FraudDiffuse: Diffusion-aided Synthetic Fraud Augmentation for Improved Fraud Detection

Ruma Roy
Mastercard
India
rumaroy8008@gmail.com

Darshika Tiwari
Mastercard
India
darshikatiwari29@gmail.com

Anubha Pandey
Mastercard
India
anubhap93@gmail.com

Abstract

Payment fraud poses a severe financial threat, with staggering global losses. Rapidly evolving fraudulent patterns challenge machine learning models, compounded by highly imbalanced training datasets with few fraud samples. Consequently, learning fraud pattern's distribution and generating synthetic fraudulent transactions is imperative. Traditional oversampling techniques like SMOTE (Synthetic Minority Over-sampling Technique) are limited by linear interpolation, failing to capture complex data manifolds. While deep generative models like GANs (Generative Adversarial Networks) have been explored, they suffer from training instability and mode collapse. Recently, denoising diffusion models have emerged as a leading generative modeling paradigm, offering stable training and the ability to learn underlying data manifolds comprehensively. This paper extends existing diffusion techniques for generating synthetic fraud patterns by utilizing the distribution of non-fraudulent samples as the prior, ensuring the model can learn intricate distributions and generate samples close to the class decision boundary. This helps in capturing the subtle nuances that distinguish fraudulent from non-fraudulent instances. Furthermore, we integrate a contrastive learning loss function, which promotes high similarity between synthetic and real fraud samples. This innovative approach not only enriches the training data with realistic fraud patterns but also strengthens the model's ability to distinguish between fraudulent and non-fraudulent transactions. The effectiveness of our proposed algorithm is validated through extensive experiments, demonstrating its superiority over traditional oversampling methods such as SMOTE, GANs, and vanilla diffusion models.

ACM Reference Format:
Ruma Roy, Darshika Tiwari, and Anubha Pandey. 2024. FraudDiffuse: Diffusion-aided Synthetic Fraud Augmentation for Improved Fraud Detection. In *5th ACM International Conference on AI in Finance (ICAIF '24), November 14–17, 2024, Brooklyn, NY, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3677052.3698658

1 Introduction

The fintech industry is experiencing rapid evolution propelled by continual advancements in payment technology. This evolution has fundamentally reshaped how consumers engage in transactions, offering seamless experiences across a multitude of platforms and payment methods. However, the swift progress of fintech has also attracted the attention of fraudsters seeking to exploit vulnerabilities within the system, resulting in significant financial losses amounting to tens of billions of dollars annually on a global scale[1] [2] [3] [4]. Consequently, combating this pervasive threat has emerged as a paramount concern for industries worldwide to preemptively flag suspicious transactions to mitigate their impact, hence safeguarding financial systems and protecting stakeholder's interests [29, 31, 34].

The ever-increasing financial losses due to payment fraud necessitate a proactive approach. Fraudulent patterns are constantly evolving, posing a challenge to traditional machine learning models. This difficulty is further compounded by highly imbalanced training datasets with a scarcity of fraud samples, resulting in machine learning models becoming biased towards the majority class and underfitting the minority class. Researchers have explored oversampling approaches to address these limitations and achieve better generalizability [4, 10]. These techniques aim to increase the representation of fraudulent transactions in the training data, allowing models to learn a more comprehensive picture of fraud patterns. Traditional oversampling techniques like SMOTE can mitigate class imbalance, they struggle to generate data in complex, real-world scenarios due to their reliance on linear interpolation [3, 10]. To address this, learning the distribution of fraud patterns and generating synthetic fraudulent transactions becomes crucial. Deep generative models like GANs offer a more sophisticated approach, capable of creating entirely new, realistic-looking fraudulent transactions [2, 7, 22, 25] . However, GANs suffer from training instability and mode collapse, limiting their practical application in fraud detection [21, 22].

Denoising diffusion models are currently becoming the leading paradigm of generative modeling for many important data modalities [12, 27]. Being the most prevalent in the computer vision community, diffusion models have also recently gained some attention in other domains, including speech [17], NLP [19, 33], and tabular datasets [16, 24, 28, 35]. Unlike traditional generative models that directly learn to create new data, diffusion models begin with a noisy version of the real data and progressively denoise it, learning the underlying patterns and relationships that define the data. By training on real fraudulent transactions and iteratively denoising them, diffusion models can learn to capture the intricate details and variations that characterize fraudulent behavior. This

ICAIF '24, November 14–17, 2024, Brooklyn, NY, USA

ACM ISBN 979-8-4007-1081-0/24/11
https://doi.org/10.1145/3677052.3698658

[1] https://tinyurl.com/2b74jf72
[2] https://tinyurl.com/hn46bc9e
[3] https://tinyurl.com/3mpmzxdv
[4] https://tinyurl.com/2abz2rw9

capability can be particularly advantageous for fraud detection as it allows the model to not only identify existing fraudulent patterns but also generalize to novel fraudulent behaviors not explicitly encountered in the training data. Furthermore, the well-defined training objective of diffusion models makes them less prone to instabilities during training, a common hurdle with GANs. Their ability to learn the underlying structure of data through a denoising process makes them well-suited for addressing the complexities and evolving nature of financial fraud.

This work introduces FraudDiffuse, a novel approach for generating synthetic fraud patterns leveraging diffusion models. FraudDiffuse leverages the distribution of legitimate transactions as a prior, enabling it to capture intricate fraudulent behaviors residing near the decision boundary between legitimate and fraudulent transactions. This focus on the decision boundary facilitates the model's ability to learn subtle nuances that distinguish fraudulent transactions. Furthermore, a contrastive learning loss function is incorporated, promoting high similarity between synthetic and real fraud samples. This approach not only enriches the training data with realistic fraud scenarios but also strengthens the model's ability to differentiate between legitimate and fraudulent transactions. The key highlights of the proposed FraudDiffuse model are as follows:

- To the best of our knowledge, this work is the first to explore the application of diffusion models in generating synthetic fraud patterns for fraud detection.
- By utilizing the distribution of legitimate transactions as a prior, FraudDiffuse focuses on learning the intricate patterns of fraud residing near the decision boundary.
- The integration of a contrastive learning loss function strengthens the model's ability to discriminate between legitimate and fraudulent transactions.
- Extensive experiments validate the effectiveness of FraudDiffuse, demonstrating its superiority over methods like SMOTE, GANs, and vanilla diffusion models.

2 Related Work

Several sampling methods have been introduced to handle imbalanced data. These can be classified into two categories: Undersampling and Oversampling. Undersampling techniques removes samples from the majority class but leads to data loss. Oversampling, particularly the SMOTE [3] algorithm, is more common but limited by its inability to generate truly new data samples. In the domain of deep generative models, several oversampling techniques such as [2, 7, 22], have used Generative Adversarial Networks (GANs) for synthetic data generation but face issues like training instability and mode collapse. To overcome this issue, many works have proposed modifications to the vanilla setup [2, 21], GAN like WGAN [1], Least Square GAN [10], RelaxedWGAN [9] have shown better performance results.

Recently, Diffusion models [26], including Denoising Diffusion Probabilistic Models (DDPM) [12] and Denoising Diffusion Implicit Models (DDIM) [27], have gained attention for their success in generative tasks and studies such as [6, 20] have highlighted the potential superiority of diffusion models in capturing complex data distributions, making them a promising alternative for oversampling in imbalanced data scenarios. Diffusion models have also been applied to tabular data, with examples like TABDDPM [16] and FinDiff [24] focusing on different data preprocessing techniques. Much of the initial work in this domain is constrained to raw tabular data, whereas our work is focused on transformed data or, in other words, the embedding space.

Contrastive learning [11, 14], known for clustering similar samples and separating dissimilar ones, has been effectively applied in diffusion models, such as in [17], which uses separate models for continuous and discrete variables. Our approach differs by using a single diffusion model for fraud oversampling, incorporating a contrastive learning loss function to enhance similarity between synthetic and real fraud samples. This method leverages class labels to guide the generation of data that better aligns with the true fraud distribution.

Many studies use anomaly detection to address data imbalance, often combined with data augmentation in diffusion models [13, 18, 30]. However, this paper introduces a novel method specifically for payment fraud detection, distinguishing it from general anomaly detection, which can result in high false positives by flagging legitimate but unusual transactions. The proposed approach learns the distribution of both legitimate and fraudulent transactions, improving the model's ability to differentiate between them.

3 Methodology

Fraud detection is critical for the fintech industry, but evolving fraud patterns challenge machine learning models. Highly imbalanced datasets further exacerbate this issue. Consequently, learning fraud pattern's distribution and generating synthetic fraudulent transactions is imperative. This paper explores the denoising diffusion model to effectively generate synthetic fraud patterns, addressing limitations of existing methods like GANs and SMOTE. A detailed description of the proposed FraudDiffuse model, including each novel component, is discussed in this section.

3.1 Problem Setup

We consider a transaction training dataset, denoted as $D = \{x_i, y_i\}_{i=1}^{N}$, containing N samples. Each sample i comprises feature vector x_i representing the transaction data and a corresponding class label y_i indicating legitimate ($y_i = 0$) or fraudulent ($y_i = 1$) transactions. We can then extract the features of existing fraud transactions within D to form a subset $x_f = \{x_i \in D | y_i = 1\}$. Similarly, a subset of legitimate transactions can be represented by $x_{nf} = \{x_i \in D | y_i = 0\}$. The objective is to generate a set of synthetic fraud samples, denoted as $\hat{x}_f$, to augment the representation of the fraudulent transactions (minority class) and improve the overall performance of fraud detection models.

3.2 Gaussian Diffusion as a Foundational Framework

Diffusion models are a type of generative model that leverages stochastic processes with latent variables represented by Markov chains. They operate by progressively corrupting a dataset with noise during a forward process and then learning to recover the original data through a reverse process using a neural network.

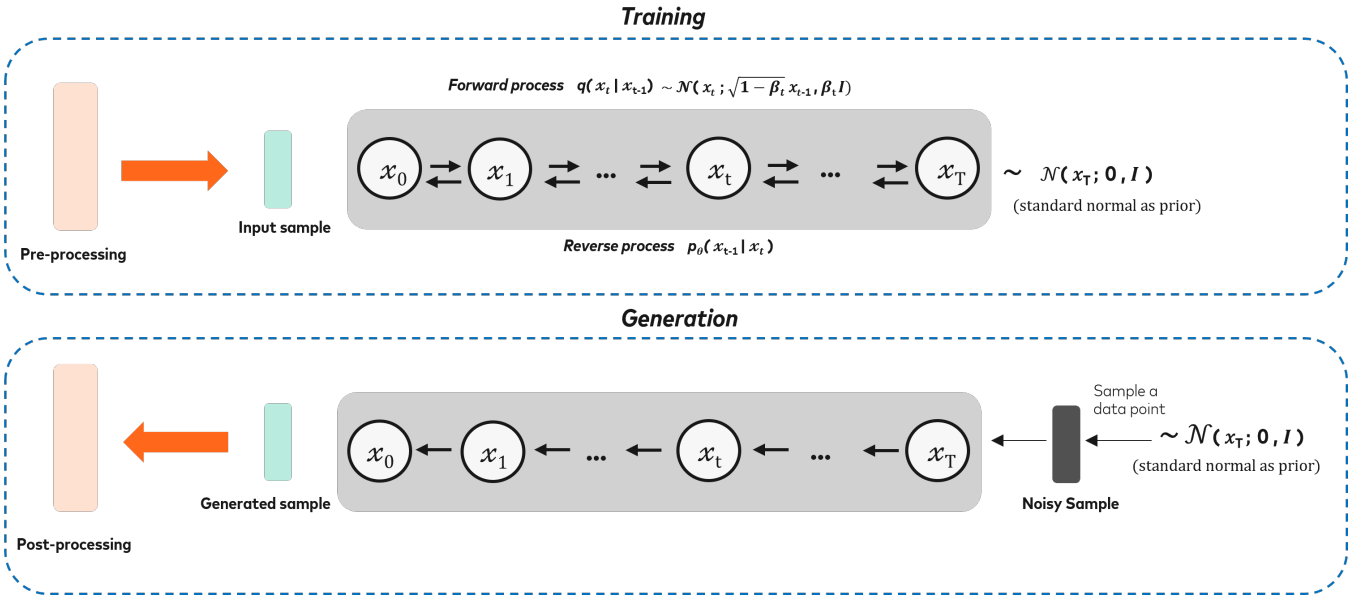


Figure 1: Depiction of the Training and Generation Stages in Vanilla Diffusion Models. The training pipeline encompasses a forward process and a reverse process, employing a standard normal distribution as the prior. The generation stage solely relies on the reverse process acquired during training.

This allows the model to effectively learn the underlying data distribution, as illustrated in Figure 1. Key aspects of diffusion models that influence their behavior include:

- **Training Steps (T_{train})** determines the number of steps used to corrupt the data in the forward process.
- **Generation Steps (T_{Gen})** specifies the number of steps used in the reverse process to generate new data points.
- **Timestep Scheduler** function defines the schedule for introducing noise at each step in the forward process. Common choices includes linear and quadratic schedulers.
- **Noise Schedule (β_t)** defines the level of noise added at each step during the forward process.
- **Multi-layer perceptron (MLP) Architecture** used in the reverse process to predict the noise added at each step.

This work focuses on Gaussian diffusion models, suitable for continuous data represented by vectors in n-dimensional space ($x_t \in \mathbb{R}^n$). Hence, for categorical features, we utilize label encoding within the network's embedding layer that learns weights to produce vector representations (embeddings) for these features. The embeddings are updated during training like any other model parameter and used in the forward pass, with refinements based on gradients in the backward pass. These embeddings of categorical features are the concatenated with the numerical features to achieve the desired representation.

Forward Process: The forward process gradually corrupts the initial data point ($x_0 \sim q(x_0)$) by adding noise sampled from a standard normal distribution. The amount of noise added at each step (t) is controlled by the noise level (β_t). This Markov process is described mathematically in the Eq. 1 and 2.

$$q(x_t|x_{t-1}) \sim \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right) \tag{1}$$

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), q(x_T) \sim \mathcal{N}(x_T; 0, I) \tag{2}$$

Reverse Process: This process can be viewed as a Markov chain moving backward in time that aims to progressively de-noise the corrupted data ($x_T \sim q(x_T)$) by estimating and removing the added noise at each step characterized by Eq. 3.

$$p_\theta(x_{t-1}|x_t) \sim \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{3}$$

Estimation of parameters for $p_\theta(x_{t-1}|x_t)$ has been formulated by [16] where Σ_θ is a constant diagonal matrix and μ_θ is describe in the Eq. 4.

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)) \tag{4}$$

Where, $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i \le t} \alpha_i$ and $\epsilon_\theta(x_t, t)$ is the predicted error component for given latent sample x_t.

This architecture is trained by minimising loss function which is mean squared-error between added error ϵ and predicted errors ϵ_θ. The loss function used for training is described in Eq. 5

$$\mathrm{L}_{\text{norm}} = \mathbb{E}_{x_0,e,t}\left[\|\epsilon - \epsilon_\theta\|_2^2\right] \tag{5}$$

3.3 Proposed FraudDiffuse Model: A Generative Model for Fraud Augmentation

This work leverages the Gaussian diffusion model framework to develop FraudDiffuse, a novel data augmentation technique for

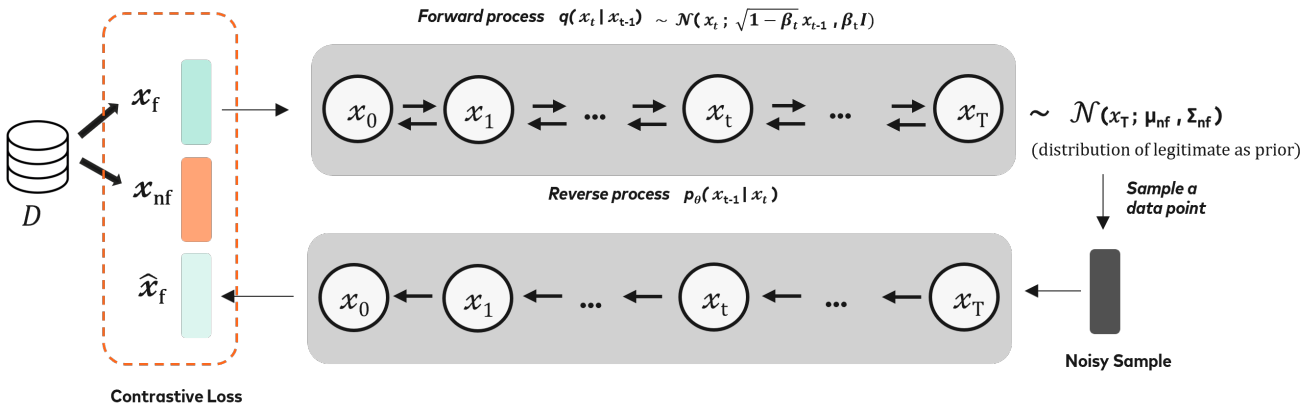


Figure 2: Architecture of FraudDiffuse: The distribution of legitimate transactions is utilized as a prior by FraudDiffuse. The influence of the prior distribution is regulated through a contrastive loss function that incorporates negative sampling.

fraud detection. FraudDiffuse generates synthetic fraudulent transaction samples ($\hat{x}_f$) to address class imbalance and evolving fraud patterns. These synthetic samples are designed to closely resemble genuine fraudulent transactions and are strategically incorporated into training datasets to enhance the representation of the minority fraud class. The optimal quantity of synthetic samples is determined by analyzing the impact of varying augmentation ratios on the validation set F1 score (Figure 3). By increasing the fraud event rate through the introduction of $\hat{x}_f$, FraudDiffuse aims to improve the overall performance of fraud detection models.

3.3.1 ***Adaptive Prior for Learning Boundary Frauds:*** Standard diffusion models utilize a Gaussian prior for noise addition. However, this may not capture complex fraud features. We address this by introducing a data-dependent non-fraud distribution as an adaptive prior. The non-fraud prior parameters $\mathcal{N}(\mu_{nf}, \Sigma_{nf})$ are estimated using non-fraud training data statistics (mean and diagonal covariance). This strategy, inspired by [8], forces the model to learn features near non-fraud samples but still categorized as fraud due to indirect complex characteristics. The estimation of $\mu_\theta(x_t, t)$ and $\Sigma_\theta(x_t, t)$ remains the same as TabDDPM, except the forward process adds noise based on the non-fraud prior. And reverse process samples x_T from $\mathcal{N}(\mu_{nf}, \Sigma_{nf})$ instead of $\mathcal{N}(0, I)$. This approach effectively encourages the model to learn features crucial for differentiating complex frauds near the decision boundary.

3.3.2 ***Improved Diffusion Model Training with Probability-Based Loss:*** Diffusion models learn complex feature spaces by estimating noise added to latent variables. Standard training utilizes distance-based loss functions like mean squared error L_{norm} described in Eq. 5. We propose a novel probability-based loss function L_{prior} to enhance training. Since the added noise follows an adaptive prior distribution $\mathcal{N}(\mu_{nf}, \Sigma_{nf})$, predicted errors should also adhere to this distribution.

L_{prior} calculates the likelihood of predicted errors originating from the prior distribution, measuring the area under the probability density function (PDF). Specifically, we employ a two-tailed z-score probability, aiming to maximize the area under the standard normal curve beyond the absolute value of the predicted error's z-score. This translates to minimizing the complement of this area.

$$\mathrm{L_{prior}} = 2 * P\,(Z \leq |\text{z-score}|) = 1 - 2 * P\,(Z \geq |\text{z-score}|) \quad (6)$$

The mathematical formulation of L_{prior} is shown in Eq. 6, where z-score is $\frac{\epsilon_{\theta_j} - \mu_j}{\sigma_j}$ and ϵ_{θ_j} is the predicted error for j-th feature. L_{prior} is calculated for each latent feature and summed across all features. This combined loss function ($w_1 * L_{prior} + L_{norm}$) encourages the model to learn complex features by improving noise estimation accuracy.

3.3.3 ***Realistic Fraud Generation with Contrastive Regularization:*** While the proposed non-fraud adaptive prior and probability loss function improve fraud generation, they can lead to overfitting. To address this, we introduce a contrastive learning mechanism using triplet loss to ensure generated data resides closer to fraud and further from non-fraud samples. This allows the model to learn smoother transitions between the two classes, capturing complex fraud patterns.

We integrate contrastive learning by incorporating triplet loss into the training pipeline. This is a novel approach, as we combine the generation process with training for the first time. During backpropagation, we generate a fraud sample by denoising estimated errors from the non-fraud prior $\mathbf{x}_T \sim \mathcal{N}(\mu_{nf}, \Sigma_{nf})$. We consider this generated sample as the "anchor," real fraud samples as "positives," and randomly chosen non-fraud samples as "negatives." And minimize the triplet loss, encouraging the model to bring the generated anchor closer to real positives while pushing it away from negatives. The triplet loss is described in the Eq. 7, where $d(.)$ denotes a distance metric.

$$\mathrm{L_{triplet}} = \max(0, d(\hat{x}_f, x_f) - d(\hat{x}_f, x_{nf})) \quad (7)$$

This additional term ($w_2 * L_{triplet}$) is minimized along with L_{norm} and L_{prior}, promoting robust feature learning and preventing overfitting. The overall loss function used to train FraudDiffuse is given by Eq. 8, where w_1 and w_2 are weights corresponding to the individual loss terms.

$$\mathrm{L_{fraudDiffuse}} = L_{norm} + w_1 * L_{prior} + w_2 * L_{triplet} \quad (8)$$

4 Datasets and Experimental Details

The proposed FraudDiffuse model has been evaluated on two datasets and compared with state-of-the-art techniques. Details regarding the dataset protocols are as follows:

European Credit Card Default Dataset[5]: This dataset contains transactions conducted by cardholders over two days. It comprises a total of 30 features consisting of *Amount, Time, 28 PCA-based features*, and a class label indicating 1 for fraudulent transactions and 0 for legitimate ones. All features are numerical, and a log transformation has been applied to the *Amount* feature to achieve a distribution closer to standard normal. This dataset does not contain any samples with missing values. It is highly imbalanced, with a fraudulent event rate of 0.172% out of a total of 284,807 transactions.

IEEE-CIS Fraud Detection dataset[6]: The dataset contains Vesta's real-world e-commerce transactions. It comprises a total of 433 features consisting of 4 categorical and 429 numerical features, with 339 as PCA-based features, along with an *isFraud* class label indicating 1 for fraudulent transactions and 0 for legitimate ones. We dropped features with more than 20% missing values and applied median imputation for numerical features and most frequent value imputation for categorical ones. The final features used in experiments consist of 175 numerical and 4 categorical features. This dataset is highly imbalanced, with a fraudulent event rate of 3.499% out of a total of 590,540 transactions.

4.1 Evaluation metrics

Classification model performance is evaluated on the test set by calculating precision, recall, and F1-score (harmonic mean of precision and recall). Accuracy is not considered a good metric for imbalanced datasets. A fraud detection classifier should accurately predict fraud while minimizing false positives to reduce financial losses. We aim to maximize the recall rate without sacrificing precision. Therefore, we focus on the F1-score as the primary evaluation metric. We performed 10 independent experiments for each oversampling technique and presented the classification metrics of the best-performing run.

4.2 Implementation Details

We utilized a two-step strategy for fraud detection: Initially, the proposed FraudDiffuse was employed to generate synthetic fraud samples. Subsequently, an XGBoost classifier [5] was trained on this balanced dataset to detect fraudulent transactions. The performance of various machine learning classifiers, including Logistic Regression, Decision Tree, Random Forest, and XGBoost was compared wherein XGBoost emerged as the best performer based on our evaluation criteria. Additionally, XGBoost is widely used for tabular datasets in real-world industrial applications, justifying our choice for performance evaluation. We divided the dataset into a 65% training set, a 15% validation set, and a 20% test set while maintaining the event rate in both subsets. Model development was conducted using the training set, with evaluation metrics reported on the test set.

[5]https://www.kaggle.com/mlg-ulb/creditcardfraud
[6]https://www.kaggle.com/c/ieee-fraud-detection

The FraudDiffuse model was implemented using the PyTorch framework [23]. Categorical features were transformed into numerical representations using 2-dimensional embeddings before feeding into the model. The MLP used in the architecture consisted of 3 layers of 256 neurons each. The learning rate was set to 0.001. The model was trained for 150 epochs on the training set, optimizing a loss function described in Eq. 8. In order to determine the value of w1 and w2, hyperparameter optimization was done where each weight was explored within the 0-1 range, and the optimal values were selected based on the validation set loss. The weights for the L_{prior} loss (w_1) were set to 0.2 for both the dataset. The weights for the $L_{triplet}$ loss (w_2) were set to 0.6 for the European credit card default dataset and 0.8 for the IEEE-CIS fraud detection dataset.

The modeling hyperparameters, such as training/generation diffusion steps, noise variance level, and the neural network architecture (number of layers/neurons), were decided using hyperparameter tuning. The values of these hyper-parameters for both datasets are specified in Table 1. Diffusion steps were scheduled linearly, and diffusion time steps were embedded using a sinusoidal function. Noise variance β followed a linear schedule, increasing from 1e-4 to 0.02 during training. Both datasets used a 3-layer feed-forward neural network with Leaky ReLU activation [32] for noise prediction. The training involved 500 epochs with a batch size of 40, utilizing the Adam optimizer [15] and a learning rate of 0.001. Finally, synthetic fraud samples were augmented to their respective datasets to double the fraud event rate.

Table 1: Modeling Hyperparameters of FraudDiffuse

Dataset	Training Diffusion Steps	Generation Diffusion Steps	Neural Network Architecture
European Credit Card Default	800	120	256-256-256
IEEE-CIS Fraud Detection	800	80	512-512-512

The XGBoost model was configured with 1200 boosting rounds and a learning rate of 0.05 for both datasets. However, the tree depth was hyperparameter-tuned based on dataset complexity: 4 for the European credit card default dataset and 8 for the IEEE-CIS fraud detection dataset. To prevent overfitting, the model was trained while monitoring AUC-PR on the validation set. Early stopping was implemented, terminating training if the AUC-PR on the validation set did not improve for 10 consecutive boosting rounds. The classification performance is reported on a default threshold of 0.5 for both datasets.

5 Results and Analysis

Table 2 present the results and analysis of the FraudDiffuse model and comparison with the state-of-the-art data augmentation techniques in fraud detection. The performance of the XGBoost model trained on the actual and augmented datasets is reported on the test set. Detailed analysis is given in the following subsections:

Table 2: The table presents fraud detection performance of FraudDiffuse model against state-of-the-art oversampling techniques. An XGBoost classifier is trained on datasets augmented with various oversampling methods. Performance metrics, including precision, recall, and F1-score are used to evaluate the effectiveness in enhancing fraud detection performance.

	European Credit Card Default Dataset			IEEE-CIS Fraud Detection Dataset		
Augmentation Method	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
None	0.900	0.762	0.824	0.795	0.289	0.424
SMOTE[7]	0.981	0.703	0.819	0.778	0.289	0.422
WGAN[2]	0.880	0.788	0.831	0.802	0.295	0.432
WCGAN[2]	0.880	0.789	0.832	0.80	0.296	0.432
WCGAN-GP + Adversarial loss[22]	0.934	0.803	0.864	0.802	0.302	0.440
WGAN-GP+ Auxiliary Classifier+DRS[21]	0.929	0.790	0.854	0.805	0.297	0.434
Vanilla Diffusion Model[24]	0.921	0.837	0.877	0.825	0.329	0.469
FraudDiffuse	**0.954**	**0.847**	**0.897**	**0.840**	**0.335**	**0.478**

5.1 Comparison with State-of-the-art Algorithms

The effectiveness of FraudDiffuse in fraud detection was evaluated by comparing it to other oversampling techniques. An XGBoost classifier is trained for this purpose on balanced datasets generated using the following methods: (i) SMOTE [7], (ii)WGAN [2], (iii)WCGAN [2], (iv) WGAN-GP+ Auxiliary Classifier+DRS [22], and (v) WCGAN-GP + Adversarial loss [21]. Additionally, FraudDiffuse was compared to a Vanilla Diffusion Model [24]. This comparison specifically assessed the efficacy of FraudDiffuse's components in improving the quality of synthetic fraud generation, ultimately leading to enhanced fraud detection performance. Performance analysis on different datasets is as follows:

(a) Comparison on the European Credit Card Default Dataset: Table 2 summarizes the XGBoost classification performance on the augmented European credit card default dataset. Augmenting with synthetic frauds increased the event rate from 0.17% to 0.34%. SMOTE showed the lowest performance, with F1-score further declining compared to the model trained on non-augmented dataset. WCGAN-GP with Adversarial Loss [21] outperformed other GAN-based techniques, achieving an absolute improvement of 4.9% in F1-Score than classfier trained on actual dataset. The proposed FraudDiffuse model achieved superior performance compared to all other methods. It demonstrated an absolute lift of 2.1% in precision, 5.5% in recall, and 3.8% in F1-score than best performing baseline WCGAN-GP + Adversarial loss [21]. Notably, this improvement surpasses GAN-based approaches despite the low event rate. This suggests that FraudDiffuse requires less training data to learn the feature distribution.

(b) Comparison on the IEEE-CIS Fraud Detection Dataset: Table 2 presents the performance for various oversampling methods on the IEEE-CIS fraud detection dataset. As with the credit card data, data augmentation doubled the event rate to 7%. SMOTE again resulted in performance degradation for all metrics, even compared to the model trained on the non-augmented dataset. Among all the baseline methods, WCGAN-GP with Adversarial Loss [21] achieved the highest absolute improvement of 3.8% in F1-Score as compared to without augmentation. The proposed FraudDiffuse model surpassed all other methods on this dataset as well. It achieved the highest absolute lifts in precision (4.7%), recall (10.9%), and F1-score (8.6%) than WCGAN-GP + Adversarial loss [21]. Notably, the vanilla diffusion model also outperformed SMOTE and GAN-based methods, with a 6.6% F1-score improvement than WCGAN-GP + Adversarial loss [21].

Table 3: The table presents the quality assessment of synthetic frauds generated from FraudDiffuse. Kolmogorov-Smirnov (KS) Distance and Kullback-Leibler (KL) Divergence are used to measure the distance of synthetic fraud samples with real fraud and legitimate samples. It is shown by the results that the synthetic fraud samples are closer to real fraud samples in comparison to the legitimate samples.

Dataset		KS Distance	KL Divergence
European Credit Card Default	**Fraud Samples**	0.142	0.018
	Legitimate Samples	0.517	0.634
IEEE-CIS Fraud Detection	**Fraud Samples**	0.257	3.521
	Legitimate Samples	0.187	7.001

5.2 Assessing the quality of the synthetic fraud samples

This section analyzes the quality of synthetic fraud samples generated by FraudDiffuse, focusing on their similarity to the distribution of real fraud data. Since our data consist of numerical features, we employed Kolmogorov-Smirnov (KS) Distance and Kullback-Leibler (KL) Divergence as the evaluation metrics. Lower values of both metrics with the real fraud samples indicate better quality generation, as they represent smaller discrepancies between the real and synthetic fraud distributions. The results are reported in Table 3.

For the European credit card default dataset, the results suggest strong evidence that synthetic fraud samples are more similar to the real fraud space compared to the legitimate feature space. Absolute KS-Distance and KL-Divergence are both significantly lower (72.5% and 97.2% lower, respectively) when comparing synthetic frauds to

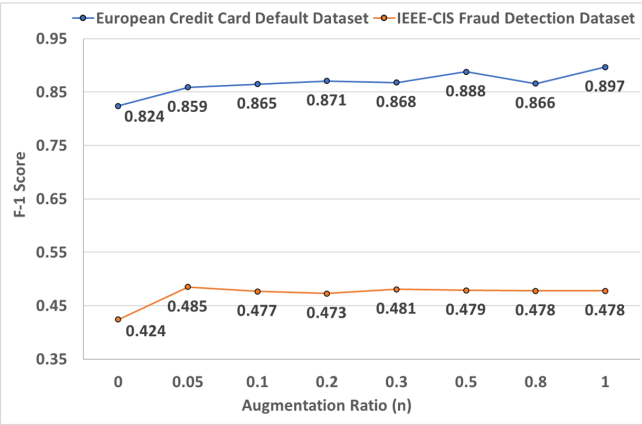
Table 4: Dissection of the FraudDiffuse model's components on a fraud detection task for European credit card default dataset. The table presents the precision, recall, and F1-score over 10 runs for a vanilla diffusion model and the incremental improvements achieved by adding each component.

Model	Precision	Recall	F1-score
Vanilla diffusion model	0.902 ± 0.012	0.823 ± 0.007	0.863 ± 0.005
Vanilla diffusion model + non-fraud prior	0.888 ± 0.004	0.835 ± 0.013	0.860 ± 0.008
Vanilla diffusion model + L_{prior}	0.919 ± 0.006	0.829 ± 0.014	0.871 ± 0.007
Vanilla diffusion model + $L_{triplet}$	0.921 ± 0.012	0.831 ± 0.017	0.873 ± 0.010
Vanilla diffusion model + L_{prior} + $L_{triplet}$	0.929 ± 0.009	0.830 ± 0.011	0.877 ± 0.009
FraudDiffuse	0.938 ± 0.013	0.836 ± 0.011	0.884 ± 0.007

real frauds compared to legitimate transactions. This indicates that FraudDiffuse successfully generated high-quality synthetic frauds for this dataset.

For the IEEE-CIS fraud detection dataset with mixed-type features, KL divergence is 50% lower for real fraud samples compared to legitimate samples. However, KS-Distance shows a higher value for real fraud samples. Despite the higher KS-Distance, the lower KL-Divergence suggests that the generated synthetic frauds are closer to the real fraud feature space. This finding highlights the potential benefits of incorporating contrastive learning in FraudDiffuse, especially for datasets with mixed feature types.

Figure 3: The impact of varying the augmentation ratio on the F1-score of an XGBoost classifier used for fraud detection.



5.3 Impact of Synthetic Fraud Augmentation Ratio

Determining the optimal number of synthetic fraud samples is crucial for balancing performance and computational cost. We investigated the effect of varying augmentation levels on classification performance (F1-score) for both datasets. The fraud event rate was increased by augmenting synthetic fraud samples $\hat{x_f} = n * x_f$, where x_f represents the real fraud samples, and n is the augmentation ratio. Figure 3 illustrates the impact of different augmentation ratios on the F1-score for both datasets. For the European credit card default dataset, the best F1-score was achieved by augmenting with an equal number of synthetic fraud samples, i.e., $n = 1$. This dataset has a very low minority class rate (0.172%), suggesting that even a modest increase in synthetic examples can be beneficial. For IEEE-CIS fraud detection dataset, a significant lift in classification metrics was observed by augmenting with 5% of real fraud samples ($n = 0.05$). Performance improvement plateaued beyond this ratio. This suggests that the model can effectively learn from a smaller number of synthetic samples due to the presence of more balanced classes in this dataset.

6 Ablation Study

This paper introduces FraudDiffuse, a modified diffusion model for generating synthetic minority fraud samples in transactional datasets. To prevent replication of existing fraud samples, we incorporate non-fraudulent transaction information through three key modifications: (1) adaptive prior for boundary fraud learning, (2) probability-based loss for improved training, and (3) contrastive regularization for realistic generation. The individual impact of each modification on model performance is assessed through experiments conducted on the European credit card default dataset. Detailed results are presented in Table 4.

Impact of Adaptive Prior: FraudDiffuse employs a non-fraud distribution as an adaptive prior to capture complex fraud features near the decision boundary. The model samples noise from this prior, forcing it to learn features close to non-fraudulent data points that are nevertheless classified as fraud due to intricate relationships. This approach leads to a 1.5% improvement in recall compared to the vanilla diffusion model.

Impact of Probability-Based Loss: The probability-based loss improves the model's ability to estimate prior errors, which is crucial for reducing false positives in fraud detection. High precision ensures legitimate customers are not inconvenienced by being mistakenly flagged as fraudulent. Our experiments demonstrate that precision improves by 1.8% when probability-based loss is incorporated with the vanilla diffusion model.

Impact of Contrastive Regularization: While the probability-based loss improves error estimation, it can lead to overfitting in predicting errors. To mitigate this and enhance the model's ability to discriminate between legitimate and fraudulent transactions, FraudDiffuse employs contrastive regularization with a triplet loss.

This enforces the generated fraudulent samples to be closer to the legitimate fraud space and further away from the legitimate transaction space. Our experiments demonstrate that incorporating contrastive loss with the vanilla diffusion model yields a 2.1% lift in precision and 1.0% lift in recall.

The combination of non-fraud adaptive prior, probability-based loss and contrastive regularization in the vanilla diffusion model leads to superior classification performance. This combination achieves the highest improvements in precision and recall, resulting in an overall 2.4% F1-score lift compared to the vanilla diffusion model. These refinements make FraudDiffuse particularly effective for highly imbalanced transactional data.

7 Conclusion

We introduce FraudDiffuse, a novel diffusion model for generating synthetic fraud patterns. By leveraging a non-fraud transaction prior, it captures intricate fraudulent behaviors near the decision boundary, crucial for effective detection. Additionally, a contrastive learning loss fosters similarity between synthetic and real fraud, enriching training data and enhancing discrimination ability. Experimental analysis and results validate FraudDiffuse's effectiveness, surpassing baseline methods.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein GAN. arXiv:1701.07875 [stat.ML]

[2] Hung Ba. 2019. Improving Detection of Credit Card Fraudulent Transactions using Generative Adversarial Networks. arXiv:1907.03355 [cs.LG]

[3] Chumphol Bunkhumpornpat, Krung Sinapiromsaran, and Chidchanok Lursinsap. 2011. DBSMOTE: Density-based synthetic minority over-sampling technique. *Applied Intelligence - APIN* 36 (04 2011). https://doi.org/10.1007/s10489-011-0287-y

[4] Fabrizio Carcillo, Yann-Aël Le Borgne, Olivier Caelen, Yacine Kessaci, Frédéric Oblé, and Gianluca Bontempi. 2021. Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences* 557 (2021), 317–331. https://doi.org/10.1016/j.ins.2019.05.042

[5] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM. https://doi.org/10.1145/2939672.2939785

[6] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 8780–8794. https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf

[7] Ugo Fiore, Alfredo De Santis, Francesca Perla, Paolo Zanetti, and Francesco Palmieri. 2019. Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences* 479 (2019), 448–455.

[8] Sang gil Lee, Heeseung Kim, Chaehun Shin, Xu Tan, Chang Liu, Qi Meng, Tao Qin, Wei Chen, Sungroh Yoon, and Tie-Yan Liu. 2022. PriorGrad: Improving Conditional Denoising Diffusion Models with Data-Dependent Adaptive Prior. arXiv:2106.06406 [stat.ML]

[9] Xin Guo, Johnny Hong, Tianyi Lin, and Nan Yang. 2021. Relaxed Wasserstein with Applications to GANs. arXiv:1705.07164 [stat.ML]

[10] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. 2005. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In *Advances in Intelligent Computing*, De-Shuang Huang, Xiao-Ping Zhang, and Guang-Bin Huang (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 878–887.

[11] Chih-Hui Ho and Nuno Nvasconcelos. 2020. Contrastive Learning with Adversarial Examples. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 17081–17093. https://proceedings.neurips.cc/paper_files/paper/2020/file/c68c9c8258ea7d85472dd6fd0015f047-Paper.pdf

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[13] Hadi Hojjati, Thi Kieu Khanh Ho, and Narges Armanfard. 2024. Self-supervised anomaly detection in computer vision and beyond: A survey and outlook. *Neural Networks* 172 (April 2024), 106106. https://doi.org/10.1016/j.neunet.2024.106106

[14] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning.. In *NeurIPS*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). http://dblp.uni-trier.de/db/conf/nips/neurips2020.html#KhoslaTWSTIMLK20

[15] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG]

[16] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. 2023. TabDDPM: Modelling Tabular Data with Diffusion Models. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 17564–17579. https://proceedings.mlr.press/v202/kotelnikov23a.html

[17] Chaejeong Lee, Jayoung Kim, and Noseong Park. 2023. Codi: Co-evolving contrastive diffusion models for mixed-type tabular synthesis. In *International Conference on Machine Learning*. PMLR, 18940–18956.

[18] Victor Livernoche, Vineet Jain, Yashar Hezaveh, and Siamak Ravanbakhsh. 2023. On Diffusion Modeling for Anomaly Detection. arXiv:2305.18593 [cs.LG] https://arxiv.org/abs/2305.18593

[19] Jiamin Luo, Jingjing Wang, and Guodong Zhou. 2024. TopicDiff: A Topic-enriched Diffusion Approach for Multimodal Conversational Emotion Detection. arXiv:2403.04789 [cs.CL]

[20] Soumik Mukhopadhyay, Matthew Gwilliam, Vatsal Agarwal, Namitha Padmanabhan, Archana Swaminathan, Srinidhi Hegde, Tianyi Zhou, and Abhinav Shrivastava. 2023. Diffusion Models Beat GANs on Image Classification. arXiv:2307.08702 [cs.CV]

[21] Anubha Pandey, Alekhya Bhatraju, Shiv Markam, and Deepak Bhatt. 2022. Adversarial Fraud Generation for Improved Detection. In *Proceedings of the Third ACM International Conference on AI in Finance* (New York, NY, USA) *(ICAIF '22)*. Association for Computing Machinery, New York, NY, USA, 123–129. https://doi.org/10.1145/3533271.3561723

[22] Anubha Pandey, Deepak L. Bhatt, and Tanmoy Bhowmik. 2020. Limitations and Applicability of GANs in Banking Domain. In *ADGN@ECAI*. https://api.semanticscholar.org/CorpusID:229357084

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv:1912.01703 [cs.LG]

[24] Timur Sattarov, Marco Schreyer, and Damian Borth. 2023. Findiff: Diffusion models for financial tabular data generation. In *Proceedings of the Fourth ACM International Conference on AI in Finance*. 64–72.

[25] Akhil Sethia, Raj Patel, and Purva Raut. 2018. Data Augmentation using Generative models for Credit Card Fraud Detection. In *2018 4th International Conference on Computing Communication and Automation (ICCCA)*. 1–6. https://doi.org/10.1109/CCAA.2018.8777628

[26] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*. PMLR, 2256–2265.

[27] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2022. Denoising Diffusion Implicit Models. arXiv:2010.02502 [cs.LG]

[28] Namjoon Suh, Xiaofeng Lin, Din-Yin Hsieh, Merhdad Honarkhah, and Guang Cheng. 2023. AutoDiff: combining Auto-encoder and Diffusion model for tabular data synthesizing. arXiv:2310.15479 [stat.ML]

[29] Yuchen Wang, Jinghui Zhang, Zhengjie Huang, Weibin Li, Shikun Feng, Ziheng Ma, Yu Sun, Dianhai Yu, Fang Dong, Jiahui Jin, Beilun Wang, and Junzhou Luo. 2023. Label Information Enhanced Fraud Detection against Low Homophily in Graphs. In *Proceedings of the ACM Web Conference 2023 (WWW '23)*. Association for Computing Machinery, New York, NY, USA, 406–416. https://doi.org/10.1145/3543507.3583373

[30] Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C. Cattin. 2022. Diffusion Models for Medical Anomaly Detection. arXiv:2203.04306 [eess.IV] https://arxiv.org/abs/2203.04306

[31] Sheng Xiang, Mingzhi Zhu, Dawei Cheng, Enxia Li, Ruihui Zhao, Yi Ouyang, Ling Chen, and Yefeng Zheng. 2023. Semi-supervised Credit Card Fraud Detection via Attribute-Driven Graph Representation. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 12 (Jun. 2023), 14557–14565. https://doi.org/10.1609/aaai.v37i12.26702

[32] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. 2015. Empirical Evaluation of Rectified Activations in Convolutional Network. arXiv:1505.00853 [cs.LG]

[33] Yiming Zhang, Avi Schwarzschild, Nicholas Carlini, Zico Kolter, and Daphne Ippolito. 2024. Forcing Diffuse Distributions out of Language Models. arXiv:2404.10859 [cs.CL]

[34] Ya-Lin Zhang, Yi-Xuan Sun, Fangfang Fan, Meng Li, Yeyu Zhao, Wei Wang, Longfei Li, Jun Zhou, and Jinghua Feng. 2023. A Framework for Detecting Frauds from Extremely Few Labels. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining* (<conf-loc>, <city>Singapore</city>, <country>Singapore</country>, </conf-loc>) *(WSDM '23)*. Association for Computing Machinery, New York, NY, USA, 1124–1127. https://doi.org/10.1145/3539597.3573022

[35] Yuan Zhong, Suhan Cui, Jiaqi Wang, Xiaochen Wang, Ziyi Yin, Yaqing Wang, Houping Xiao, Mengdi Huai, Ting Wang, and Fenglong Ma. 2024. Meddiffusion: Boosting health risk prediction via diffusion-based data augmentation. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*. SIAM, 499–507.