Information & Security

Published since 1998

Y. Pushkarenko & V. Zaslavskyi
vol. 55, no. 2 (2024): 185-198
https://doi.org/10.11610/isij.5534

ISSN 0861-5160 (print), ISSN 1314-2119 (online)

Research Article

Synthetic Data Generation for Fraud Detection Using Diffusion Models

Yurii Pushkarenko (✉) ***and Volodymyr Zaslavskyi***

Faculty of Computer Science and Cybernetics, Taras Shevchenko National University of Kyiv, Ukraine, https://knu.ua/en/

ABSTRACT:

Detection of fraudulent transactions in payment and banking systems using credit cards is a significant challenge, primarily due to the limitations in accessing real-world data necessary for training models and developing algorithms to analyze transaction streams for accuracy. Real data related to contractual relationships between financial systems and their clients is confidential, which influences both the formation of the data recorded in transactions and the analysis of transaction flows to identify fraudulent activities.

This paper explores the potential of using diffusion models to generate realistic synthetic transaction data aimed at improving the performance of fraud detection algorithms. Particular emphasis is placed on processing datasets that contain a mix of categorical (textual) and numerical attributes and exhibit a pronounced class imbalance between legitimate and fraudulent transactions.

A comparison is presented between the effectiveness of traditional fraud detection methods on real transaction data and the proposed approach, which actively employs synthetic data generated using diffusion models. The results demonstrate significant improvements in the reliability of models in accurately detecting fraud, highlighting the potential of diffusion models as a powerful tool in the development of more effective fraud detection systems.

ARTICLE INFO:

RECEIVED: 06 SEP 2024
REVISED: 06 OCT 2024
ONLINE: 31 OCT 2024

KEYWORDS:

banking transactions, fraud detection, diffusion models, synthetic dataset generation

✉ E-mail: yurii.pushkarenko@gmail.com

Introduction

Fraud detection remains a pivotal concern in numerous sectors, including finance, healthcare, and online services, and since the popularity of the gambling industry where observed extensive money laundering. As fraudulent activities become increasingly sophisticated, the need for advanced detection techniques has never been greater. Traditional methods often rely on supervised learning algorithms that necessitate extensive labeled data for training. However, these methods face significant challenges, particularly the class imbalance problem,[1] where legitimate transactions far outnumber fraudulent ones.

Generative models have emerged as powerful tools for addressing the limitations of traditional fraud detection systems.[2] Among these, diffusion models have shown exceptional promise in generating high-quality synthetic data. Unlike other generative approaches, diffusion models iteratively refine data through transformations, resulting in realistic and diverse synthetic samples. This characteristic makes them particularly suitable for augmenting imbalanced datasets, thereby enhancing the training process of fraud detection algorithms.

This paper explores the application of diffusion models for synthetic data generation in the context of fraud detection. Our approach leverages these models to create synthetic instances that mimic the statistical properties of genuine fraudulent activities. Integrating this synthetic data with real-world datasets aims to improve the performance and robustness of fraud detection systems.

It conducts comprehensive experiments using benchmark datasets and real-world scenarios to evaluate the efficacy of our proposed method. The results demonstrate substantial improvements in detection accuracy and model resilience, underscoring the potential of diffusion models as a valuable asset in the fight against fraud.

Related Work

Fraud detection has been extensively studied, with numerous approaches proposed to tackle this pervasive issue. Traditional methods, including statistical techniques and rule-based systems, laid the groundwork for more sophisticated machine-learning algorithms that were proposed by Zaslavskyi et al.[3] in their research. These early methods, while effective to some extent, often struggled with scalability and adaptability to evolving fraudulent behaviors.[4]

In recent years, machine learning and artificial intelligence have revolutionized fraud detection. Supervised learning algorithms, such as logistic regression,[5] decision trees, and neural networks, have been widely adopted due to their ability to learn from historical data and make predictions on new, unseen instances. However, these approaches are significantly hindered by the class imbalance problem, where fraudulent instances are rare compared to legitimate ones, leading to biased models that favor the majority class.

To address this issue, researchers have explored various techniques for data augmentation and resampling.[6] Synthetic Minority Over-sampling Technique (SMOTE) [7] and its variants have been particularly popular, creating synthetic

samples of the minority class to balance the training dataset. While effective, these methods sometimes generate unrealistic samples that can degrade model performance.

Generative Adversarial Networks (GANs) have emerged as a powerful alternative for synthetic data generation.[8] GANs consist of two neural networks—a generator and a discriminator—that work in tandem to produce realistic data samples. GANs have been successfully applied in various domains, including image generation and anomaly detection. However, training GANs can be challenging due to issues such as mode collapse and instability during training.[9]

Diffusion models represent a newer class of generative models that address some of the limitations of GANs. These models generate data by iteratively refining samples through a series of stochastic transformations, resulting in highly realistic and diverse outputs, one of the most popular proposals was explored by Sattarov et.al. in their research paper related to financial tabular synthetic data.[10] Diffusion models have shown great promise in fields such as image synthesis and natural language processing, but their application in fraud detection remains relatively unexplored.

This paper proposes leveraging diffusion models for synthetic data generation to enhance fraud detection systems by training networks on synthetic datasets. By generating high-quality synthetic fraudulent instances, it aims to mitigate the class imbalance problem and improve the overall performance and robustness of detection algorithms. This work builds on the foundation of previous research[11] while introducing novel applications of diffusion models in the realm of fraud detection.

Methods

This experiment aims to generate and test fraud synthetic data based on the diffusion model algorithm to improve the security of the transaction process by training models on the data augmented through ground truth.

The approach depicted in the diagram is an adaptation of the Duo-GAN technique, modified to employ a diffusion process. This methodology utilizes two separate diffusion blocks that independently learn from positive samples. Unlike the traditional GAN-based approach, the diffusion models in this case are designed to capture the class-conditional distributions and correlations within each class, allowing for a more nuanced understanding of the actual data distribution.

In this setup, the real data is first filtered to retrieve positive samples, which are then fed into each of the diffusion models operating in parallel. These models generate synthetic positive samples independently. The subsequent fusion of these samples results in a final synthetic dataset that amalgamates the strengths of both diffusion processes.

The ultimate goal of this process is to create an augmented or synthetic dataset that can be used to train classical fraud recognition systems, thereby improving their performance by enhancing the representation of minority classes

within the data (see Fig. 1). This diffusion-based approach offers a robust alternative to traditional methods, particularly in scenarios involving highly imbalanced datasets.

Figure 1 illustrates a dual-track process of synthetic data generation using diffusion models for fraud detection. The two parallel tracks represent independent diffusion processes, one for generating synthetic data from legitimate (positive) transaction samples and the other for generating synthetic data from fraudulent (negative) transaction samples.

In the first track, positive samples are taken from the real data and undergo a forward diffusion process where noise is gradually added, transforming them into less structured, noisy versions. This forward process allows the model to generalize the patterns within the legitimate transactions. In the reverse diffusion process, the model denoises the data to generate realistic synthetic legitimate transaction data.

The second track follows the same process but focuses on negative (fraudulent) samples. The diffusion model trained on these fraudulent samples learns the characteristics and statistical patterns of fraudulent behavior. Through the same forward and reverse diffusion process, realistic synthetic fraudulent transactions are generated.

Once the synthetic data from both tracks is generated, the two streams are combined to form a final dataset. This dataset integrates both legitimate and fraudulent transactions, balancing the original data's class distribution. The inclusion of both positive and negative samples is essential for addressing the class imbalance issue, which is a major challenge in fraud detection systems.

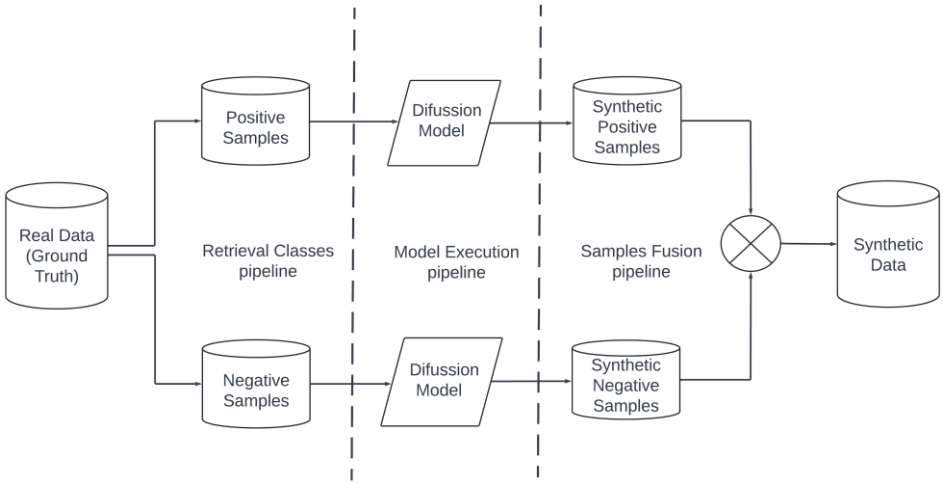

Figure 1: The blueprint of SDG using a diffusion model.

This combined synthetic dataset enhances the performance of detection algorithms by providing more diverse and realistic examples of both types of transactions.

A fraudulent transaction formalized on (1) can be modeled based on various characteristics and behaviors that differentiate it from legitimate transactions. Let's denote a transaction as a vector:

$$\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_\mathrm{m}), \tag{1}$$

Let, $\mathbf{W}_l = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_\mathrm{n}\}$ represent the set of weight vectors corresponding to legitimate transactions. Each $\mathbf{w}_\mathrm{l}$ is a vector representing typical values for the features. Let also μ_l (2) be the mean vector and Σ_l (3) be the covariance matrix of the legitimate transaction profile $\mathbf{W}_l$:

$$\mu_l = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{w}_i, \tag{2}$$

$$\boldsymbol{\Sigma}_l = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{w}_i - \mu_l)(\boldsymbol{w}_i - \mu_l)^T, \tag{3}$$

The similarity between a transaction and the legitimate transaction profile $\mathbf{W}_l$ is measured using the Mahalanobis distance, see (4) - (5), which takes into account the correlations between features and is more sensitive to multi-dimensional outliers:

$$D_m(\mathbf{x}, \mu_l, \boldsymbol{\Sigma}_l) = \sqrt{(\mathbf{x} - \mu_l)^T \boldsymbol{\Sigma}_l^{-1}(\mathbf{x} - \mu_l)}, \tag{4}$$

$$f(\epsilon_f) = \begin{cases} D_m(\mathbf{x}, \mu_l, \boldsymbol{\Sigma}_l), > \epsilon_f \implies \mathbf{x} \text{ is fraudulent} \\ D_m(\mathbf{x}, \mu_l, \boldsymbol{\Sigma}_l), \leq \epsilon_f \implies \mathbf{x} \text{ is legitimate} \end{cases}, \tag{5}$$

Where, ϵ_f a threshold for classifying a transaction as fraudulent.

Then, once the transactions are defined nature it assumes that the forward diffusion process gets started (6), let's define the initial data and forward diffusion process:

$$\mathbf{x}_0 \sim q(\mathbf{x}_0), \tag{6}$$

where, $\mathbf{x}_0$ - represents a data point in the original (undistributed) data space; $\mathrm{q}(\mathbf{x}_0)$ - represents the probability distribution from which the data point $\mathbf{x}_0$ is sampled.

In the forward process, data is diffused by gradually adding Gaussian noise (7). For a given data sample x_0, the noisy version at step t, x_t:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\alpha_t}\ \mathbf{x}_{t-1}, (1-\alpha_t)\mathbf{I}\right), \quad (7)$$

where, $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ - represents the probability distribution of the noisy data point x_t given the previous step x_{t-1};

$\mathcal{N}\left(\mathbf{x}_t; \sqrt{\alpha_t}\ \mathbf{x}_{t-1}, (1-\alpha_t)\mathbf{I}\right)$ - specifies that x_t is drawn from a Gaussian distribution with mean $\sqrt{\alpha_t}\ \mathbf{x}_{t-1}$ and variance $(1-\alpha_t)\mathbf{I}$;

α_t – is defined as $1-\beta_t$, where β_t represents a variance schedule that controls the noise level at each step.

This is a step where Gaussian noise is gradually added to the data to produce increasingly noisy versions. The goal is to transform the original data into a distribution that is close to pure noise.

Important to note that the Gaussian noise is applied uniformly across all types of data, including both numerical and categorical (textual) values. For the numerical features, Gaussian noise is directly added as part of the diffusion process. For categorical (textual) data, these values are first encoded into numerical representations using techniques like one-hot encoding or embedding vectors. Once the textual data is converted into numerical form, Gaussian noise is applied similarly to these encoded values during the forward diffusion process. This ensures that both numerical and textual data are treated uniformly and integrated into the synthetic data generation pipeline. This approach enhances the diversity and realism of the generated data across both feature types.

Then, over multiple steps, this becomes (8):

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\ \mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right), \quad (8)$$

where, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

Let's define the reverse diffusion process. The reverse process aims to denoise $\mathbf{x}_t$ to reconstruct $\mathbf{x}_0$ in (9). The reverse process is defined by:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_\theta^2(\mathbf{x}_t, t)\mathbf{I}\right), \quad (9)$$

where, the parameters μ_θ and σ_θ^2 are learned through the neural network with parments θ. The parameters μ_θ and σ_θ^2 are estimated by minimizing the Kullback-Leibler divergence (10) between the true posterior q and the approximate posterior p_θ, which can be formulated as:

$$\mathcal{L}(\theta) = \mathbb{E}_q[\textstyle\sum_{t=1}^{T} D_{KL}(q(\mathbf{x}_{t-1}|\,\mathbf{x}_t, \mathbf{x}_0)\ ||\ p_\theta(\mathbf{x}_{t-1}|\ (\mathbf{x}_t))], \quad (10)$$

Here the model parmeter θ are updated iteratively (11) to minimize the loss function using gradient descent:

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\theta), \quad (11)$$

Then, let's formalize synthetic data generation:

1. Sampling noise generates random noise vectors $\mathbf{z} = \mathcal{N}(0, \mathbf{I})$, here, $\mathbf{z} = (z_1, z_2, \dots, z_d)$ is a d-dimensional vector where each $\{z_i\}$ is an independent sample from the normal distribution $\mathcal{N}(0, 1)$;
2. Apply the reverse diffusion process to transform noise $\mathbf{z}$ into synthetic data samples, as described in (12):

$$\mathbf{x}_{\text{synthetic}}: \mathbf{x}_{t-1} \sim p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t),$$

$$\mathbf{x}_{t-2} \sim p_\theta(\mathbf{x}_{t-2}|\mathbf{x}_{t-1}), \dots, \mathbf{x}_0 \sim p_\theta(\mathbf{x}_0|\mathbf{x}_1); \quad (12)$$

3. The generated sample is described as: $\mathbf{x}_{\text{synthetic}} = \mathbf{x}_0$.

Experiments

The experiments aim to evaluate the effectiveness of Synthetic Data Generation (SDG) using diffusion models as an oversampling technique for enhancing fraud detection in imbalanced datasets. We compared the performance of various classification algorithms when trained on datasets augmented with synthetic data generated by diffusion models.

To test the following hypotheses:

- **H1**: The use of diffusion models, especially Denoising Diffusion Probabilistic Models (DDPMs), introduced by Sohl-Dickstein et al., and later improved by Ho et al., for generating synthetic data will improve algorithmic performance in baseline experiments on benchmark imbalanced datasets.
- **H2**: The synthetic data generated by diffusion models will enhance the performance of classification algorithms in real-world fraud detection scenarios.

These hypotheses were tested by augmenting the original datasets with synthetic data generated using the following methods:

1. **Diffusion Models**: Generates synthetic data by iteratively refining samples through a series of transformations.
2. **SMOTE (Synthetic Minority Over-sampling Technique)**: Generates synthetic samples by interpolating between existing minority class instances.
3. **ADASYN (Adaptive Synthetic Sampling Approach)**: Similar to SMOTE but focuses more on difficult-to-learn examples.
4. **Borderline-SMOTE**: Generates synthetic samples near the border between classes to define the decision boundary better.

Hyperparameters Settings

Let's summarize the model's configuration settings (see Tab.1). N/A means that it's not applicable for traditional sampling methods.

The fraud detection is performed using classical supervised learning models. The experimental setting involved training classification algorithms on datasets augmented with synthetic data generated by the diffusion models. Specifically, we used Logistic Regression (LR), Random Forest (RF), XGBoost, and Multi-Layer Perceptron (MLP) models to identify fraud after they were trained on these augmented datasets.

Table 1. Environment configuration.

Hyperparameters	Diffusion Models	SMOTE	ADASYN	Borderline-SMOTE
Learning Rate	1×10^{-4}	N/A	N/A	N/A
Optimizer	Adam	N/A	N/A	N/A
Epochs	50	N/A	N/A	N/A
Batch Size	64	N/A	N/A	N/A
Diffusion Steps	1000	N/A	N/A	N/A
Noise Schedule	Linear	N/A	N/A	N/A
k_neighbors	N/A	5	5	5
m_neighbors	N/A	N/A	N/A	10
Activation Function	ReLU	N/A	N/A	N/A
Noise Distribution	N(0, 1)	N/A	N/A	N/A
Sampling Strategy	N/A	auto	auto	auto

The synthetic data generated by diffusion models is integrated with real-world datasets described below, providing a balanced training set that helps these classification models perform more robustly against fraud detection, especially for minority classes such as fraudulent transactions.

This clarified experimental setting highlights that the fraud detection is done using these well-known classifiers, while the diffusion models generate the synthetic data to balance the dataset and improve detection accuracy.

Datasets Used

Benchmark Datasets

The method was evaluated on several benchmark imbalanced datasets,[12] see Tab. 2. There are a popular existing datasets that contain different aspects of the particular domain.

Table 2. There are different datasets containing transaction samples from a different domain.

Dataset Name	Domain	Number of features	Number of Instances	Imbalance Ratio (IR)
Credit Card Fraud Detection Dataset	Finance	30	284,807	1:577
Online Retail II Dataset	Retail	8	541,909	1:25
E-commerce Transaction Dataset	E-commerce	10	100,000	1:20

Real-World Datasets

The IEEE-CIS Fraud Detection Dataset is utilized as the real-world dataset. This dataset contains anonymized real-world e-commerce transactions provided by Vesta Corporation. It includes a wide range of features from device type to product information.[13] See the Tab.3.

Table 3. There is one of the popular e-commerce datasets with real-world scenario data.

Dataset Name	Domain	Number of features	Number of Instances	Imbalance Ratio (IR)
IEEE-CIS Fraud Detection Dataset	E-commerce	67	561,013 (training), 28,527 (testing)	1:28.6

Before using the datasets from the Tab.1 and Tab.2, their attribute values were scaled to the interval [0, 1] using the min-max normalization method to standardize the range of all attributes and prevent any single attribute from dominating the others due to its scale.

Results

The baseline model serves as a reference point for evaluating the performance of other models or methods. It is the model trained on the original, unaugmented dataset without any synthetic data generation techniques applied.

The use of synthetic data generated by diffusion models significantly improves the accuracy of fraud detection algorithms. Among the various methods tested, diffusion models provided the highest improvements in terms of precision, recall, F1-score, and ROC-AUC, demonstrating the effectiveness of using these advanced generative models for generating synthetic data and enhancing fraud detection systems. See Tab.4.

Table 4. Transaction detection results: recall, precision, and F1 measure.

Method	Precision	Recall	F1-Score	ROC-AUC
Baseline Model	0.85	0.80	0.82	0.87
(SDG) Diffusion Models	0.90	0.88	0.89	0.92
SMOTE	0.88	0.85	0.86	0.89
ADASYN	0.89	0.86	0.87	0.90
Borderline-SMOTE	0.87	0.84	0.85	0.88

It worth to mention that the Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) is a performance measurement for classification problems at various threshold settings. ROC is a graphical plot that illustrates the diagnostic ability of a binary classifier, while AUC measures the area under the ROC curve, providing a single scalar value that indicates the overall performance of the model. A higher AUC indicates better model performance, with 1.0 being the optimal score and 0.5 indicating random guessing.

Figure 2 presents the measurement results across all evaluated methods. Below, the chart results are analyzed in detail:

1. **Diffusion Models:** The diffusion models significantly outperform the other methods across all metrics (see Fig. 2). They achieve the highest scores in Precision, Recall, F1-Score, and ROC-AUC, demonstrating their effectiveness in improving the performance of fraud detection algorithms. This suggests that diffusion models are highly effective in generating synthetic data that enhances the accuracy and robustness of the detection systems.
2. **SMOTE and ADASYN:** These two methods also improve the performance of the fraud detection algorithms compared to the baseline. However, their performance is slightly lower than that of the diffusion models (see Fig.2). SMOTE and ADASYN are traditional oversampling techniques that help to balance the dataset, but they might generate less realistic synthetic samples compared to diffusion models.
3. **Borderline-SMOTE:** This method shows better performance (see Fig.2) than the baseline but is generally less effective than both diffusion models and the other oversampling techniques (SMOTE and ADASYN). It focuses on generating samples near the decision boundary, which helps in improving the detection rate but doesn’t match the overall performance of the diffusion models.

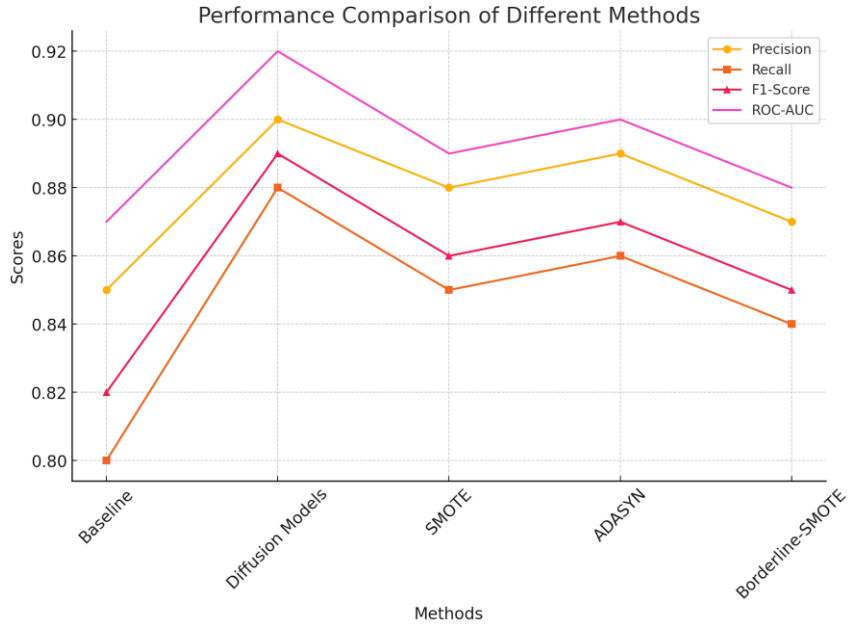

Figure 2: Experiments results Precison & Recall, F1, ROC-AUC.

4. **Baseline Model:** The baseline model, which does not use any synthetic data generation, shows the lowest performance across all metrics (see Fig.2). This indicates the importance of synthetic data generation in enhancing the detection of fraudulent transactions, especially in imbalanced datasets.

Conclusions

The experiments conducted in this study aimed to evaluate the effectiveness of SDG using diffusion models as an oversampling technique for enhancing fraud detection in imbalanced datasets. The results demonstrate that the denoising diffusion models significantly improve the accuracy of fraud detection algorithms compared to traditional methods such as SMOTE, ADASYN, and Borderline-SMOTE.

Key Findings:

1. **Improved Detection Performance**: The diffusion models outperformed other methods in terms of precision, recall, F1-score, and ROC-AUC, indicating a higher accuracy in detecting fraudulent transactions.
2. **Enhanced Model Robustness**: The use of diffusion models led to more robust fraud detection systems capable of handling various types of fraud scenarios effectively.

3. **Addressing Class Imbalance**: By generating high-quality synthetic data that closely mimic real fraudulent transactions, diffusion models effectively mitigated the class imbalance problem, which is a significant challenge in fraud detection.
4. **Baseline Comparison**: The baseline model, trained on the original unaugmented dataset, provided a reference point, showing significant improvements when synthetic data from diffusion models was introduced.

Hypotheses Validation:

- **H1**: Using diffusion models for generating synthetic data improved algorithmic performance in baseline experiments on benchmark imbalanced datasets.
- **H2**: Synthetic data generated by diffusion models enhanced the performance of classification algorithms in real-world fraud detection scenarios.

Summary

The findings from this research highlight the potential of diffusion models as a valuable tool in the field of fraud detection. By generating realistic synthetic data, these models can significantly enhance the efficacy of fraud detection systems, making them more accurate and robust. The results suggest that implementing diffusion models can be a promising approach to overcoming the limitations of traditional data augmentation techniques, particularly in scenarios involving highly imbalanced datasets.

Future Work

Future research could explore the application of diffusion models to other domains where class imbalance is a significant issue. Additionally, further optimization of diffusion model parameters and integration with other advanced deep learning techniques such as transformers, by using attention mechanisms could yield even better performance in fraud detection and other anomaly detection tasks.

This study demonstrates the practical benefits of using advanced generative models to improve the detection and prevention of fraudulent activities, providing a strong foundation for future advancements in this critical area.

References

[1] Nitesh Chawla, Nathalie Japkowicz and Aleksander Kołcz, “Editorial: Special Issue on Learning from Imbalanced Data Sets,” *SIGKDD Explorations* 6, no. 1 (2004): 1-6, http://dx.doi.org/10.1145/1007730.1007733.

[2] Emilija Strelcenia and Simant Prakoonwit, “A Survey on GAN Techniques for Data Augmentation to Address the Imbalanced Data Issues in Credit Card Fraud

Detection," *Machine Learning and Knowledge Extraction* 5, no. 1 (2023): 304-329, https://doi.org/10.3390/make5010019.

[3] Vladimir Zaslavsky and Anna Strizhak, "Credit Card Fraud Detection Using Self-Organizing Maps," *Information & Security: An International Journal* 18 (2006), https://doi.org/10.11610/isij.1803.

[4] Alexander Schuchter and Levi Michael, "The Fraud Triangle Revisited," *Security Journal* 29 (2016): 107–121, https://doi.org/10.1057/sj.2013.1.

[5] Hala Alenzi and Nojood O. Aljehane, "Fraud Detection in Credit Cards using Logistic Regression," *International Journal of Advanced Computer Science and Applications* 11, no. 12 (2020), http://dx.doi.org/10.14569/IJACSA.2020.0111265.

[6] Saksham Jain, Gautam Seth, Arpit Paruthi, Umang Soni, and Girish Kumar, "Synthetic data augmentation for surface defect detection and classification using deep learning," *J. Intell. Manuf*. 33, no. 4 (2022): 1007–1020, https://doi.org/10.1007/s10845-020-01710-x.

[7] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research* 16 (2002): 321-357, https://doi.org/10.1613/jair.953.

[8] Charitos Charitou, Simo Dragicevic, and Artur d'Avila Garcez, "Synthetic Data Generation for Fraud Detection using GANs," *arXiv preprint*, arXiv:2109.12546 (2021), https://doi.org/10.48550/arXiv.2109.12546.

[9] Mengran Zhu, Yulu Gong, Yafei Xiang, Hanyi Yu, and Shuning Huo, "Utilizing GANs for Fraud Detection: Model Training with Synthetic Transaction Data," *International Conference on Image, Signal Processing, and Pattern Recognition (ISPP 2024)*, vol. 13180, 2024, pp. 887-894, https://doi.org/10.48550/arXiv.2402.09830.

[10] Timur Sattarov, Marco Schreyer, and Damian Borth, "Findiff: Diffusion Models for Financial Tabular Data Generation," *Proceedings of the Fourth ACM International Conference on AI in Finance*, 2023, pp. 64-72, https://doi.org/10.48550/arXiv.2309.01472.

[11] Chaocheng Yang, Tingyin Wang, and Xuanhui Yan, "DDMT: Denoising Diffusion Mask Transformer Models for Multivariate Time Series Anomaly Detection," *arXiv preprint* arXiv:2310.08800, 2023, https://doi.org/10.48550/arXiv.2310.08800.

[12] Prince Grover, Julia Xu, Justin Tittelfitz, Anqi Cheng, Zheng Li, Jakub Zablocki, Jianbo Liu, Hao Zhou "Fraud Dataset Benchmark and Applications," *arXiv preprint arXiv:2208.14417* (2022), https://doi.org/10.48550/arXiv.2208.14417.

[13] Addison Howard, Bernadette Bouchon-Meunier, IEEE CIS, inversion, John Lei, Lynn@Vesta, Marcus2010, Hussein Abbass, *IEEE-CIS Fraud Detection*, *Kaggle*, 2019, https://kaggle.com/competitions/ieee-fraud-detection.

About the Authors

Yurii **Pushkarenko** is an AI Architect / Technical Lead with 15+ years of experience in software engineering and solution architecture. He specializes in AI, ML, and DL solutions with deep expertise in cloud systems (AWS), routing/ navigation, search, and computer vision for remote sensing systems. Yurii has led teams of 50+ engineers, managing large-scale projects in the automotive, telecom, and banking sectors. Researcher and AI consultant with experience in data fusion, synthetic data generation, and data processing. He has consulted some military units on sensor data fusion, visual (ontology) search, and situational awareness tools (like Palantir). Currently, he is a PhD candidate in Computer Science at Taras Shevchenko National University of Kyiv and lectures on formal methods, data mining, and Explainable AI.
E-mail: yurii.pushkarenko@gmail.com, https://orcid.org/0009-0007-2560-2971

Prof. Vladimir A. **Zaslavskyi** is Head of the "Mathematical Methods for Ecological and Economic Research" department, Faculty of Cybernetics, Taras Schevchenko National University of Kiev. He was Vice Dean of the Faculty of Cybernetics in the period 2000-2004. Dr. Zaslavsky received his PhD in Mathematics from the Faculty of Cybernetics in 1984. He has been an Associate Professor since 1992 and has more than 100 publications in the areas of system analysis of complex systems, risk analysis, reliability optimization and redundancy, and decision support systems. He is a member of IIASA Society and President of the AFCEA –Ukraine Chapter. E-mail: zas@unicyb.kiev.ua.
https://orcid.org/0000-0001-6225-1313