Related Work

Generating Synthetic Credit Card Fraud Data with Advanced Generative Models for Enhanced Fraud Detection

Akash Murali, Anish Rao, Raghu Ram Sattanapalle

Paper 1: FraudDiffuse: Diffusion-aided Synthetic Fraud Augmentation for Improved Fraud Detection
The selected paper, "FraudDiffuse: Diffusion-aided Synthetic Fraud Augmentation for Improved Fraud Detection" by Ruma Roy, Darshika Tiwari, and Anubha Pandey, serves as our primary reference. Our project would use the above reference with a different dataset (Fraud ecommerce + Sparkov dataset).

Comparative Analysis:

1. FraudDiffuse employs diffusion models to generate synthetic fraud data using the IEEE-CIS fraud dataset.
2. Our work will utilize the Fraud Ecommerce and Sparkov datasets, providing a new benchmark for diffusion-based fraud detection models.

Our Findings:

1. The FraudDiffuse paper demonstrated the effectiveness of synthetic data augmentation using diffusion models, leading to improved fraud detection performance.
2. Our contribution lies in expanding the scope of diffusion models to different datasets, potentially uncovering unique fraud patterns and enhancing generalizability.
3. We aim to introduce novel preprocessing techniques and evaluate model performance with enhanced evaluation metrics.

Our work contributes by applying diffusion models to a new combination of datasets, comparing results with existing approaches, and exploring the potential of diffusion models in fraud detection across diverse datasets. This work will provide insights into model generalization and adaptability across different fraud detection contexts.

Paper 2: FinDiff: Diffusion Models for Financial Tabular Data Generation
The above study uses a different dataset and same model (Diffusion model) albeit with some small differences.

The diffusion model used in the above study is Financial Tabular Diffusion (FinDiff as they call it). It is a diffusion-based generative approach designed for financial tabular datasets for a

variety of tasks. FinDiff has been applied across tasks such as economic scenario modeling, stress testing, and fraud detection. Our planned approach also uses a diffusion model—FraudDiffuse—from a separate study, but with some small differences. While FinDiff is a general-purpose diffusion model for financial tasks, the model that we aim to utilize is tailored specifically for fraud detection. We plan to adapt FraudDiffuse to a new dataset, such as the Sparkov Fraud Detection or the Fraud Ecommerce dataset, both of which present significant class imbalance challenges. By exploring the adaptability of the Diffusion model on a dataset that has not been previously tested with this model, we plan to extend its application and demonstrate its usefulness in handling real-world data scenarios.

Paper 3: Imb-FinDiff: Conditional Diffusion Models for Class Imbalance Synthesis of Financial Tabular Data

The above study also uses a diffusion model but applies it to a different dataset.

Their diffusion model, Imb-FinDiff, is a denoising diffusion framework specifically designed to address class imbalance in financial tabular datasets. It works by generating synthetic samples for the minority class while preserving the statistical properties of the original data. The model incorporates the features using embeddings and focuses on two main objectives: minimizing noise at each diffusion time step and predicting class labels. Our planned approach also uses a diffusion model but focuses specifically on fraud detection tasks. While Imb-FinDiff is designed for general financial data, FraudDiffuse introduces fraud-specific optimizations, such as the contrastive learning loss function, to capture patterns near the decision boundary. We aim to apply the Diffusion model to new datasets like the Spark or Fraud Ecommerce datasets, which both have significant class imbalance challenges. By exploring the adaptability of the Diffusion model on a dataset that has not been previously tested with this model, we plan to extend its application and demonstrate its usefulness in handling real-world data scenarios.

Paper 4: TabDDPM: Modelling Tabular Data with Diffusion Models

The selected paper, "TabDDPM: Modelling Tabular Data with Diffusion Models," also serves as a valuable reference for our study.

The TabDDPM study applies diffusion models to generic tabular datasets with mixed numerical and categorical data, emphasizing performance on benchmark datasets for general-purpose tabular data modeling. Our work applies diffusion models on the Fraud Ecommerce and Sparkov datasets, targeting e-commerce fraud detection and focusing on fraud pattern identification in online transactions. Our datasets contain complex transactional behaviors unique to e-commerce fraud, requiring tailored preprocessing and feature engineering techniques that differ from those used in the generic datasets employed by TabDDPM.

By tailoring diffusion models to fraud-specific datasets, we aim to contribute novel insights into improving fraud detection models with synthetic data augmentation. Our study will demonstrate the potential benefits of using diffusion models in high-stakes financial fraud applications.

Paper 5: Synthetic Data Generation for Fraud Detection Using Diffusion Models

The paper "Synthetic Data Generation for Fraud Detection Using Diffusion Models" by Pushkarenko and Zaslavskyi (2024) explores the potential of diffusion models for generating high-quality synthetic transaction data, to address the problem of class imbalance and improve the performance of fraud detection systems. Their experiments include several benchmark datasets, as well as the IEEE-CIS dataset for real world evaluation. The models were trained using a combination of synthetic and real data, and the fraud detection was then performed using classification models. Their evaluation included the standard metrics of Precision, Recall, F1-Score and ROC-AUC. Unlike our work, which focuses primarily on the Fraud e-commerce and Sparkov datasets, their experiments use benchmark financial data, and a popular real-world e-commerce dataset (IEEE-CIS). Furthermore, while they demonstrate the effectiveness of diffusion models for synthetic data generation, their method is a custom implementation that uses two separate diffusion models, and not the FraudDiffuse model that our project aims to evaluate. Also, they do not use TABDDPM, or focus on specific parameters of a diffusion model, or the effects of the architectural differences. To validate our approach and further understand the contribution of this paper, we plan to implement a simplified version of their dual-track approach to compare against the performance of our TABDDPM, and FraudDiffuse models on our chosen datasets. This will give a strong point of comparison and allow us to understand the differences in model performance. Additionally, we will also use their results on the IEEE-CIS data to validate our own model performance. By focusing specifically on the FraudDiffuse model, along with TABDDPM, we will explore the unique properties of diffusion models for generating high-quality synthetic data, for the specific task of fraud detection. Furthermore, we will be testing this model on the Fraud e-commerce and Sparkov datasets, and not solely relying on the IEEE-CIS or benchmark datasets, adding to the body of knowledge in this area. Finally, we also plan on highlighting the importance of using evaluation metrics beyond the standard benchmarks in order to more fully understand model performance. This will allow our study to contribute a new perspective in evaluating diffusion models for synthetic fraud data generation, and also help inform best practices on how these models should be evaluated for the task of fraud detection.