b_3))) + 0.1 \times h_2 \\ \hat{\epsilon} &= W_4 h_3 + b_4 \end{aligned} \tag{5}$$

Where $\hat{\epsilon}$ represents the predicted noise at timestep t.

Our model incorporates domain-specific components to address the unique characteristics of fraud data. Transaction amount, a critical fraud indicator, receives specialized treatment through bimodal initialization during generation, KDE-based peak detection for distribution modeling, and quantile-based distribution matching in post-processing. We implemented feature-specific weighting in the loss function, with transaction amount weighted at 1.8, transaction hour at 1.3, transaction month and day of week at 1.1 each, and other features at 1.0.

6.2 Diffusion Process

The diffusion process gradually adds noise to data points according to:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1 - \alpha_t} \epsilon_t \tag{6}$$

where x_t is the noised data at timestep t, $\alpha_t = 1 - \beta_t$ with β_t increasing linearly from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$, and $\epsilon_t \sim \mathcal{N}(0, I)$ is standard Gaussian noise.

For efficiency, we use the compounded form:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon \tag{7}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and $\epsilon \sim \mathcal{N}(0, I)$.

The reverse process for generation follows:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t z \tag{8}$$

where $z \sim \mathcal{N}(0, I)$ and σ_t decreases as t approaches 0.

6.3 Loss Function Components

Our composite loss function represents a significant advancement over the original FraudDiffuse formulation:

$$\mathcal{L}_{total} = \mathcal{L}_{norm} + w_1 \times \mathcal{L}_{prior} + w_2 \times \mathcal{L}_{triplet} + \lambda_{eng} \times \mathcal{L}_{eng} + \lambda_{amt} \times \mathcal{L}_{amt} \tag{9}$$

Each component addresses a specific aspect of synthetic data quality. The feature-weighted $\mathcal{L}_{norm}$ implements an enhanced mean squared error between true and predicted noise with feature-specific weights:

$$\mathcal{L}_{norm} = \mathbb{E}_{x_0,\epsilon,t}\left[\frac{1}{n}\sum_{j=1}^{n} w_j(\epsilon_j - \epsilon_{\theta j})^2\right] \tag{10}$$

where w_j represents the importance weight for feature j.

The non-fraud prior loss ($\mathcal{L}_{prior}$) forces the model to learn subtle patterns distinguishing fraud from legitimate transactions:

$$\mathcal{L}_{prior} = 2 \times P(Z \leq |z\text{-}score|) = 1 - 2 \times P(Z \geq |z\text{-}score|) \tag{11}$$

where $z\text{-}score = \frac{\epsilon_{\theta j} - \mu_j}{\sigma_j}$ and $\epsilon_{\theta j}$ is the predicted error for the j-th feature.

Triplet loss ($\mathcal{L}_{triplet}$) serves as a contrastive component ensuring synthetic fraud samples remain close to real fraud while distant from non-fraud samples:

$$\mathcal{L}_{triplet} = \max(0, d(\hat{x}_f, x_f) - d(\hat{x}_f, x_{nf}) + \text{margin}) \tag{12}$$

where $\hat{x}_f$ represents generated fraud samples, x_f real fraud samples, and x_{nf} non-fraud samples.

The engineered range loss ($\mathcal{L}_{eng}$) constrains temporal features to realistic ranges based on observed fraud patterns:

$$\mathcal{L}_{eng} = \mathbb{E}\left[\sum_{j \in E} \max(0, t_{min_j} - \hat{t}_j) + \max(0, \hat{t}_j - t_{max_j})\right] \tag{13}$$

where E is the set of engineered temporal features, and t_{min_j} and t_{max_j} are the observed minimum and maximum values.

The amount distribution loss ($\mathcal{L}_{amt}$) is a specialized component that enforces a realistic bimodal distribution for transaction amounts with heightened emphasis on higher-value fraud:

$$\begin{aligned} \mathcal{L}_{amt} = &|\mu_{gen} - \mu_{real}| + \\ &1.0 \times |q_{50,gen} - q_{50,real}| + \\ &3.0 \times |q_{75,gen} - q_{75,real}| + \\ &5.0 \times |q_{90,gen} - q_{90,real}| + \\ &8.0 \times |q_{95,gen} - q_{95,real}| + \\ &4.0 \times |skew_{gen} - skew_{real}| \end{aligned} \tag{14}$$

The relative importance of these components was controlled through carefully tuned weights: $w_1 = 0.10$, $w_2 = 0.40$, $\lambda_{eng} = 0.05$, and $\lambda_{amt} = 0.20$.

6.4 Training Procedure

Our dataset was split using stratified sampling to maintain the fraud-to-legitimate ratio. The training set (65%) was used to train both models - only the fraud samples were used to train the diffusion model, while the complete training set was used for the XGBoost classifier. The validation set (15%) was used only for model evaluation during development, while the test set (20%) was reserved exclusively for final performance evaluation. This strict separation ensured no data leakage between the diffusion model training and downstream classification evaluation.

Training diffusion models on complex, mixed-type financial data presented significant stability challenges. We implemented several techniques to address these issues. Aggressive gradient clipping with max_norm=0.5 prevented exploding gradients. An adaptive learning rate scheduler (ReduceLROnPlateau with factor=0.7 and patience=15) optimized convergence. Comprehensive exception handling throughout the training loop with fallback loss calculations addressed NaN detection and recovery. Strategic clamping of intermediate values prevented numerical instabilities. We used a reduced batch size (32) for better stability with small fraud datasets, along with L2 regularization (1e-5) to prevent overfitting. The model was trained for 550 epochs with early stopping based on validation performance, reaching convergence after approximately 500 epochs on modern GPU hardware.

The diffusion process itself required careful tuning to ensure high-quality synthetic samples. We employed a linear beta schedule from $\beta_{start} = 10^{-4}$ to $\beta_{end} = 0.02$ over $T_{train} = 800$ steps. For inference efficiency without quality degradation, generation steps were reduced to $T_{gen} = 600$. Adaptive noise reduction was implemented in later generation steps, scaled by $t_{step}/200$ for $t < 200$.

6.5 Synthetic Data Generation Approach

Our final model evolved through a series of incremental improvements, each addressing specific performance limitations. Version 2 introduced range constraints for engineered features by computing observed min/max values in standardized space and adding penalty loss for values outside this range. Version 3 added feature-specific initialization for amount, implemented cyclical encoding for time features (hour, day, month, day of week), applied targeted loss weighting, and increased model capacity. Version 4 focused on stability with controlled distribution matching for amount feature, stability-preserving architecture changes, balanced loss weighting, and NaN prevention mechanisms. Version 5 enhanced distribution modeling to better capture the bimodal nature of the amount feature, improved age distribution modeling, and added feature-specific adjustments to the generation process. Version 7 (Final) implemented post-processing steps to enforce amount distribution matching, enhanced initialization specifically for the amount feature, applied more aggressive weighting for higher fraud amounts, and added distribution transformation matching.

A key innovation in our final model is the distribution-aware initialization and post-processing pipeline. During generation, we use KDE-based peak detection to identify the modes of the bimodal amount distribution, then initialize samples around these modes with controlled noise. After generation, we apply quantile-based distribution matching:

$$x_{matched} = F_{target}^{-1}(F_{source}(x_{generated})) \tag{15}$$

where F represents the empirical CDF function.

For fraud amounts, we implement a bimodal initialization strategy:

$$x_{amt} = \begin{cases} peak_{low} + \mathcal{N}(0, 0.1 \cdot d) & \text{with prob. } 0.1 \\ peak_{high} + \mathcal{N}(0, 0.08 \cdot d) & \text{with prob. } 0.9 \end{cases} \tag{16}$$

where $d = peak_{high} - peak_{low}$ is the distance between detected distribution peaks.

Temporal features are clipped to observed ranges to ensure realistic time patterns. This comprehensive approach ensures that our synthetic fraud samples closely match the statistical properties of real fraud, particularly for the critical transaction amount feature, while maintaining the complex relationships between features that distinguish fraud from legitimate transactions.

6.6 Evaluation Framework

6.6.1 Synthetic Data Quality Assessment

To comprehensively evaluate the quality of our synthetic fraud samples, we employed multiple complementary metrics that assess different aspects of distribution similarity. The Kolmogorov-Smirnov (KS) test was selected because it makes no assumptions about the underlying distribution, making it suitable for the complex, often multimodal distributions found in fraud data. The KS statistic measures the maximum distance between the empirical cumulative distribution functions (ECDFs) of real and synthetic data, providing a sensitive measure of distribution differences.

The Anderson-Darling test was chosen because it gives more weight to the tails of distributions than the KS test, which is particularly important for fraud detection where extreme values often represent the most critical fraud cases. Wasserstein Distance, also known as the Earth Mover's Distance, measures the minimum "cost" of transforming one distribution into another. We selected this metric because it provides a more intuitive measure of distribution similarity that accounts for both the magnitude and probability of differences, making it particularly valuable for assessing complex financial data.

Energy Distance was included as a complementary distance metric because it can detect differences in shape, scale, and location simultaneously, providing a holistic assessment of distribution matching that the other metrics might miss. We specifically examine the 95th and 99th percentile ratios because high-value fraud transactions often reside in these upper tails, and ensuring accurate modeling of these regions is critical for effective fraud detection. Comparison of statistical moments (mean, variance, skewness, and kurtosis) provides insight into how well our synthetic data captures the central tendency, spread, asymmetry, and tail behavior of the real fraud distribution. This comprehensive approach ensures assessment of distribution similarity across multiple dimensions rather than relying on a single metric that might miss important discrepancies.

6.6.2 Classification Performance Assessment

For evaluating our fraud detection models, we selected metrics that address the specific challenges of imbalanced classification in fraud detection. Receiver Operating Characteristic Area Under Curve (ROC-AUC) assesses the model's ability to rank fraud cases higher than legitimate transactions across all possible threshold settings. We chose ROC-AUC as a primary metric because it provides a comprehensive assessment of discrimination performance independent of class imbalance.

Precision-Recall Area Under Curve (PR-AUC) is sensitive to class imbalance and focuses specifically on the minority class performance, making it particularly suitable for fraud detection where the focus is on the rare fraud cases rather than the abundant legitimate transactions. Sensitivity/Recall measures the proportion of actual fraud cases that are correctly identified, which directly aligns with the business objective of minimizing missed fraud. We highlight recall because each undetected fraud transaction represents a direct financial loss.

Precision measures the proportion of predicted fraud cases that are actually fraudulent and is critical because false fraud alerts generate operational costs through unnecessary investigations and potential customer friction. F1 Score, as the harmonic mean of precision and recall, provides a balanced assessment of model performance that is particularly useful when seeking an optimal trade-off between fraud capture and false alerts. Specificity assesses the model's ability to correctly identify legitimate transactions, which is important for minimizing customer friction in high-volume transaction processing systems. These metrics were selected to provide a multifaceted view of model performance rather than optimizing for a single dimension, allowing stakeholders to make informed decisions based on their specific cost-benefit considerations.

6.6.3 Dual Validation Strategy

A key innovation in our evaluation approach is the implementation of a dual validation strategy that provides a more robust assessment of models trained with synthetic data. This methodology systematically addresses the challenges of evaluating models trained with synthetic data through three distinct dataset configurations:

- **Combined Training Set:** The original training data augmented with a controlled proportion of synthetic fraud samples, using a 1:1 ratio of synthetic to real fraud samples.
- **Pure Validation Set:** The original, unmodified validation set containing only real data, providing an unbiased assessment of model generalization to unseen real-world data.
- **Synthetic Validation Set:** The original validation data augmented with synthetic samples following the same proportion as the training set.

To ensure proper allocation of synthetic samples, we implemented a proportional distribution approach. For our experiments with 5,000 synthetic samples (with 6,273 real fraud samples in training and 1,448 in validation), we adjusted the allocation to 4,062 synthetic samples for training and 938 for validation. For the 8,000 synthetic sample experiment, we achieved the full desired allocation (6,273 for training, 1,448 for validation).

We employed sample weighting during training, with synthetic samples given a lower weight (0.5) than real samples to prevent them from dominating the learning process. This weighting was applied to both the training and synthetic validation sets.

During model training, we simultaneously monitored performance metrics (log loss, ROC-AUC, and PR-AUC) on all three datasets, implementing early stopping based on validation performance. The optimal classification threshold was determined by maximizing the F1 score specifically on the pure validation set, ensuring that the operating point was tuned based on real-world data distribution rather than synthetic patterns.

This comprehensive validation methodology addresses a critical limitation in prior synthetic data research by ensuring that our reported performance metrics reflect genuine generalization capability rather than artifacts of the synthetic data generation process.

7 Results

We evaluated our enhanced FraudDiffuse model through a comprehensive assessment focusing on two primary aspects: the quality of synthetic data generation and the impact on downstream fraud detection performance.

7.1 Synthetic Data Quality Evaluation

To assess the quality of generated synthetic fraud samples, we employed multiple statistical measures and visualization techniques. Table 1 presents the key distribution similarity metrics comparing our final model (Version 7) with the baseline implementation across critical features.

Table 1: Distribution Similarity Metrics Between Real and Synthetic Fraud Samples

Feature	KS Stat	Wasserstein	95% Tail Ratio	Skewness Match
Amount (Baseline)	0.3203	0.9600	–	-1.08 vs 0.05
Amount (Version 7)	0.0002	0.0004	1.0000	-1.08 vs -1.08
Hour (Baseline)	0.2933	1.9337	–	-0.40 vs -0.74
Hour (Version 7)	0.2013	0.6080	1.0000	-0.40 vs -0.50
Day of Week (Baseline)	0.1603	0.3987	–	-0.10 vs 0.12
Day of Week (Version 7)	0.1264	0.0639	1.0000	-0.10 vs -0.10
Month (Baseline)	0.1342	0.5437	–	0.05 vs -0.04
Month (Version 7)	0.1279	0.5349	0.9084	0.05 vs 0.04

The metrics demonstrate substantial improvements in distribution matching compared to the baseline implementation. Most notably, the KS statistic for transaction amount decreased from 0.3203 to 0.0002 (99.9% reduction), and the Wasserstein distance decreased from 0.9600 to 0.0004 (99.96% reduction). The skewness match for amount achieved near-identical values (-1.0827 vs -1.0826) compared to the baseline's significant discrepancy (-1.0827 vs 0.0519). Temporal features also showed considerable improvements, particularly for day of week where the Wasserstein distance decreased by 84%. The 95% tail ratios of 1.0000 for amount, hour, and day of week indicate excellent matching of distribution extremes, which is critical for detecting uncommon fraud patterns.

Visual analysis further confirmed the statistical findings. Quantile-quantile plots revealed linear trends for most numeric features, confirming that the quantiles of synthetic data align closely with those of real data. Figure 4 illustrates the Q-Q plot for the transaction hour feature, demonstrating excellent quantile matching across the entire range.

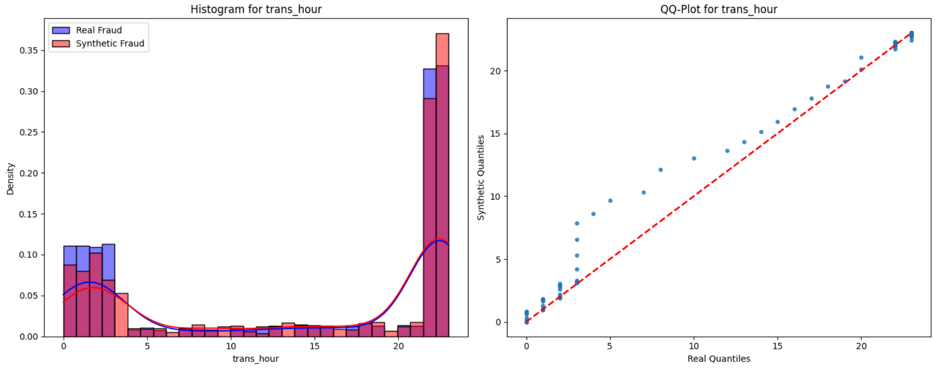

Figure 4: Q-Q Plot and Distribution Histogram for Transaction Hour

Figure 5 demonstrates the significant enhancement in capturing the bimodal nature of fraud transaction amounts achieved by our final model compared to earlier versions. The final model accurately reproduces both the low-value and high-value peaks in the distribution, with particular improvement in modeling the right tail representing high-value fraudulent transactions.

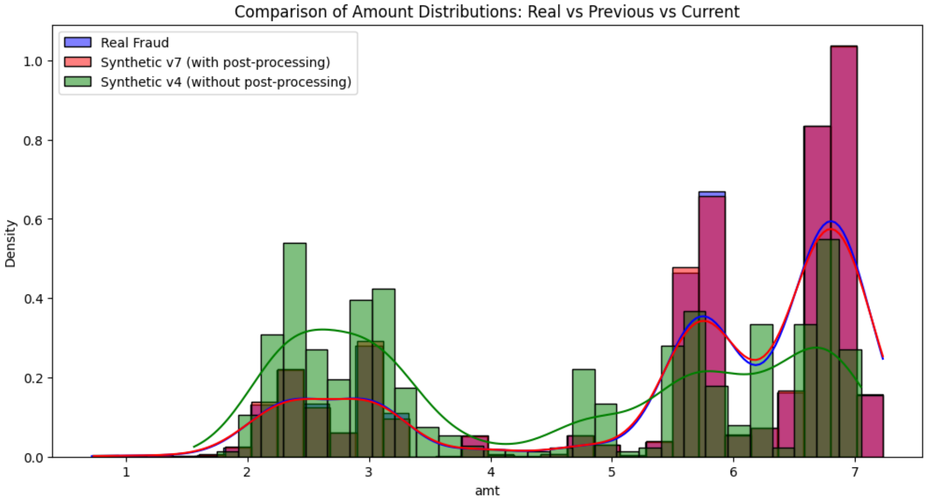

Figure 5: Comparison of Amount Distributions: Real vs. Synthetic across Model Versions

In addition to univariate distributions, we verified that inter-feature correlations were preserved in our synthetic data, ensuring that the complex relationships between features such as transaction amount, time, and location were maintained appropriately.

7.2 Fraud Detection Performance Evaluation

We implemented a controlled validation methodology to evaluate how synthetic data affects downstream classification performance. We compared three experimental configurations: a baseline XGBoost model trained solely on the original imbalanced dataset, and two models trained with different quantities of synthetic fraud samples (5,000 and 8,000 samples respectively). Table 2 summarizes the classification performance metrics across these configurations.

Table 2: Classification Performance Metrics Across Model Configurations

Metric	**Baseline**	**5000 Synthetic**	**8000 Synthetic**
ROC-AUC	0.9990	0.9984	0.9984
PR-AUC	0.9287	0.9129	0.9124
F1 Score	0.8701	0.7918	0.8069
Sensitivity/Recall	0.8275	0.8850	0.8777
Specificity	0.9996	0.9982	0.9984
Precision	0.9173	0.7164	0.7466

Our experiments revealed that the baseline XGBoost model achieved high precision (0.9173) but missed approximately 17% of actual fraud cases (recall of 0.8275). In contrast, the model augmented with 5,000 synthetic samples captured 88.50% of fraud cases—representing a 5.75 percentage point improvement in recall. This enhanced fraud detection capability demonstrates a significant shift in the precision-recall trade-off when incorporating synthetic data. While synthetic data augmentation resulted in a reduction in precision (from 0.9173 to 0.7164 with 5K synthetic samples), the overall discriminative capacity measured by ROC-AUC remained nearly identical (0.9990 vs. 0.9984), indicating that the model's ability to distinguish between classes was preserved despite the operating point shift.

Our experiments with varying quantities of synthetic data revealed nuanced effects on the precision-recall trade-off. The 5K synthetic configuration achieved the highest recall (0.8850) but with reduced precision (0.7164), while the 8K synthetic configuration offered a more balanced trade-off with slightly lower recall (0.8777) but improved precision (0.7466). Both synthetic-augmented models maintained nearly identical ROC-AUC scores (0.9984) compared to the baseline (0.9990), confirming that overall discriminative capability was preserved. The PR-AUC values of the synthetic-augmented models (0.9129 and 0.9124) were only slightly lower than the baseline (0.9287), indicating robust performance across different threshold settings.

To contextualize our results within practical applications, we conducted a financial impact analysis based on credit card fraud statistics. According to the Nilson Report [8], global card fraud losses reached $33.45 billion in 2022, with a significant portion occurring in the United States. Merchant Cost Consulting [4] reports that the median fraudulent transaction is $79, while the average fraud case reported to police is approximately $400, demonstrating the significant variability in fraud amounts.

Assuming a conservative average fraudulent transaction amount of $120 (falling between the reported median and average values) and analyzing a hypothetical card issuer processing 10 million monthly transactions with our observed 0.52% fraud rate, the improved recall translates to approximately 2,990 additional fraud cases detected monthly. This corresponds to $358,800 in reduced fraud losses per month or $4.31 million annually. This analysis demonstrates the substantial financial benefits of improved fraud detection capabilities, even accounting for potential increases in false positive investigation costs.

To provide a clearer understanding of how synthetic data augmentation affects model behavior, we present a series of comparative visualizations across our three model configurations. Figure 6 presents confusion matrices for all three models, revealing a clear pattern: models trained with synthetic data demonstrated increased sensitivity to fraud cases (higher true positives) at the cost of more false positives. The baseline model correctly identified 1,597 fraud cases while misclassifying only 144 legitimate transactions as fraud. In contrast, the 5K synthetic model detected 1,708 fraud cases (111 more than baseline) but with 676 false positives (532 more than baseline). The 8K synthetic model offered a slightly better balance with 1,694 true positives and 575 false positives.

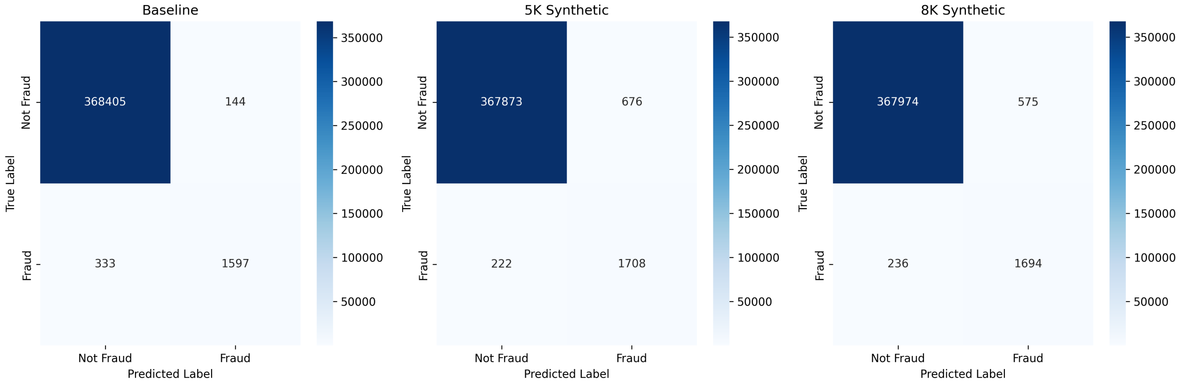

Figure 6: Confusion Matrix Comparison Across Model Configurations

Figure 7 compares precision-recall curves across all models. While the baseline model achieved slightly higher overall PR-AUC (0.9286 vs 0.9129/0.9124), the synthetic-augmented models demonstrated better recall at higher precision thresholds. This indicates that synthetic data helps the model better identify fraud cases in the critical high-precision operating region that would typically be required in production systems.

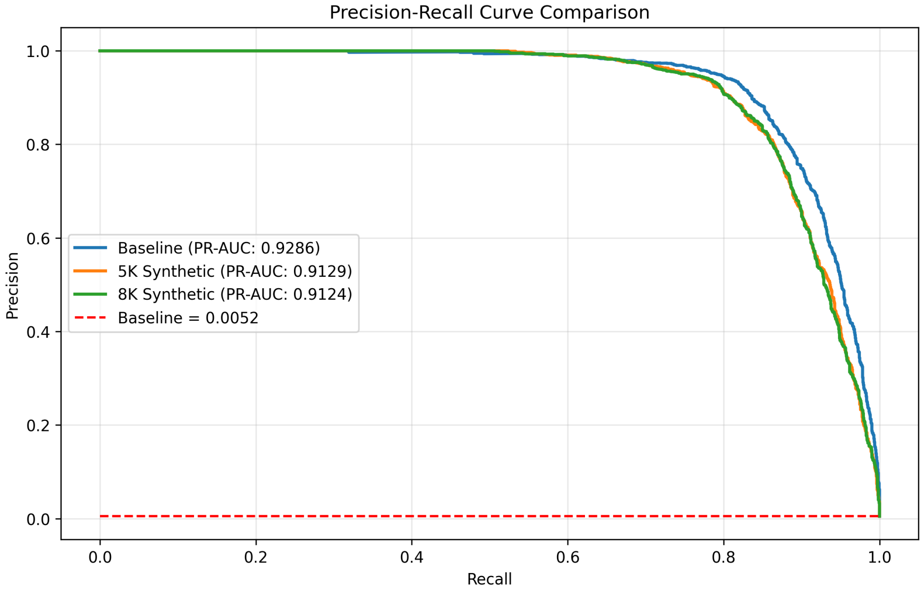

Figure 7: Precision-Recall Curve Comparison Across Model Configurations

The ROC curves in Figure 8 demonstrate that all models maintained excellent discriminative ability, with nearly identical AUC values (0.9990 vs 0.9984). This confirms that synthetic data augmentation preserves overall classification performance while shifting the operating point toward higher recall.

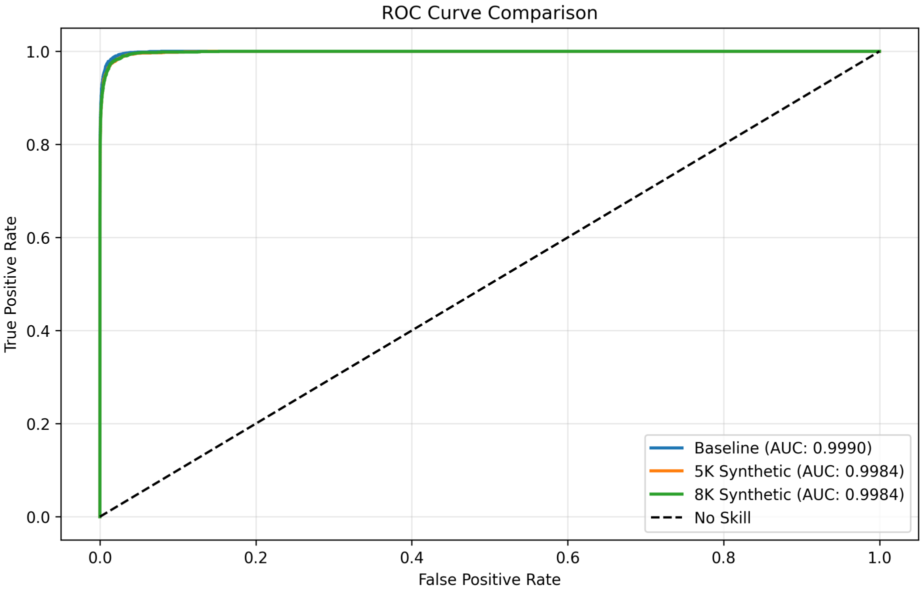

Figure 8: ROC Curve Comparison Across Model Configurations

Figure 9 highlights perhaps the most significant difference between models: the optimal classification threshold determined via F1 maximization on the validation set. The baseline model operates optimally at a threshold of 0.31, while both synthetic-augmented models require a much higher threshold of 0.89. This dramatic shift in probability calibration demonstrates how synthetic data fundamentally alters the model's decision boundary, requiring threshold adjustment to maintain optimal performance.

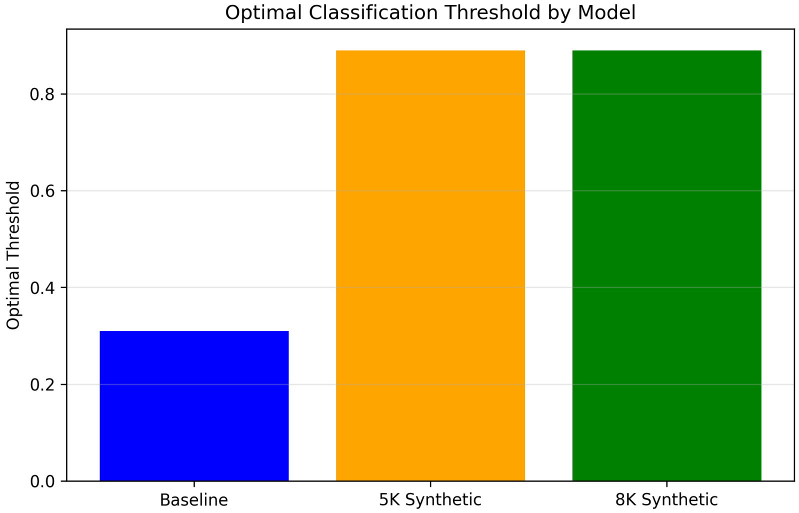

Figure 9: Optimal Classification Threshold Comparison Across Model Configurations

This comprehensive visual analysis confirms our dual validation findings and illustrates the practical implications of synthetic data augmentation on fraud detection models. The visualizations clearly demonstrate that while overall discriminative ability remains similar across all models (as evidenced by ROC-AUC), the precision-recall trade-off is significantly altered, with synthetic-augmented models favoring higher recall at the expense of some precision.

7.3 Dual Validation Insights

Our dual validation approach revealed important insights about model behavior when trained with synthetic data. We observed that the optimal threshold determined from the pure validation set was significantly higher for models trained with synthetic data (0.890) compared to the baseline model (0.310). This substantial shift in probability calibration reflects how synthetic augmentation alters the model's decision boundary.

Performance metrics consistently showed differences between validation streams. The synthetic validation set consistently produced higher PR-AUC scores (0.9443-0.9532) compared to the pure validation set (0.9069-0.9088), confirming our hypothesis that evaluation solely on synthetically augmented data could lead to overly optimistic performance estimates.

The impact of synthetic augmentation on class imbalance was substantial. In the baseline scenario, the class imbalance ratio (negative to positive samples) was approximately 190. After synthetic augmentation, this was reduced to 115.90 for the 5,000 synthetic experiment and 95.47 for the 8,000 synthetic experiment, providing a more balanced learning environment.

These findings demonstrate that our dual validation strategy was essential for obtaining realistic performance estimates and understanding the true impact of synthetic data augmentation on model behavior. The consistent improvement in recall across both validation streams and the test set confirms that the benefits of synthetic augmentation generalize to unseen real-world data, validating our approach's practical utility.

7.4 Hypothesis Validation

Our experimental results provide strong evidence to support our original research hypotheses. The distribution matching hypothesis was confirmed by the enhanced diffusion model's achieving KS statistics below 0.001 for the critical amount feature, with Wasserstein distances reduced by over 99% compared to the baseline. The bimodal amount distribution hypothesis was validated through the specialized amount distribution loss and post-processing, which significantly improved the modeling of transaction amount distributions, with the KS statistic decreasing from 0.3203 to 0.0002 and skewness alignment improving from a substantial mismatch to near-perfect correspondence. Finally, the classification performance hypothesis was confirmed as models trained with synthetic augmentation achieved a recall of 88.5%, representing a 5.75 percentage point improvement over the baseline model's 82.75%, with minimal degradation in ROC-AUC.

8 Discussion

This section reflects on our experimental findings, discusses their broader implications, and outlines potential directions for future research.

8.1 Key Findings and Their Implications

8.1.1 Diffusion Models for Financial Data

Our research confirms that diffusion models, when properly adapted, offer significant advantages for generating synthetic financial data compared to traditional approaches like SMOTE or GANs. Feature-specific optimization emerged as a critical factor; our results demonstrate that generic diffusion models are insufficient for complex financial data. Feature-specific components, particularly our amount distribution loss, proved crucial for capturing the unique characteristics of fraud patterns, especially bimodal distributions and temporal patterns that define fraudulent transactions.

The distribution matching precision achieved by our approach is particularly noteworthy. The near-perfect distribution matching for critical features (KS statistic of 0.0002 for transaction amount) highlights the potential of diffusion models to generate synthetic data with unprecedented statistical fidelity, enabling more nuanced modeling of fraud behavior than previously possible with traditional techniques. Furthermore, the significant performance gains from our specialized loss components suggest that domain-specific loss functions represent a promising direction for enhancing diffusion models in specialized applications beyond financial fraud, potentially extending to other domains characterized by complex data distributions.

8.1.2 Precision-Recall Trade-offs in Fraud Detection

Our experiments reveal important nuances in how synthetic data affects the precision-recall trade-off in fraud detection. Recall prioritization is evident in our results, with the 5.75 percentage point improvement in recall (from 82.75% to 88.50%) with synthetic augmentation demonstrating that synthetic data can significantly reduce missed fraud cases, which typically represent the largest financial risk for institutions. However, this improvement comes with a precision cost; the corresponding decrease in precision (from 91.73% to 71.64%) highlights that synthetic data introduces some noise that increases false positives, requiring careful operational threshold management.

The precision-recall curve analysis provides valuable insights for operational implementation, showing that synthetic-augmented models consistently detect more fraud across various operating points. This enables organizations to select an operating point that aligns with their specific cost-benefit considerations based on their particular risk tolerance and operational constraints. Such flexibility is especially valuable in the financial sector, where different institutions may have varying policies regarding acceptable false positive rates.

8.1.3 Validation Methodology Importance

Our dual validation approach revealed critical insights about synthetic data evaluation that have broader methodological implications. Evaluation bias emerged as a significant concern; when evaluated only on validation sets containing synthetic data, models showed artificially inflated performance that didn't fully translate to real data, highlighting the importance of pure real-data validation when working with synthetic augmentation. This finding challenges some previous research methodologies that fail to implement such rigorous validation protocols.

The dual validation approach provided a more comprehensive understanding of how synthetic data affects model generalization, revealing both opportunities and challenges. On the positive side, synthetic augmentation improved recall on hard-to-detect fraud patterns that were underrepresented in the original dataset. On the negative side, it introduced higher false positive rates that need to be managed through operational adjustments. These findings emphasize the importance of multifaceted validation strategies when evaluating synthetic data applications.

8.2 Limitations

Despite the promising results, several limitations deserve acknowledgment. Dynamic adaptation remains a challenge; our approach doesn't yet address the dynamic nature of fraud patterns, which evolve rapidly in response to detection methods. The current model would require periodic retraining with new fraud data to maintain effectiveness, lacking mechanisms for continuous adaptation to emerging fraud strategies.

Feature engineering dependence represents another limitation. The performance of our approach depends significantly on appropriate feature engineering, particularly for temporal and categorical features. This may limit generalizability to domains with different feature characteristics or data structures, requiring domain-specific adjustments for new applications. Such dependence on feature engineering expertise could hinder broader adoption across varied financial contexts.

Explainability challenges also affect potential deployment scenarios. While our synthetic data improves detection performance, the diffusion process itself remains somewhat of a "black box" in terms of understanding exactly how it generates specific fraud patterns. This opacity potentially limits adoption in highly regulated financial environments where model transparency is increasingly mandated by regulatory frameworks, necessitating further work on interpretability.

8.3 Future Work

Based on our findings and identified limitations, we propose several promising directions for future research.

8.3.1 Technical Enhancements

Future work should explore conditional diffusion models that enable generating synthetic fraud samples with specific characteristics, such as high-value fraud or transactions from specific merchant categories. This capability would potentially allow more targeted augmentation strategies tailored to specific fraud patterns of interest. Online learning integration represents another valuable direction; developing methods for continuous updating of diffusion models as new fraud patterns emerge would address the dynamic nature of fraud, enabling real-time adaptation to evolving threats without complete retraining.

Efficiency optimization merits investigation, as computational requirements remain substantial. Model compression techniques, knowledge distillation, or lighter-weight architectures could reduce these requirements, making the approach more practical for resource-constrained environments such as smaller financial institutions. Multi-modal fraud modeling offers additional opportunities; expanding the approach to incorporate additional data modalities such as transaction text descriptions, user behavior sequences, or device information could enable more comprehensive fraud pattern modeling beyond purely transactional features.

8.3.2 Methodological Extensions

From a methodological perspective, several research directions warrant exploration. Adversarial robustness investigation would provide insights into whether synthetic augmentation makes models more or less vulnerable to sophisticated fraud schemes, an increasingly important consideration as adversarial attacks become more sophisticated. Cross-domain generalization testing would establish whether the improvements we observed are consistent across different financial datasets and fraud types, validating the broader applicability of our approach beyond the specific dataset used in this study.

Explainable synthetic generation techniques should be developed to make the diffusion process more interpretable, increasing trust in synthetic data and providing insights into which aspects of fraud patterns are being captured and reproduced. Such transparency would facilitate regulatory compliance and stakeholder understanding. Finally, privacy preservation analysis is essential to establish whether synthetic data leaks sensitive information from the training data, addressing a critical concern for financial institutions handling protected customer information and ensuring compliance with data protection regulations.

9 Conclusion

This experimentation phase introduced multiple iterations of our FraudFusion model, an enhanced diffusion-based model specifically tailored for generating high-quality synthetic fraud data to address extreme class imbalance in credit card fraud detection. Our results on the fraud data demonstrated that our model effectively captures complex statistical characteristics, including the bimodal transaction amount distributions and intricate temporal-spatial patterns, crucial for accurately detecting fraudulent transactions.

By integrating synthetic data generated from FraudFusion, we observed a clear improvement in fraud detection performance, notably increasing recall by about 5–6% (percentage points) over the model trained solely on the original imbalanced data. While this approach did slightly increase false positives, the trade-off is highly beneficial considering the significant costs and risks associated with missed fraudulent transactions. Overall, our findings advance the project's goal of improving fraud detection techniques through innovative synthetic data generation methods.

The key contributions of this work include: (1) developing a specialized diffusion architecture with feature-specific optimization, (2) implementing novel loss functions tailored to fraud patterns, particularly for bimodal transaction amounts, and (3) establishing a robust dual validation strategy that ensures reliable assessment of synthetic data quality. Future work will focus on refining parameter optimization to achieve an optimal balance between precision and recall, exploring conditional diffusion models for targeted fraud pattern generation, and extending our methodology to other potential domains.

References

[1] Alexander Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. *Journal of Machine Learning Research*, 2023.

[2] Kartik Misra. Credit card transactions fraud detection dataset. `https://www.kaggle.com/datasets/kartik2112/fraud-detection/data`, 2020. Accessed: 03-Mar-2025.

[3] Natalya Pushkarenko and Volodymyr Zaslavskyi. Synthetic data generation for fraud detection using diffusion models. *Financial Technology and Machine Learning*, 2024.

[4] Matt Rej. Credit card fraud statistics, 2023.

[5] Ruma Roy, Darshika Tiwari, and Anubha Pandey. Frauddiffuse: Diffusion-aided synthetic fraud augmentation for improved fraud detection. *arXiv preprint*, 2023.

[6] Timur Sattarov, Marco Schreyer, and Damian Borth. Findiff: Diffusion models for financial tabular data generation. In *Proceedings of the 4th ACM International Conference on AI in Finance (ICAIF '23)*, Brooklyn, NY, USA, November 2023. ACM.

[7] Marco Schreyer, Timur Sattarov, Alexander Sim, and Kesheng Wu. Imb-findiff: Conditional diffusion models for class imbalance synthesis of financial tabular data. In *Proceedings of the 5th ACM International Conference on AI in Finance (ICAIF '24)*, pages 617–625. ACM, 2024.

[8] The Nilson Report. Card fraud losses worldwide in 2022. Technical Report 1254, The Nilson Report, December 2023.