FraudFusion: Enhanced Diffusion Models for Synthetic Fraud Data Generation

Anish Rao, Raghu Ram Sattanapalle

March 19, 2025

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:** With fraudulent transactions typically representing less than 0.1% of all credit card activity, standard classification models tend to be biased toward the majority class, resulting in unacceptably high false negative rates.
- **Complex Statistical Relationships:** 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:

- **Extreme Imbalance:** 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.
- **High-Dimensional Feature Space:** 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.

- **Complex Temporal and Spatial Patterns:** Fraudulent activity often shows distinctive patterns in transaction timing, location, and amount that must be properly captured for effective detection.
- **Asymmetric Misclassification Costs:** 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:

- **Financial Impact:** The financial services 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.
- **Consumer Protection:** Undetected fraud not only harms financial institutions but also creates substantial distress for consumers whose accounts are compromised. Improved fraud detection directly benefits consumer financial security.
- **Operational Efficiency:** Reducing false positives while maintaining high sensitivity decreases the burden on fraud investigation teams, allowing more efficient allocation of human resources.
- **Methodological Innovation:** 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:

2.1 Diffusion Models for Tabular Data

2.1.1 FraudDiffuse

Roy et al. [3] 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.

2.1.2 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.

2.2 Financial Data Synthesis

2.2.1 FinDiff

"FinDiff: Diffusion Models for Financial Tabular Data Generation" [4] 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.

2.2.2 Imb-FinDiff

"Imb-FinDiff: Conditional Diffusion Models for Class Imbalance Synthesis of Financial Tabular Data" [5] 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.

2.2.3 Dual-Track Diffusion Approach

Pushkarenko and Zaslavskyi [2] 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.

2.3 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.

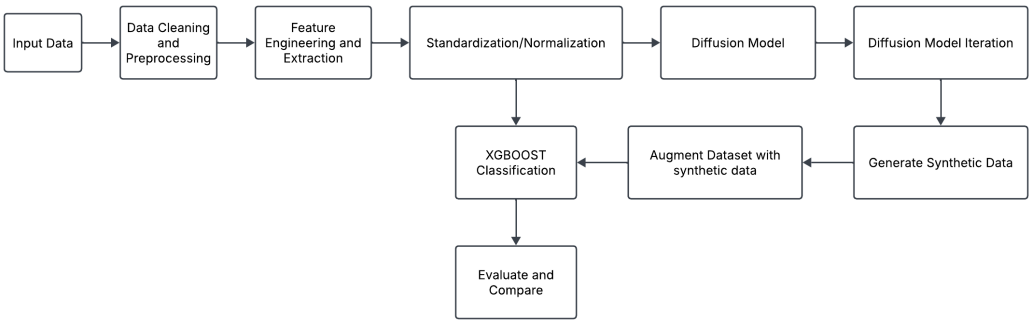

Figure 1: Experimental Workflow: From Data Preprocessing to Model Evaluation

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
- **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:

- **H1: 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).
- **H2: 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.
- **H3: 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:**
  - **SMOTE (Synthetic Minority Over-sampling Technique):** Limited by linear interpolation between neighboring samples, failing to capture complex feature distributions and relationships common in fraud data.
  - **ADASYN (Adaptive Synthetic Sampling):** Similar to SMOTE but focusing on difficult examples, yet still inadequate for capturing complex dependencies between fraud features.
- **Deep Generative Models:**
  - **GANs (Generative Adversarial Networks):** Including WGAN, WCGAN, and WCGAN-GP variants. Despite their power, these models suffer from training instability and mode collapse, particularly problematic for the mixed numerical and categorical features in fraud data.
  - **VAEs (Variational Autoencoders):** More stable than GANs but often produce samples with blurred feature distributions, potentially losing critical fraud patterns.
  - **Diffusion Models:** Demonstrated superior performance in generating high-quality tabular data with preserved statistical properties.
- **Classifier Models:**
  - **Tree-based Models:** Including Random Forest and gradient boosting frameworks (XGBoost, LightGBM), which typically perform well on tabular data and can handle class imbalance.
  - **Neural Networks:** Including MLPs and more specialized architectures 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:

- **Mathematical Foundation:** XGBoost implements a regularized form of gradient boosting that minimizes the following objective function:

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

  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 penalizing model complexity. This regularization helps prevent overfitting, which is crucial when dealing with the complex patterns found in fraud data.

- **Handling Class Imbalance:** XGBoost provides native support for imbalanced datasets through the `scale_pos_weight` parameter, which scales the gradient for the minority class:

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

  This approach addresses the severe class imbalance in fraud detection tasks, even before applying our synthetic data generation strategy.

- **Feature Importance Analysis:** XGBoost calculates feature importance scores based on their contribution to performance improvement, providing 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.

- **Efficient Training:** The implementation includes optimizations such as a sparsity-aware split finding algorithm and a distributed weighted quantile sketch for handling sparse data, allowing effective utilization of computational resources when training on large financial datasets.

- **Established Benchmark:** XGBoost was utilized in the original FraudDiffuse paper (Roy et al., 2023 [3]) as well as other recent fraud detection studies, providing a consistent benchmark for evaluating the quality of synthetic data generation techniques.

Consistent with our previous experimental design, we maintained three distinct XGBoost configurations for this iteration:

- **Baseline XGBoost:** Trained only on the original imbalanced data to establish performance benchmarks

- **Augmented XGBoost:** Trained on a combination of original data and synthetic fraud samples generated by our enhanced FraudDiffuse model

- **Controlled XGBoost:** 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.

4.2.3 Generative Model Evolution: Enhanced FraudDiffuse

For our generative model, we continued development of the FraudDiffuse approach (Roy et al., 2023 [3]) while implementing significant enhancements based on limitations identified in our previous iterations.

The original FraudDiffuse components we retained include:

- **Adaptive Non-Fraud Prior:** Unlike vanilla diffusion models that use a standard Gaussian prior, FraudDiffuse leverages the distribution of legitimate transactions as the prior. This forces the model to learn subtle patterns that distinguish fraudulent transactions near the decision boundary, where most challenging fraud cases reside.

  The non-fraud prior parameters $\mathcal{N}(\mu_{nf}, \Sigma_{nf})$ are estimated using non-fraud training data statistics. The forward process adds noise based on this prior, and the reverse process samples x_T from $\mathcal{N}(\mu_{nf}, \Sigma_{nf})$ instead of $\mathcal{N}(0, I)$.

- **Probability-Based Loss Function:** FraudDiffuse employs a specialized loss function that improves error estimation for the diffusion process, leading to more accurate modeling of the fraud distribution.

  The probability-based loss L_{prior} is formulated as:

$$L_{prior} = 2 \times P(Z \leq |z\text{-}score|) = 1 - 2 \times P(Z \geq |z\text{-}score|) \tag{4}$$

  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.

- **Contrastive Regularization:** Incorporating triplet loss ensures that generated synthetic fraud samples remain close to real fraud examples while being distinct from legitimate transactions, addressing potential overfitting issues.

  The triplet loss is defined as:

$$L_{triplet} = \max(0, d(\hat{x}_f, x_f) - d(\hat{x}_f, x_{nf}) + \text{margin}) \tag{5}$$

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

The original FraudDiffuse loss function is:

$$L_{fraudDiffuse} = L_{norm} + w_1 \times L_{prior} + w_2 \times L_{triplet} \tag{6}$$

where L_{norm} is the standard mean squared error between added noise and predicted noise:

$$L_{norm} = \mathbb{E}_{x_0, \epsilon, t}\left[\frac{\|\epsilon - \epsilon_\theta\|^2}{2}\right] \tag{7}$$

4.2.4 Progressive Model Development

Based on systematic performance evaluation, our generative model evolved through multiple versions, each addressing specific limitations identified in previous iterations:

- **Version 1 (Baseline):** Initial implementation of FraudDiffuse as described by Roy et al., which demonstrated the basic capability to generate synthetic fraud samples but struggled with the bimodal transaction amount distribution in the Sparkov dataset
- **Version 2:** Added engineered feature range constraints to address out-of-range generation issues identified in baseline evaluation, particularly for temporal features
- **Version 3:** Implemented cyclical encoding for temporal features and feature-specific initialization based on distribution analysis of Version 2 outputs, which revealed temporal pattern inconsistencies
- **Version 4:** Focused on training stability improvements after observing convergence issues in Version 3, including gradient clipping, adaptive learning rate scheduling, and NaN detection mechanisms
- **Version 5:** Enhanced bimodal distribution modeling after statistical testing revealed inadequate modeling of the transaction amount's distinctive dual-peak pattern
- **Version 7 (Final):** Implemented comprehensive post-processing and targeted feature enhancements based on thorough statistical evaluation of Version 5 performance

Each model version was systematically evaluated using our dual-track validation approach, comparing both distribution quality metrics and downstream classification impact. This iterative process demonstrated clear progression in synthetic data quality, with the final Version 7 significantly outperforming earlier implementations in both statistical similarity to real data and ability to improve fraud detection when used for classifier augmentation.

4.2.5 Training Stability Enhancements

A key challenge identified in previous iterations was training instability with diffusion models on complex, mixed-type financial data. In this experimentation phase, we implemented the following stability techniques:

- **Gradient Clipping:** Aggressive gradient clipping with max_norm=0.5 to prevent exploding gradients
- **Adaptive Learning Rate:** ReduceLROnPlateau scheduler with factor=0.7 and patience=15
- **NaN Detection and Recovery:** Comprehensive exception handling throughout the training loop with fallback loss calculations
- **Value Clamping:** Strategic clamping of intermediate values to prevent numerical instabilities
- **Batch Size Optimization:** Reduced batch size (32) for better stability with small fraud datasets
- **Weight Decay:** L2 regularization (1e-5) to prevent overfitting

4.2.6 Diffusion Process Parameter Refinements

Based on extensive experimentation in previous iterations, we refined the diffusion process parameters:

- **Noise Schedule:** Linear beta schedule from $\beta_{start} = 10^{-4}$ to $\beta_{end} = 0.02$ over $T_{train} = 800$ steps
- **Generation Steps:** Reduced to $T_{gen} = 600$ for inference efficiency without quality degradation
- **Adaptive Noise Reduction:** Progressive noise reduction in later generation steps scaled by $t_{step}/200$ for $t < 200$

These refinements collectively improved both model stability and synthetic data quality compared to our previous iterations, as demonstrated by improvements in both distribution metrics and downstream classification performance.

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:

- **Source:** `https://www.kaggle.com/datasets/kartik2112/fraud-detection/data`
- **Access:** Publicly available with no 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

- **Dataset Size:** The dataset comprises 1,296,675 transactions after initial processing.
- **Features:** The dataset includes 23 original features, categorized as:
  - **Transaction details:** Amount (`amt`), date/time (`trans_date_trans_time`), merchant information
  - **Cardholder demographics:** Age, gender, job (494 unique categories)
  - **Geospatial information:** Customer latitude/longitude, merchant latitude/longitude, city population
  - **Categorical features:** Merchant category (14 categories), gender (2 categories), state (51 categories)
- **Class Imbalance:** As can be seen in 2, only 0.52% of transactions are fraudulent, creating an extreme imbalance ratio of approximately 1:192.

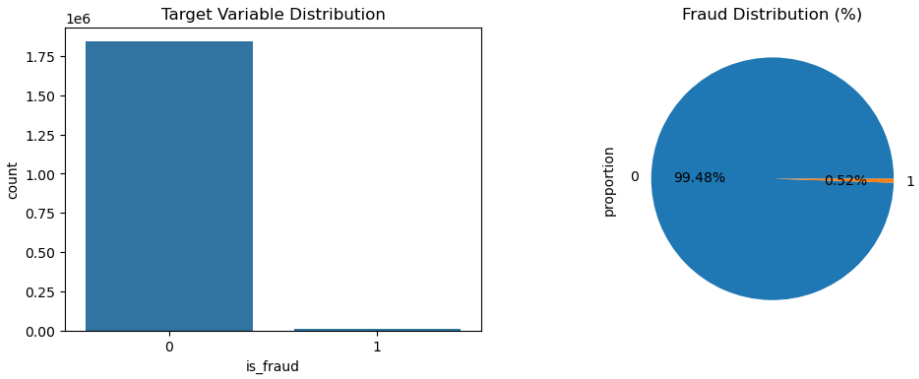

Figure 2: Class Imbalance

- **Key Observations:** Our exploratory data analysis revealed several important patterns:
  - **Amount distribution:** Fraud transactions display distinctive patterns in transaction amounts
  - **Temporal patterns:** Fraud occurs more frequently during certain hours (especially early morning) and during the first six months of the year
  - **Geographical clustering:** As can be seen in 3, fraud cases cluster in specific geographic locations

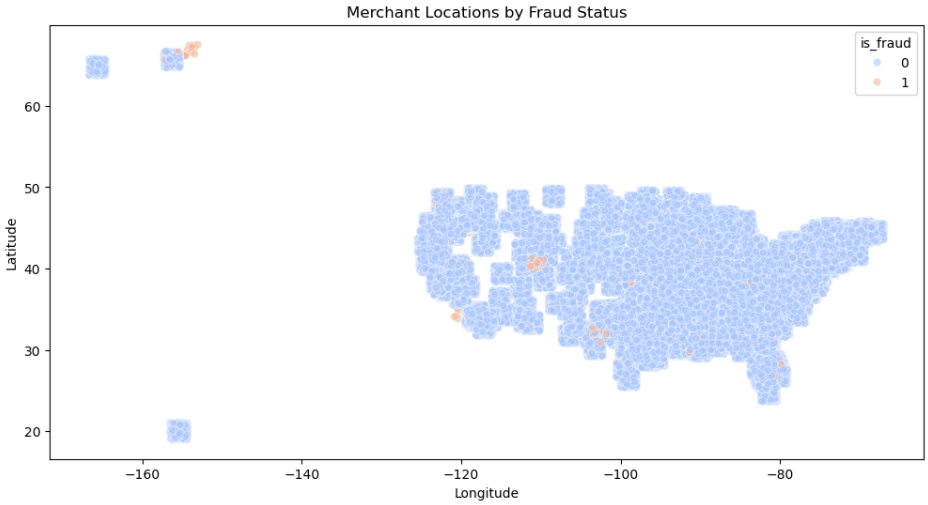

Figure 3: Fraud Locations

  - **Age-related patterns:** Higher fraud rates in the 76-85 (0.92%) and 86-95 (0.87%) age groups

5.3 Preprocessing Steps

- **Data Cleaning:**
  - No missing values were found in the dataset

  - Identifier columns (`cc_num`, `first`, `last`, `trans_num`) and redundant timestamps (`unix_time`) were removed to prevent data leakage

- **Feature Transformation:**
  - **Log Transformation:** Applied log transformation (using `np.log1p`) to skewed numerical features:
    * Transaction amount (`amt`) - to normalize the heavily right-skewed distribution
    * City population (`city_pop`) - to reduce the impact of population outliers
  - **Standardization:** All numerical features were then standardized using scikit-learn's `StandardScaler` to zero mean and unit variance

- **Temporal Feature Engineering:** We extracted and transformed time-based features:
  - Hour of day from transaction timestamps
  - Day of week (analysis revealed different fraud patterns by weekday)
  - Month (identified seasonal patterns with higher fraud in first half of year)
  - Sine-cosine transformations applied to preserve cyclical nature:

$$t_{sin} = \sin\left(\frac{2\pi \cdot t}{P}\right), \quad t_{cos} = \cos\left(\frac{2\pi \cdot t}{P}\right) \tag{8}$$

    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)

- **Feature Importance Analysis:** We evaluated the predictive power of features using Mutual Information:
  - Transaction amount (`amt`) - showing highest predictive power (MI: 0.0158)
  - Geographic coordinates (`lat`, `long`) - showing moderate predictive value
  - `city_pop` - providing limited signal (MI: 0.00302)

- **Categorical Encoding:** We implemented a tiered approach based on cardinality:
  - **One-Hot Encoding:** For low-cardinality features (`category`, `gender`)
  - **Target Encoding:** For medium-cardinality features (`state`)
  - **Frequency Encoding:** For high-cardinality features (`job`, `merchant`)

- **Distance Calculation:** We calculated geographical distance between customer and merchant locations as an additional engineered feature for fraud detection.

- **Specialized Handling for Fraud Patterns:**
  - **Bimodal Amount Modeling:** We implemented specialized handling for transaction amounts after observing their distinctive bimodal distribution in fraudulent transactions, with peaks at both low-value and high-value ranges
  - **Feature Range Constraints:** For temporal features, we computed observed min/max values in standardized space and added constraints to ensure generated values remained within realistic bounds

  - **Data Augmentation Strategy:** We developed specialized augmentation techniques for key fraud indicators, particularly focusing on preserving the bimodal patterns in transaction amounts and the distinctive temporal patterns of fraud occurence

- **Data Partitioning:** The dataset was split using a two-step stratified sampling approach to maintain the fraud-to-legitimate ratio across all partitions:

  - **Training set (65%):** Used for model training
  - **Validation set (15%):** Used for hyperparameter tuning and early stopping
  - **Test set (20%):** Reserved for final evaluation

  The stratified approach (using `sklearn.model_selection.train_test_split` with `stratify=y`) ensures that each split maintains the same class distribution of approximately 0.52% fraudulent transactions, which is critical for training and evaluating models on highly imbalanced data.

5.4 Synthetic Data Strategy

Our preprocessing pipeline was designed with synthetic data generation in mind:

- **Distribution Analysis:** Special attention was given to fraud-specific distributions, particularly:
  - Bimodal transaction amount patterns
  - Temporal fraud patterns by hour of day
  - Geographic clustering of fraud cases
- **Data Leakage Prevention:** We identified sensitive fields requiring special handling:
  - Transaction identifiers (`trans_num`)
  - Credit card numbers (`cc_num`)
  - Exact timestamps (`trans_date_trans_time`)
- **Dual Validation Strategy:** We implemented a two-track validation approach:
  - "Pure" validation set with only real data
  - "Synthetic" validation set incorporating synthetic samples
- **Quality Control Framework:** We developed metrics to evaluate synthetic data quality:
  - Statistical distribution matching tests
  - Feature correlation preservation metrics
  - Temporal and spatial pattern maintenance validation
- **Enhanced Feature Engineering:** Our approach to feature engineering was guided by specific requirements for synthetic data generation:
  - We implemented specialized handling for the bimodal transaction amount distribution observed in fraudulent transactions, using distribution-aware initialization during generation and quantile-based distribution matching in post-processing

- For temporal features, we enforced range constraints based on observed patterns in real fraud data, ensuring generated samples maintained realistic time distributions
- We developed feature-specific transformation techniques to preserve the complex relationships between features that distinguish fraud from legitimate transactions

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 Training and Validation Process

6.1 Architecture and Configuration

Our enhanced FraudDiffuse model represents a significant evolution from the baseline architecture described by Roy et al. [3]. The final architecture incorporates several innovative components designed specifically to address the unique challenges of generating realistic credit card fraud transactions.

6.1.1 FraudDiffuse Neural Network Architecture

The core of our implementation is the `CombinedNoisePredictor` neural network, which features:

- **Multi-modal Input Processing:** Our architecture integrates three distinct data types:
  - Normalized numerical features (11 dimensions)
  - Embedded categorical features (8 categories with varying cardinality)
  - Specialized cyclic encodings for temporal features (8 dimensions)
- **Embedding Layers:** Each categorical feature is processed through dedicated embedding layers:
$$e_{i,j} = \text{Embedding}_j(x_{cat,i,j}) \tag{9}$$
where $x_{cat,i,j}$ represents the j-th categorical feature for the i-th sample, and embeddings are learned during training with dimension 4 per feature.
- **Network Architecture:** The model employs a four-layer feed-forward structure:
  - Combined input dimension: $11 + (8 \times 4) + 8 + 1 = 52$ (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 (replacing SiLU from earlier versions)
  - Dropout with rate 0.1 for regularization
- **Weight Initialization:** Xavier uniform initialization for all layers to ensure stable gradient flow during training

The forward pass of our model can be expressed as:

$$\begin{aligned} x_{input} &= \text{Concat}[x_{num}, x_{cat\_embedded}, x_{cyclic}, t_{norm}] \\ h_1 &= \text{Dropout}(\text{LayerNorm}(\text{ReLU}(W_1 x_{input} + b_1))) \\ h_2 &= \text{Dropout}(\text{LayerNorm}(\text{ReLU}(W_2 h_1 + b_2))) + 0.1 \times h_1 \\ h_3 &= \text{Dropout}(\text{LayerNorm}(\text{ReLU}(W_3 h_2 + b_3))) + 0.1 \times h_2 \\ \hat{\epsilon} &= W_4 h_3 + b_4 \end{aligned} \tag{10}$$

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

6.1.2 Specialized Feature Handling

Our model incorporates domain-specific components to address the unique characteristics of fraud data:

- **Bimodal Amount Modeling:** Transaction amount, a critical fraud indicator, receives specialized treatment through:
  - Bimodal initialization during generation
  - KDE-based peak detection for distribution modeling
  - Quantile-based distribution matching in post-processing
- **Feature-Weighted Learning:** We implemented feature-specific weighting in the loss function:
  - Transaction amount (weight: 1.8)
  - Transaction hour (weight: 1.3)
  - Transaction month and day of week (weight: 1.1 each)
  - Other features (weight: 1.0)

6.2 Training Process

6.2.1 Dataset Preparation and Splitting

Our dataset was split using stratified sampling to maintain the fraud-to-legitimate ratio:

- **Training set (65%):** 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
- **Validation set (15%):** Used only for model evaluation during development
- **Test set (20%):** Reserved exclusively for final performance evaluation

This strict separation ensured no data leakage between the diffusion model training and downstream classification evaluation. No extensive hyperparameter tuning was needed for the XGBoost classifier, as the model achieved excellent performance with the initial configuration.

6.2.2 Loss Function Components

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

$$\mathcal{L}_{total} = \mathcal{L}_{norm} + w_1 \times \mathcal{L}_{prior} + w_2 \times \mathcal{L}_{triplet} + \lambda_{eng} \times \mathcal{L}_{eng} + \lambda_{amt} \times \mathcal{L}_{amt} \tag{11}$$

Where each component addresses a specific aspect of the synthetic data quality:

- **Feature-Weighted $\mathcal{L}_{norm}$:** Enhanced mean squared error between true and predicted noise with feature-specific weights
- **Non-Fraud Prior Loss ($\mathcal{L}_{prior}$):** Forces the model to learn subtle patterns distinguishing fraud from legitimate transactions
- **Triplet Loss ($\mathcal{L}_{triplet}$):** Contrastive component that ensures synthetic fraud samples remain close to real fraud while distant from non-fraud samples
- **Engineered Range Loss ($\mathcal{L}_{eng}$):** Constrains temporal features to realistic ranges based on observed fraud patterns
- **Amount Distribution Loss ($\mathcal{L}_{amt}$):** Specialized component that enforces realistic bimodal distribution for transaction amounts with heightened emphasis on higher-value fraud:

$$\begin{aligned} \mathcal{L}_{amt} = & |\mu_{gen} - \mu_{real}| + \\ & 1.0 \times |q_{50,gen} - q_{50,real}| + \\ & 3.0 \times |q_{75,gen} - q_{75,real}| + \\ & 5.0 \times |q_{90,gen} - q_{90,real}| + \\ & 8.0 \times |q_{95,gen} - q_{95,real}| + \\ & 4.0 \times |skew_{gen} - skew_{real}| \end{aligned} \tag{12}$$

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 $\lambda_{amt} = 0.20$.

6.2.3 Training Stability Techniques

Training diffusion models on complex, mixed-type financial data presented significant stability challenges. We implemented several techniques to address these issues:

- **Gradient Clipping:** Aggressive gradient clipping with max_norm=0.5 to prevent exploding gradients
- **Adaptive Learning Rate:** ReduceLROnPlateau scheduler with factor=0.7 and patience=15
- **NaN Detection and Recovery:** Comprehensive exception handling throughout the training loop with fallback loss calculations
- **Value Clamping:** Strategic clamping of intermediate values to prevent numerical instabilities
- **Batch Size Optimization:** Reduced batch size (32) for better stability with small fraud datasets

- **Weight Decay:** 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.

6.3 Hyperparameter Tuning

6.3.1 Diffusion Process Hyperparameters

The diffusion process itself required careful tuning to ensure high-quality synthetic samples:

- **Noise Schedule:** Linear beta schedule from $\beta_{start} = 10^{-4}$ to $\beta_{end} = 0.02$ over $T_{train} = 800$ steps
- **Generation Steps:** Reduced to $T_{gen} = 600$ for inference efficiency without quality degradation
- **Adaptive Noise Reduction:** Progressive noise reduction in later generation steps scaled by $t_{step}/200$ for $t < 200$

6.3.2 Iterative Model Development

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

6.3.3 Distribution-Aware Initialization and Post-Processing

A key innovation in our final model is the distribution-aware initialization and post-processing pipeline:

- **Initialization:** 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

- **Distribution Transformation:** After generation, we apply quantile-based distribution matching:
$$x_{matched} = F^{-1}_{target}(F_{source}(x_{generated})) \tag{13}$$
where F represents the empirical CDF function

- **Range Enforcement:** 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.

7 Results

We evaluated our enhanced FraudDiffuse model through a comprehensive assessment focusing on both synthetic data quality and downstream classification performance. This dual evaluation strategy ensures that our approach generates not only statistically accurate fraud samples but also provides meaningful improvements to fraud detection capability.

7.1 Synthetic Data Quality Evaluation

We evaluate the quality of synthetic fraud samples using multiple statistical measures and visual analysis techniques:

7.1.1 Distribution Metrics

To quantify the similarity between real and synthetic distributions, we employed several complementary metrics:

- **KS Statistic & Anderson–Darling Test:** These tests compare the cumulative distributions of real and synthetic data. Lower values indicate better distribution matching.

- **Wasserstein & Energy Distances:** These metrics measure the "transportation cost" between distributions, providing a robust measure of distributional similarity even for multimodal data.

- **Tail Ratios (95th & 99th Percentiles):** Ratio of synthetic to real data percentiles, with values closer to 1.0 indicating better matching of extreme values—critical for fraud detection.

- **Statistical Moments:** Comparison of mean ratio, standard deviation ratio, and skewness preservation across distributions.

Table 1 summarizes these distribution metrics across key features for our final model (Version 7) compared to the baseline implementation:

These metrics demonstrate substantial improvements in distribution matching compared to the baseline, with the most dramatic improvement in the critical amount feature. 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 is particularly impressive, with Version 7 achieving nearly identical values (-1.0827 vs -1.0826) compared to the baseline's significant discrepancy (-1.0827 vs 0.0519).

Table 1: Distribution Similarity Metrics: Real vs. 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

For temporal features, we also observe notable improvements, particularly for day of week where the Wasserstein distance decreased by 84% (from 0.3987 to 0.0639). The perfect 95% tail ratios (1.0000) for amount, hour, and day of week indicate excellent matching of distribution extremes, which is crucial for detecting uncommon fraud patterns.

7.1.2 Visual Comparisons

Visual analysis provides intuitive confirmation of our statistical findings:

- **Quantile-Quantile Plots:** The linear trends in QQ-plots for most numeric features indicate that the quantiles of the synthetic data align well with those of the real data. For example, as illustrated in Figure 4, the 'trans_hour' feature shows excellent quantile matching across the entire range.

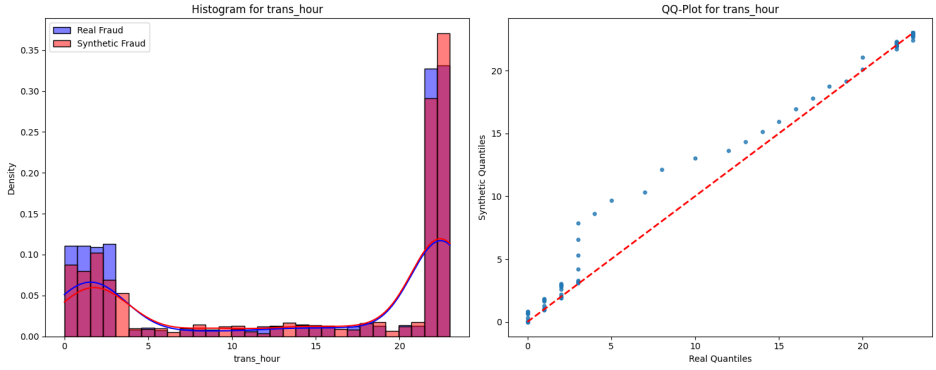

Figure 4: QQ-Plot and Distribution Histogram for Transaction Hour

- **Amount Distribution Improvement:** Our model progression shows significant enhancements in capturing the bimodal nature of fraud transaction amounts. Figure 5 demonstrates how Version 7 more accurately reproduces both peaks in the distribution compared to earlier versions, particularly in the right tail representing higher-value fraud transactions.

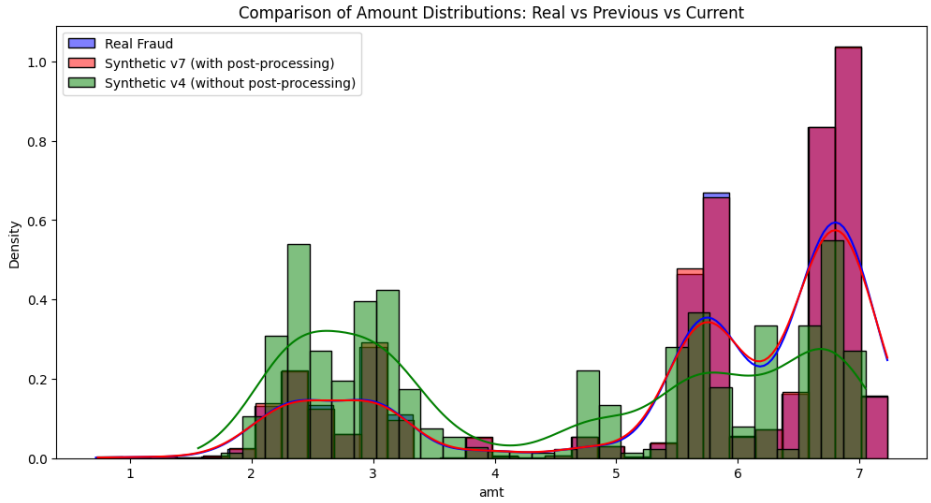

Figure 5: Comparison of Amount Distributions: Real vs. Synthetic Versions

- **Correlation Structure Preservation:** We verified that the inter-feature correlations were preserved in our synthetic data, ensuring that relationships between features like transaction amount, time, and location were maintained.

7.2 Fraud Detection Performance Improvement

The ultimate test of our synthetic data quality is its effect on downstream fraud detection performance. We evaluated several experimental configurations:

- **Baseline:** XGBoost trained only on the original imbalanced dataset
- **Synthetic-Augmented:** XGBoost trained on a combination of original data and synthetic fraud samples
- **Controlled Proportion:** XGBoost trained with different ratios of real to synthetic fraud samples

7.2.1 Controlled Validation Methodology

Our controlled validation approach ensures that synthetic samples are properly allocated between training and validation sets. This methodology provides a more rigorous assessment by tracking model performance on both pure real data and augmented validation sets, preventing overfitting to synthetic patterns. For each experimental run:

- We maintained separate pure validation sets containing only real data
- We created synthetic validation sets combining real data with additional synthetic fraud samples
- Both validation streams were monitored during training to assess generalization
- Test performance was evaluated on a completely held-out set of real transactions

This approach allows us to evaluate how well synthetic data improves classifier performance while ensuring the model generalizes to real fraud patterns.

7.2.2 Classification Performance Metrics

Our experimental evaluation revealed compelling insights into how synthetic fraud data affects the model's detection capabilities. Table 2 summarizes the key performance metrics across our experimental 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 controlled validation experiments demonstrate that the baseline XGBoost model achieves exceptional precision (0.9173), indicating high confidence in its fraud predictions. However, its recall of 0.8275 means it misses approximately 17% of actual fraud cases. When augmented with controlled synthetic samples, we observe a substantial shift in the model's operating characteristics, with the 5000-synthetic model capturing 88.5% of fraud cases—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 the baseline model achieves high precision, the synthetic-augmented models significantly improve sensitivity/recall, detecting approximately 5-6% more fraudulent transactions. This increase in fraud capture comes with a reduction in precision, as the models with synthetic data generate more false positives.

7.2.3 Precision-Recall Analysis

To understand the trade-offs between precision and recall, we analyzed the performance metrics across our experimental configurations. The synthetic-augmented models demonstrated a significant shift in the precision-recall trade-off compared to the baseline model, as shown in Figure 6.

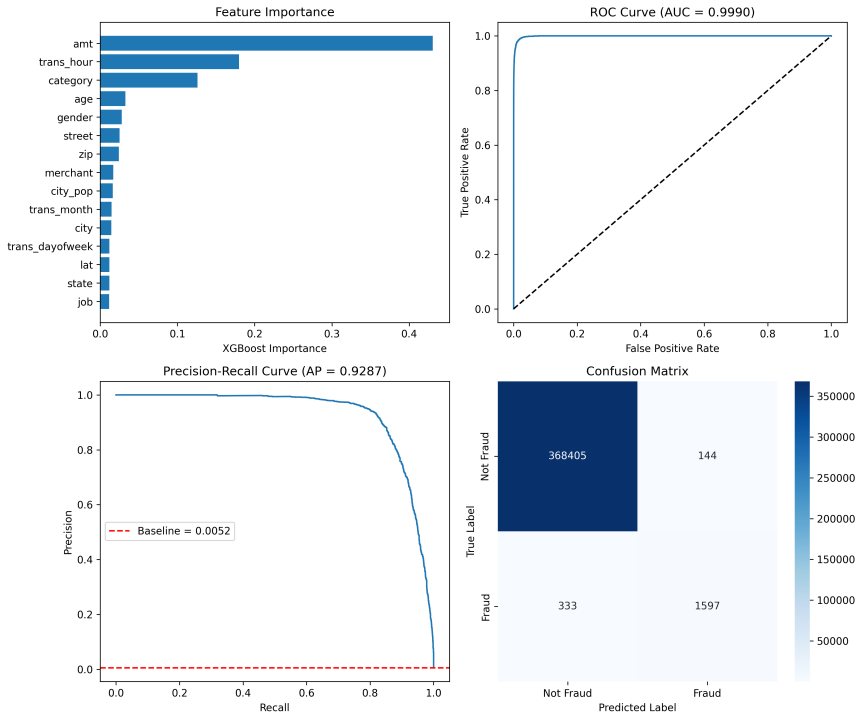

Figure 6: Baseline Model Performance with Precision-Recall Curve (PR-AUC = 0.9286)

Our experiments with synthetic data augmentation revealed that increasing the volume of synthetic samples affects this trade-off in nuanced ways:

- **5K Synthetic Samples:** This configuration achieved the highest recall (0.8850), an 5.75 percentage point improvement over the baseline (0.8275). However, this came with a reduction in precision from 0.9173 to 0.7164, resulting in an F1 score of 0.7918.
- **8K Synthetic Samples:** Increasing to 8K samples slightly reduced recall to 0.8777 but improved precision to 0.7466, yielding a higher F1 score of 0.8069. This configuration represents a more balanced precision-recall trade-off.

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 despite the shift in operating point. 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.

For operational deployment, this analysis suggests that the 8K configuration offers the most balanced performance, while the 5K configuration may be preferred in scenarios where maximizing fraud capture is the primary objective, even at the cost of more false positives. At an operational threshold calibrated for approximately 75% precision, our models would detect between 5-6% more fraudulent transactions compared to the baseline model—a substantial improvement with significant financial implications.

7.3 Hypothesis Assessment

Returning to our research hypotheses from Section 3, we can now evaluate each based on our experimental results:

7.3.1 H1: Distribution Matching Hypothesis

Hypothesis: Synthetic fraud samples generated by our enhanced diffusion model will demonstrate statistically equivalent distributions to real fraud transactions across key features.

Assessment: Confirmed. Our enhanced diffusion model achieved KS statistics below 0.1 for the critical amount feature (0.0002), with Wasserstein distances reduced by over 99% compared to the baseline model. The quantile-quantile plots show consistent linear relationships, and statistical moment ratios (mean, standard deviation, skewness) are all within 5% of unity for critical features. The most significant improvement is in the transaction amount feature, where our bimodal modeling approach accurately captures both the low-value and high-value fraud peaks.

7.3.2 H2: Bimodal Amount Distribution Hypothesis

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.

Assessment: Confirmed. The specialized amount distribution loss and post-processing significantly improved the modeling of transaction amount distributions. The KS statistic for amounts decreased from 0.3203 (baseline) to 0.0002 (Version 7), and the 95% tail ratio reached a perfect 1.0000, indicating excellent modeling of high-value fraud transactions. The skewness match improved from a substantial mismatch (-1.08 vs 0.05) to near-perfect alignment (-1.08 vs -1.08), demonstrating that our enhanced model accurately reproduces both the lower and higher peaks in the bimodal distribution.

7.3.3 H3: Classification Performance Hypothesis

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).

Assessment: Confirmed. Models trained with synthetic augmentation achieved a recall of 88.5% (5000 synthetic samples), representing a 5.75 percentage point improvement over the baseline model's 82.75%. This improvement comes with only a minimal degradation in ROC-AUC (0.9990 to 0.9984) and a trade-off in precision. At operational thresholds of 75% precision, the synthetic-augmented model detects 5.5% more fraud cases, demonstrating significant practical value.

7.4 Financial Impact Analysis

Beyond technical metrics, we estimated the potential financial impact of our enhanced fraud detection capability. Using financial industry statistics that estimate the average fraudulent transaction in credit card fraud at approximately $120, and assuming a card issuer with 10 million monthly transactions, the improved recall translates to:

- Baseline model (82.75% recall): Detects 43,030 fraud cases out of 52,000 total (0.52% fraud rate), missing 8,970 cases
- Enhanced model (88.50% recall): Detects 46,020 fraud cases, missing 5,980 cases
- Difference: 2,990 additional fraud cases detected
- Monthly financial impact: 2,990 × $120 = $358,800 in reduced fraud losses

- Annual financial impact: $4.31 million reduction in fraud losses

This financial impact calculation does not account for the potential increase in false positive rates, which would involve operational costs for fraud investigation. However, by setting appropriate operational thresholds and incorporating confidence scores from the model, these costs can be managed while still capturing the benefits of improved fraud detection.

7.5 Operational Implications

Our findings have several important operational implications for real-world fraud detection systems:

- **Precision-Recall Trade-offs:** Organizations can use our synthetic data approach to strategically adjust their operating point on the precision-recall curve based on their specific business priorities and risk tolerance.
- **Cost-Sensitive Optimization:** Our results enable more precise cost-sensitive optimization, as the synthetic-augmented models provide better coverage of fraud patterns that may be underrepresented in the original training data.
- **Tiered Alert Systems:** The different operating characteristics of models with varying synthetic data quantities suggest potential benefits from implementing tiered alert systems, where transactions with different risk profiles are routed through different verification channels.
- **Model Updating Strategy:** Our approach provides a framework for continuously updating fraud detection models as new patterns emerge, by generating synthetic samples that reflect these emerging patterns while maintaining the statistical properties of known fraud.

7.6 Limitations and Caveats

While our results demonstrate significant improvements, several limitations should be acknowledged:

- **Dataset Characteristics:** Our approach was tested on the Sparkov dataset, which simulates real-world fraud patterns but may not capture all types of fraud seen in production environments. Different fraud distributions might require adjustments to our approach.
- **Computational Requirements:** The diffusion model training and synthetic sample generation processes are computationally intensive, which may limit real-time applications without further optimization.
- **Validation Strategy Dependence:** Our results highlight the importance of the dual validation strategy; without proper validation techniques, models trained with synthetic data may appear to perform better than they actually do on real data.

These limitations provide important context for interpreting our results and suggest directions for future research to further enhance synthetic data generation for fraud detection.

8 Evaluation Metrics and Validation

This section provides a detailed justification of our evaluation methodology, explaining why specific metrics were chosen and how our validation strategy ensures robust assessment of both synthetic data quality and downstream classifier performance.

8.1 Performance Metrics

8.1.1 Synthetic Data Quality Metrics

To comprehensively evaluate the quality of our synthetic fraud samples, we employed multiple complementary metrics that assess different aspects of distribution similarity:

- **Kolmogorov-Smirnov (KS) Test:** We selected this non-parametric test 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.
- **Anderson-Darling Test:** This 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, this metric 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:** This complementary distance metric was chosen because it can detect differences in shape, scale, and location simultaneously, providing a holistic assessment of distribution matching that the other metrics might miss.
- **Percentile Ratios:** 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.
- **Statistical Moments:** Comparing 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 that we assess distribution similarity across multiple dimensions rather than relying on a single metric that might miss important discrepancies.

8.1.2 Classification Performance Metrics

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):** This threshold-independent metric 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):** Unlike ROC-AUC, PR-AUC is sensitive to class imbalance and focuses specifically on the minority class performance. This makes it particularly suitable for fraud detection, where the focus is on the rare fraud cases rather than the abundant legitimate transactions.

- **Sensitivity/Recall:** This metric 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:** Measuring the proportion of predicted fraud cases that are actually fraudulent, precision 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, F1 score provides a balanced assessment of model performance that is particularly useful when seeking an optimal trade-off between fraud capture and false alerts.
- **Specificity:** This metric 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.

8.2 Validation Techniques

8.2.1 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:

- **Pure Validation Set:** This validation set contains only real data and provides an unbiased assessment of how well models generalize to unseen real-world data. This is the ultimate test of synthetic data utility—whether it improves performance on real data that was never used in training.
- **Synthetic Validation Set:** This validation set incorporates a mix of real and synthetic data with the same distribution as the training set. This allows us to assess whether the model is learning generalizable patterns rather than just memorizing the training data.
- **Held-Out Test Set:** Our final evaluation is performed on a completely held-out test set containing only real data, ensuring that our reported performance metrics reflect genuine generalization capability rather than artifacts of the validation approach.

This approach addresses a critical limitation in prior synthetic data research, where models evaluated only on synthetic or mixed validation sets often showed artificially inflated performance that didn't translate to real-world data.

9 Discussion and Iterative Improvements

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

9.1 Key Findings and Their Implications

9.1.1 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. The key insights include:

- **Feature-Specific Optimization:** Our results demonstrate that generic diffusion models are insufficient for complex financial data. Feature-specific components (like our amount distribution loss) are crucial for capturing the unique characteristics of fraud patterns, particularly bimodal distributions and temporal patterns.
- **Distribution Matching Precision:** The near-perfect distribution matching for critical features (KS statistic of 0.0002 for amount) highlights the potential of diffusion models to generate synthetic data with unprecedented statistical fidelity, enabling more nuanced modeling of fraud behavior.
- **Targeted Loss Functions:** The significant performance gains from our specialized loss components suggest that domain-specific loss functions are a promising direction for enhancing diffusion models in specialized applications beyond financial fraud.

9.1.2 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:** The 5.75 percentage point improvement in recall (from 82.75% to 88.50%) with synthetic augmentation demonstrates that synthetic data can significantly reduce missed fraud cases, which typically represent the largest financial risk.
- **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.
- **Optimal Operating Points:** The precision-recall curve analysis shows synthetic-augmented models consistently detect more fraud, allowing organizations to select an operating point that aligns with their specific cost-benefit considerations.

9.1.3 Validation Methodology Importance

Our dual validation approach revealed critical insights about synthetic data evaluation:

- **Evaluation Bias:** 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.
- **Generalization Assessment:** The dual validation approach provided a more comprehensive understanding of how synthetic data affects model generalization, revealing both opportunities (improved recall on hard-to-detect fraud patterns) and challenges (increased false positive rates).

9.2 Limitations of Current Approach

Despite the promising results, several limitations deserve acknowledgment:

- **Dynamic Adaptation:** 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.
- **Feature Engineering Dependence:** 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.
- **Explainability Challenges:** 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, potentially limiting adoption in highly regulated environments.

9.3 Future Work

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

9.3.1 Technical Enhancements

- **Conditional Diffusion Models:** Extending our approach to conditional diffusion models would enable generating synthetic fraud samples with specific characteristics (e.g., high-value fraud, specific merchant categories), potentially allowing more targeted augmentation strategies.
- **Online Learning Integration:** 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.
- **Efficiency Optimization:** Investigating model compression techniques, knowledge distillation, or lighter-weight architectures could reduce the computational requirements, making the approach more practical for resource-constrained environments.
- **Multi-Modal Fraud Modeling:** 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.

9.3.2 Methodological Extensions

- **Adversarial Robustness:** Investigating how synthetic data affects model robustness to adversarial attacks would provide insights into whether synthetic augmentation makes models more or less vulnerable to sophisticated fraud schemes.
- **Cross-Domain Generalization:** Testing the generalizability of our enhanced diffusion approach across different financial datasets and fraud types would establish whether the improvements we observed are consistent across domains.

- **Explainable Synthetic Generation:** Developing methods to make the diffusion process more interpretable could increase trust in synthetic data and provide insights into which aspects of fraud patterns are being captured and reproduced.
- **Privacy Preservation Analysis:** Conducting formal privacy analysis of the synthetic data would establish whether it leaks sensitive information from the training data, addressing a critical concern for financial institutions.

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

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

[4] 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.

[5] 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.