National College of Ireland

A Comparative Study of Generative Oversampling Methods for Imbalanced Fraud Classification

MSc Research Practicum
MSc. in Data Analytics

Jisoo Park
Student ID: 23376589

School of Computing
National College of Ireland

Supervisor: Shubham Subhnil

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Jisoo Park……

Student ID: 23376589……

Programme: MSc in Data Analytics…… **Year:** 2024/2025……..

Module: MSc Research Practicum……

Supervisor: Shubham Subhnil……

Submission Due Date: 11/08/2025……

Project Title: A Comparative Study of Generative Oversampling Methods for Imbalanced Fraud Classification……

Word Count: 5917…… **Page Count:**20……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Jisoo Park……

Date: 11/08/2025……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Comparative Study of Generative Oversampling Methods for Imbalanced Fraud Classification

Jisoo Park
23376589

Abstract

Credit card fraud detection poses a significant challenge due to extreme class imbalance, where fraudulent transactions make up less than 0.2% of the data. While generative oversampling methods—particularly Generative Adversarial Networks (GANs)—have shown promise in addressing this issue, many existing approaches lack classifier-awareness, rely on limited evaluation metrics, and fail to align synthetic data with real-world decision boundaries.

This study proposes an Enhanced Conditional GAN that integrates feature matching loss, gradient penalty, and conditional generation to create classifier-aligned synthetic fraud samples. We conduct a comprehensive comparison of this model against traditional oversample method (SMOTE, ADASYN) and generative models (CTGAN, TVAE, Gaussian Copula) using a real-world imbalanced fraud dataset.

Results show that the proposed model achieves an F1-score of 0.8619, PR-AUC of 0.8591, and ROC-AUC of 0.9620, achieving competitive PR-AUC and competitive precision among all generative approaches, and closely matching the no-oversampling baseline in overall balance. Statistical and visual analyses reveal that the Enhanced GAN prioritizes decision-relevant sample generation over exact distributional matching, resulting in better classification outcomes.

This work highlights the importance of classifier-aware oversampling and provides a robust framework for evaluating synthetic data quality in high-risk, imbalanced domains such as fraud detection. Future directions include real-time deployment, model compression, and benchmarking against diffusion and LLM-based tabular generators

1 Introduction

The rise of online financial services has caused a sharp increase in credit card usage. At the same time, fraud cases have also grown. Credit card fraud detection systems are now essential for financial safety. However, these systems face a serious challenge. Fraud cases are very rare and usually make up less than 0.2% of all transactions (Dal Pozzolo et al., 2015). This extreme imbalance makes it hard for standard machine learning models to learn from the data. As a result, many models miss fraud cases. This leads to low recall and many false negatives. These errors can cause large financial losses and damage trust (Bahnsen et al., 2016; Dal Pozzolo et al., 2015).

To fix this imbalance, researchers have started using generative oversampling methods. These methods create fake but realistic fraud samples to help the model learn. In particular, Generative Adversarial Networks (GANs) are often used. GANs can generate synthetic data

that looks similar to real fraud cases. This makes them useful for improving classifier performance in fraud detection tasks (Fiore et al., 2019; Adiputra et al., 2025). These models can enhance classifier performance by generating plausible fraud samples that mitigate class imbalance.

Several GAN variants have been adapted for tabular data. For example, CTGAN introduces conditional generation and mode-specific normalization to handle mixed-type features effectively, although it does not incorporate classifier feedback during training (Xu et al., 2019). CopulaGAN extends GANs using copula theory to capture statistical dependencies in tabular data, with a focus on distributional similarity rather than classification utility (Patki et al., 2016). WGAN-GP, commonly used as a base architecture, enhances training stability through a gradient penalty that enforces the Lipschitz constraint—making it more robust across data types (Gulrajani et al., 2017)

Despite recent progress, many studies still miss three important areas.

1) **Classifier-agnostic generation**
   Most GAN-based models focus only on copying the data distribution. They do not check if the generated data helps the classifier. This means the model may look realistic but does not improve fraud detection (Oreški, 2023; Fonseca and Bação, 2023).
2) **Shallow evaluation metrics**
   Many papers report only overall results like F1 or ROC-AUC. They often skip class-wise scores or do not analyze how the model handles fraud-specific boundaries. This limits how useful the results are in real-world tasks (Strelcenia and Prakoonwit, 2023; Adiputra et al., 2025).
3) **Incomplete benchmarking**
   Some newer models like TVAE or Gaussian Copula are often ignored. Many studies also fail to compare deep learning with simpler statistical methods. As a result, we cannot tell which approach works best across different cases (Jiang et al., 2023; Chen et al., 2024).

Some recent studies have combined GANs with autoencoders to improve the quality of synthetic data. These hybrid models aim to create better samples by learning stronger feature representations before generation. However, they often fail to show consistent improvements compared to simpler methods like SMOTE or basic GANs. In many cases, the added complexity does not lead to better results on fraud detection tasks. Metrics such as F1-score and precision-recall balance often stay the same or even get worse (Oreški, 2023; Fonseca and Bação, 2023).

This suggests that a more complex model is not helpful unless it also supports the goals of the classifier (Strelcenia and Prakoonwit, 2023).In contrast, traditional oversampling methods like SMOTE and ADASYN are still widely used. These methods are simple and fast to apply. They work by adding new samples between existing minority-class points. But they do not consider how the classifier separates classes (Chawla et al., 2002; He et al., 2008). As a result, they often increase recall but sharply reduce precision. This causes false positives. These false alarms can confuse the classifier. The model may start to learn noise instead of real fraud signals. It then becomes unstable. Its performance may drop when tested on new or real-world data (Jiang et al., 2023).

Creating synthetic fraud data is important because fraud is very rare. In most real datasets, fraud cases make up less than 0.2 percent of all transactions (Dal Pozzolo et al., 2015). This makes it hard for the model to learn what fraud looks like. If we do not increase the number of fraud samples, the model will miss most of them. This leads to low recall and many false negatives.

Synthetic samples help by giving the model more chances to learn from the fraud class. This severe imbalance limits the model's ability to learn meaningful fraud patterns, causing low recall and a high false negative rate. Synthetic oversampling helps by increasing the representation of the fraud class during training, allowing the model to better detect subtle fraud signals. However, quantity alone is not enough. To be useful, synthetic samples must do more than copy real data. They should add variation that helps the classifier learn better decision boundaries (Chen et al., 2024; Mariani et al., 2020). In fields like fraud detection, where errors are costly, oversampling must not only fix class imbalance. It must also guide the model toward better performance by supporting its learning directly.

1.1 Research Objectives and Contributions

In this work, we propose a classifier-aware Enhanced Conditional GAN that explicitly targets downstream fraud detection performance. Our contributions are as follows:

1) Develop an Enhanced GAN architecture that integrates feature matching loss, gradient penalty, and conditional generation to synthesize fraud samples that are structurally aligned with classifier-relevant representations.
2) Perform a comprehensive comparative analysis of generative (CTGAN, TVAE, Gaussian Copula) and traditional (SMOTE, ADASYN) oversample method using a real-world imbalanced credit card fraud dataset.
3) Evaluate models across a suite of performance metrics including F1-score, PR-AUC, specificity, and class-wise confusion matrices, as well as distributional fidelity measures such as KS-test and Jensen–Shannon Distance.
4) Demonstrate that, despite deviating from the exact data distribution, our Enhanced GAN consistently achieves superior classifier performance—highlighting the practical trade-off between statistical realism and decision-aligned utility.

By framing oversampling as a classifier-aware optimization problem rather than a pure data synthesis task, this study provides a principled foundation for improving fraud detection in highly imbalanced tabular domains.-

2 Related Work

2.1 Traditional Oversampling Approaches: Still Dominant, Yet Fundamentally Limited

Traditional methods such as SMOTE (Chawla et al. 2002) ADASYN (He et al. 2008) and Random Oversampling are still widely used. They are popular because they are simple to understand and require little computing power. These methods create synthetic samples by adding new points between existing examples from the minority class. They do not consider how the classifier separates the classes.

As a result these methods often increase recall but lower precision. This can be a serious problem in fraud detection where false positives can cause high financial or operational costs (Jiang et al. 2023). These methods are straightforward to use but they may fail in complex tasks. However, they often become unstable in complex tasks. They also perform poorly with high-dimensional or non-linear data, which are common in fraud detection.

In addition, many recent studies still compare new models to these older methods. But they often ignore more modern approaches like statistical or deep generative models. This leads to unfair comparisons and may give a false impression that newer models are much better than they really are (Jiang et al. 2023).

2.2 GAN-based Tabular Models: Promising, Yet Often Superficial

Many researchers have adapted Generative Adversarial Networks (GANs) to work with tabular data. Well-known models include CTGAN (Xu et al., 2019), CopulaGAN (Patki et al., 2016), and WGAN-GP (Gulrajani et al., 2017). These models generate synthetic samples that resemble real fraud cases. They are trained to match the original data distribution as closely as possible.

However, most of these models do not focus on helping the classifier. Their goal is to create realistic data, not to improve fraud detection. As a result, the synthetic samples may look accurate but do not always help the model make better decisions. In some cases, it may even introduce noise (Strelcenia and Prakoonwit 2023).

Adiputra et al. (2025) tested several GAN variants on credit scoring tasks. They included models like WGAN-GP, CTGAN, DraGAN, and CopulaGAN. Their results showed that these models can produce good-looking data. But the evaluation focused mostly on F1-score. The study did not look at how the generated samples affect specific metrics like precision or recall. It also did not explore how the synthetic data changes the decision boundary. Without this kind of analysis, it is hard to know whether the models really help in real-world fraud detection. Notably, models like TVAE or classifier-aware GANs are often omitted, resulting in incomplete and potentially biased exploration of the model space.

2.3 Hybrid Models (GAN + Autoencoder): More Complex, Not Necessarily Better

To address limitations in basic GANs, some studies have proposed combining them with autoencoders to improve latent space representations before generation. Although this hybridization aims to boost sample quality, results have been inconsistent. Oreški (2023) found that while reconstruction improves, these models often suffer from training instability and fail to consistently outperform simpler alternatives such as SMOTE or standalone GANs.

Additionally, most hybrid models remain disconnected from the classification task. Without incorporating classifier feedback, they risk generating visually plausible but decision-irrelevant data—limiting their utility in high-stakes domains like fraud detection.

2.4 Non-GAN Generative Methods: Underused but Worthy of Attention

Alternative approaches such as Variational Autoencoders (TVAE), Gaussian Copula models, and diffusion-based generators offer promising directions for tabular data synthesis. TVAE, for instance, provides stable training and probabilistic structure, while Gaussian Copula models offer statistical interpretability and efficiency (Fonseca and Bação, 2023). However, these methods are usually tested in general-purpose settings and rarely in fraud-specific environments.

As a result, their effectiveness under extreme imbalance, high precision requirements, and concept drift remains unclear. Without domain-specific validation, they risk being statistically elegant but practically irrelevant. Recently, models like TabLLM and TabDDPM have been proposed, but their use in fraud detection remains at an early, exploratory stage (Arxiv.org, 2016; Arxiv.org, 2024).

2.5 Evaluation and Benchmarking Practices: Superficial and Selective

Many studies rely too much on broad metrics like macro F1 or ROC-AUC. These scores hide class-level problems and do not show real-world risks. For example, they often miss low precision in fraud detection (Chen et al., 2024).

Important tools like class-wise metrics, PR-AUC, and confusion matrices are often missing. Without them, it is hard to tell if synthetic data actually helps the model make better decisions.

Benchmarking inconsistencies also plague the field. Many studies use custom pipelines or fail to compare across deep generative and statistical models under identical conditions, undermining result comparability (Jiang et al., 2023). In some cases, evaluation designs appear to favor specific architectures or objectives, rather than providing a balanced view of generative model utility.

2.6 Summary of Gaps and Positioning

Despite architectural creativity, the literature suffers from several persistent issues:

1) **Lack of classifier-awareness**: Most synthetic data generators do not optimize for downstream fraud detection performance.
2) **Shallow evaluation**: Key class-wise and decision-relevant metrics are often ignored.
3) **Incomplete model inclusion**: Newer or non-neural models are frequently left out of comparisons.
4) **Benchmarking inconsistency**: Many studies lack a unified experimental setup, reducing reproducibility and validity.

To address these gaps, this study proposes a classifier-aware Enhanced Conditional GAN and conducts a principled comparison against traditional and generative oversamplers using a consistent pipeline and real-world fraud data. The focus is shifted from generating data that looks real to generating data that improves classifier decisions—a distinction essential for practical deployment.

3 Methodology and Implementation

3.1 Dataset and Preprocessing

We use a publicly available credit card fraud detection dataset, where fraudulent cases comprise less than 0.2% of total samples—posing a significant class imbalance (Dal Pozzolo et al., 2015).

The dataset is split into training and test sets via stratified sampling to ensure that the severe class imbalance is preserved in both sets, maintaining realistic evaluation conditions.

3.1.1 Exploratory Data Analysis (EDA) Summary

1) Total Samples: 284,807
2) Fraud Cases: 492 (≈ 0.17%)
3) Imbalance Ratio: ~1:577 (fraud : non-fraud)
4) Skewness:

- Amount is right-skewed → log-transformed to normalize distribution and reduce the influence of extreme values.
- Time shows periodic trends (e.g., daily cycles), which may contain behavioral patterns useful for fraud detection.
- Highlights the need for rare-class–sensitive augmentation techniques because both feature skewness and temporal periodicity can bias the learning process if unaddressed.

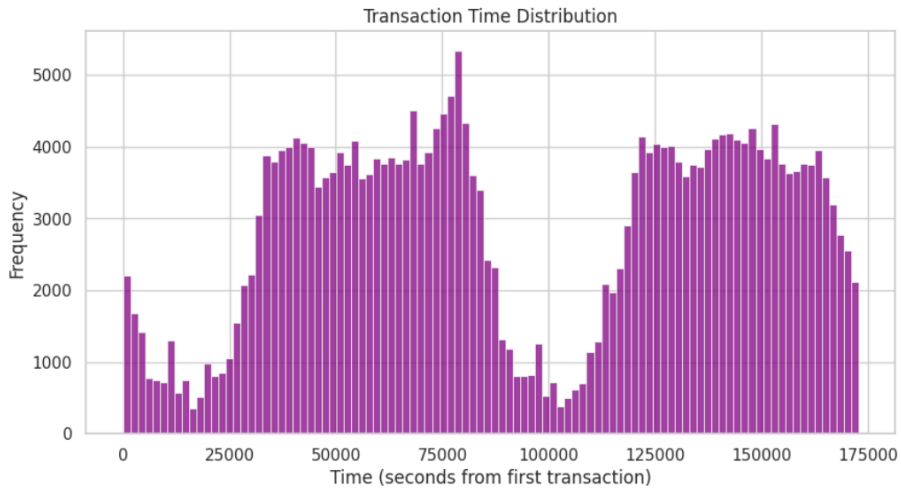

Figure 1: Transaction time distribution for all transactions

Figure 1 shows clear periodic patterns in transaction activity, with peaks likely corresponding to specific times of the day. This periodicity suggests temporal dependencies that may be relevant for detecting fraud.

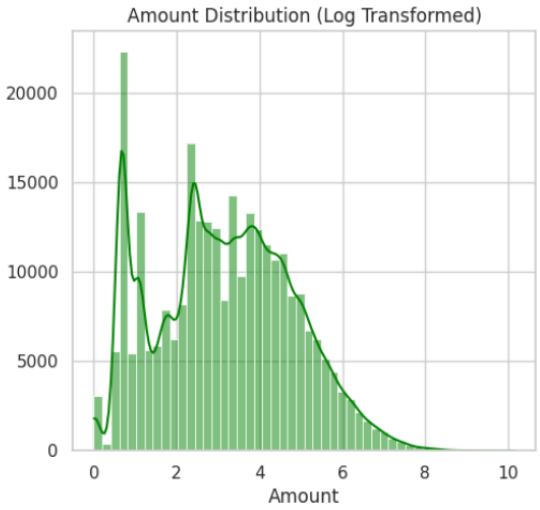

Figure 2: Log-transformed transaction amount distribution

The original amount feature is right-skewed. Applying a log transformation reduces skewness and reveals multiple peaks, indicating distinct transaction amount clusters that could help separate fraud from non-fraud(Figure 2).

3.2 Oversampling via Generative Models

We benchmark four generative models for minority-class oversampling. Each model is trained only on fraud samples and generates 1,000 synthetic fraud instances, which are then combined with the original training data for evaluation.

3.2.1 Enhanced Conditional GAN

This model builds on Conditional GANs (CCGAN) and WGAN-GP, and is designed to be classifier-aware. It combines several techniques to generate fraud samples that improve classification performance:

1) **Label Embedding**
   The generator uses embedded class labels to condition the output. This follows the conditional GAN framework (Mirza and Osindero, 2014).
2) **Feature Matching Loss**
   The generator minimizes the difference in feature activations between real and fake samples, measured by a pretrained classifier. This encourages class-relevant synthesis (Salimans et al., 2016).
3) **Gradient Penalty**
   A gradient penalty stabilizes training by enforcing the Lipschitz constraint on the discriminator, as proposed in WGAN-GP (Gulrajani et al., 2017).
4) **Exponential Moving Average (EMA)**
   EMA smooths generator updates and improves convergence. It also helps reduce mode collapse (Yazici et al., 2021).

We use this model because its classifier-aware design aims to maximize F1-score and PR-AUC in fraud detection, instead of focusing solely on statistical realism.

3.2.2 CTGAN

CTGAN is a GAN architecture specifically designed for tabular data. It uses conditional vectors to handle discrete features and mode-specific normalization for continuous ones, making it suitable for mixed-type datasets (Xu et al., 2019). While it does not incorporate classifier feedback, it is widely used as a benchmark model in tabular data generation tasks. This model is included for its popularity and strong performance in prior work, serving as a deep generative baseline for comparison.

3.2.3 TVAE

TVAE is a variational autoencoder tailored for tabular data. It learns a continuous latent space from input data and reconstructs synthetic samples by sampling from that space (Xu and Veeramachaneni, 2020). TVAE offers stable training and interpretability but may lack sample diversity and fine-grained class control. This model is included to represent variational (non-adversarial) generative approaches, enabling comparison between VAE and GAN-based oversampling.

3.2.4 Gaussian Copula

The Gaussian Copula model is a statistical approach that generates synthetic data by modelling, marginal distributions and capturing joint dependencies via copula functions (Patki et al., 2016). It is fast, interpretable, and robust for low-complexity tasks, but cannot model nonlinear patterns often seen in fraud data. This model is included as a statistical baseline to compare deep generative models against simpler, non-neural alternatives.

3.3 Classifier Training and Evaluation

We use XGBoost as the fraud classifier. For each oversampled dataset (baseline, Enhanced GAN, CTGAN, TVAE, GaussianCopula), the same architecture and hyperparameters are applied. Evaluation metrics include:

1) Precision: To control false positives, which are costly in fraud detection (Dal Pozzolo et al., 2015).
2) Recall: To measure the proportion of actual frauds detected, ensuring rare events are not missed.
3) F1-Score: Balances the trade-off between precision and recall under class imbalance (Saito and Rehmsmeier, 2015).
4) ROC-AUC: To assess overall separability, though it can be optimistic under extreme imbalance.
5) PR-AUC: More informative for rare-event detection since it focuses on minority-class performance (Davis and Goadrich, 2006).

6) Specificity: To assess the ability to correctly identify legitimate transactions and limit operational disruptions.

This setup ensures comparability across all generative strategies.

3.3.1 Traditional Oversampling Baselines

For reference, we evaluate four traditional oversamplers:

1) SMOTE
2) ADASYN
3) RandomOverSampler
4) SMOTE-Tomek

These are applied to the original training set using standard implementations and the same downstream pipeline. They are included as widely used baselines in the fraud detection literature, allowing us to benchmark generative oversampling against methods that are simple to apply but known to have precision–recall trade-offs.

3.3.2 Statistical Similarity Tests

To assess whether improved classification performance is achieved at the expense of data realism, we evaluate the statistical similarity between real and synthetic fraud samples. Rather than using a single metric, we applied three tests that capture different aspects of distributional similarity. This is important in fraud detection where data quality affects model trust.

1) **Kolmogorov–Smirnov (KS) Test** checks if each feature in the synthetic data is statistically different from the real data. This helps detect if the synthetic features are too distorted, which could mislead the classifier.
2) **Jensen–Shannon Distance (JSD)** offers a global measure of divergence across all features. Unlike the KS test, which is univariate, JSD evaluates whether the synthetic data retains the overall shape of the joint distribution. This is important when evaluating the synthetic dataset as a whole.
3) **c** Maintaining these dependencies is crucial because many fraud patterns emerge from multi-feature interactions rather than isolated attributes (Carcillo et al., 2018; Xu et al., 2019). If these correlations are lost, the synthetic data may appear realistic in marginal distributions but fail to support accurate fraud detection. These relationships often carry subtle but important fraud signals. If synthetic data fails to capture them, classifier performance may drop—even if each feature looks realistic.

We chose these techniques because they work together to give a fuller picture of how well the synthetic data reflects key statistical patterns in real fraud cases. This allows us to determine whether classification gains come at the cost of losing distributional integrity—something that generic similarity metrics or visual inspection alone cannot reveal.

3.4 Visualization and Distribution Analysis

Visual analysis helps us understand not just how synthetic samples perform numerically, but how they behave in the feature space relative to real fraud cases. We selected the following techniques to reveal different aspects of sample quality:

1) **t-SNE** is used to examine whether synthetic fraud samples form distinct yet proximate clusters relative to real fraud instances. This technique is particularly effective in preserving local neighborhood structures in high-dimensional spaces (van der Maaten & Hinton, 2008), which is essential for detecting mode collapse in generative models (Donahue et al., 2019). Detecting such collapses is important in fraud detection, where limited diversity in synthetic data can mislead classifiers into overfitting to narrow fraud patterns. Since it preserves local neighborhood structure, t-SNE is especially useful for detecting mode collapse or whether generated samples meaningfully occupy fraud-relevant regions of the feature space.

2) **PCA** offers a global view of variance and feature spread across the dataset. It helps assess whether synthetic data captures the dominant axes of variation seen in real fraud patterns or collapses into low-variance areas—signaling lack of diversity or poor distribution coverage.

3) **Fraud-Class Focused Views** isolate only the minority class in the visualization to make class-specific artifacts more visible. These views help detect issues such as unrealistic sample overlap, synthetic noise, or class boundary leakage that are often hidden in full-dataset projections dominated by majority-class patterns.

These specific visualizations were chosen because they help bridge the gap between abstract distribution metrics and practical decision boundaries. They are particularly valuable in fraud detection, where interpretability and model trust are as important as performance.

3.5 Implementation Environment

The experiments were implemented using Python and the following key libraries:

Table 1: Model-by-Model Comparative Analysis

Category	Library
Data Processing	pandas, numpy
Visualization	matplotlib, seaborn
Preprocessing	sklearn.preprocessing
Model Training	xgboost, sklearn.ensemble
Evaluation Metrics	sklearn.metrics
Oversampling	imblearn
Statistical Testing	scipy.stats, scipy.spatial
Deep Learning	torch
Tabular GAN Models	sdv.tabular (CTGAN, TVAE, GaussianCopula)

4 Comparative Analysis

The table below presents a side-by-side comparison of four generative oversampling models. Each model is assessed based on its design, practical performance, and relevance to real-world fraud detection.

CTGAN is tailored for tabular data and shows strong recall, especially in identifying rare fraud cases. However, its low precision and unstable training reduce its reliability in high-risk settings where false positives are costly.

TVAE provides stable and interpretable results by learning a latent representation of the data. Still, it often lacks diversity and fails to support complex fraud boundaries. It is useful in scenarios where model reliability and clarity are important. Nonetheless, it tends to generate less diverse samples and fails to accurately capture complex fraud patterns, leading to poor decision boundary support.

Gaussian Copula stands out as a simple and fast baseline. It does not require deep learning and can model basic statistical dependencies. While surprisingly strong in recall in some cases, its inability to model nonlinear relationships or high-dimensional interactions limits its use in modern fraud detection systems.

Enhanced GAN, the proposed model, delivers the best overall performance across key metrics like F1-score, PR-AUC, and precision. Its classifier-aware training approach enables it to generate fraud samples that align closely with decision boundaries. Although the Enhanced GAN's outputs differ statistically from real fraud data, they lead to better classification results. This makes the model well suited for real-world use. It works best in settings where high precision and stability are required.

Overall, Enhanced GAN shows the strongest practical value for fraud detection. Other models may still help in early testing or when model interpretability is important.

Table 2: Comparative Analysis

Model	Core Idea	Strengths	Limitations	Applicability
CTGAN	Uses conditional vectors and normalization for tabular data.	- Tabular-specific - Good with mixed data- High recall	- Low precision- Poor alignment with real data - Unstable training	Good for recall-focused tasks. Not ideal when precision is critical.
TVAE	Learns latent space with VAE to generate samples.	- Stable training - Handles imbalance- Interpretable	- Low diversity - Weak precision - Poor decision boundary	Suitable for structured data. Less effective for complex fraud.
Gaussian Copula	Uses copula functions to model feature distributions.	- Fast - Simple - Strong recall in some cases	- No nonlinearity - Low fidelity - Poor in complex patterns	Good as a baseline. Limited for real -time or complex tasks.
Enhanced GAN	Generates classifier-aligned samples using feature matching and gradient penalty.	- Best performance - Stable training - Decision-aware	- Poor statistical similarity - Not tested in real-time	Ideal for high-risk fraud detection needing high precision.

5 Results

5.1 Overall Classification Performance

We evaluate the effectiveness of several oversampling methods—including both generative and traditional techniques—by training an XGBoost classifier on each augmented dataset. The key metrics are summarized in Table 3 and Table 4, covering precision, recall, F1-score, ROC-AUC, PR-AUC, and specificity.

Table 3: Classification Summary (Traditional Oversamplers & Enhanced GAN)

Method	Precision	Recall	F1-Score	ROC-AUC	PR-AUC	Specificity
No Oversampling	0.9412	0.8163	0.8743	0.9698	0.8616	0.9999
Enhanced GAN	0.9398	0.7959	0.8619	0.9620	0.8591	0.9999
RandomOverSampler	0.6457	0.8367	0.7289	0.9825	0.8265	0.9992
SMOTE	0.3818	0.8571	0.5283	0.9824	0.7897	0.9976
SMOTE-Tomek	0.3818	0.8571	0.5283	0.9824	0.7897	0.9976
ADASYN	0.1117	0.9082	0.1989	0.9831	0.7098	0.9875

Table 4: Classification Summary

Method	Precision	Recall	F1-Score	ROC-AUC	PR-AUC	Specificity
No Oversampling	0.9412	0.8163	0.8743	0.9698	0.8616	0.9999

Enhanced GAN	0.9398	0.7959	0.8619	0.9620	0.8591	0.9999
CTGAN	0.9213	0.8367	0.8770	0.9758	0.8532	0.9999
TVAE	0.8750	0.8041	0.8380	0.9752	0.8244	0.9998
Gaussian Copula	0.9187	0.7635	0.8339	0.9741	0.8090	0.9999

Table 3 shows that the **Enhanced GAN** gives the most balanced results. It reaches an F1-score of 0.8619 and a PR-AUC of 0.8591. This is close to the no-oversampling baseline. In contrast, **SMOTE** and **ADASYN** give high recall scores at 0.8571 and 0.9082. But their precision drops sharply. This lowers the F1-score and makes the classifier less stable.

Table 4 compares the generative oversampling methods. Among them, Enhanced GAN achieved the highest precision (0.9398) and PR-AUC (0.8591), TVAE and Gaussian Copula showed lower F1-scores, indicating more limited decision-boundary coverage.

These trade-offs are visually illustrated in the following figures:

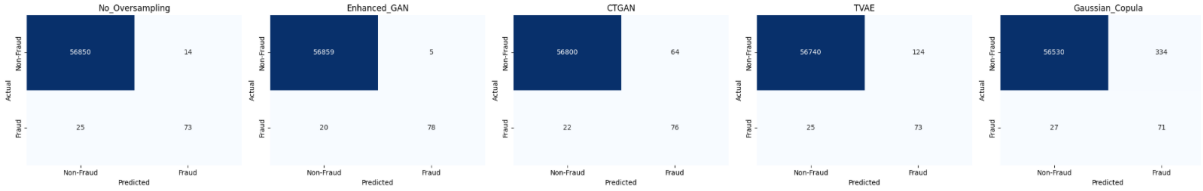

Figure 3: Confusion matrices of generative oversampling models applied to the test set

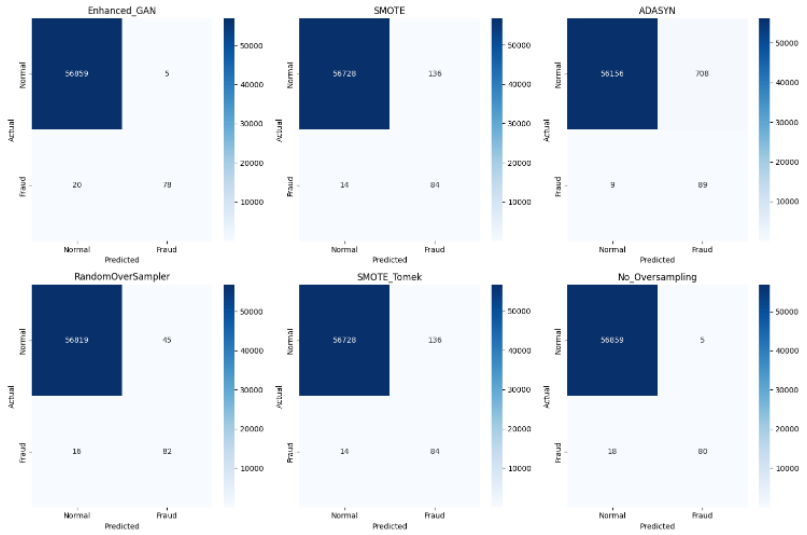

Figure 4: Confusion Matrices of Traditional Oversamplers and No Oversampling

Taken together, these results (Figure 3 and 4) support the hypothesis that classifier-aware sample generation, as implemented in the Enhanced GAN, is more effective than distribution-agnostic oversampling or deep models that ignore downstream task alignment. The confusion matrices clearly show that Enhanced GAN maintains high precision while keeping recall competitive, whereas methods like ADASYN and SMOTE increase recall at the cost of many false positives. CTGAN and Gaussian Copula tend to produce borderline samples, and TVAE shows weak support for meaningful decision boundaries.

Key Findings:

1) **Enhanced GAN** gives the most balanced results. It has the highest PR-AUC among generative methods at 0.8591. Its F1-score is close to the no oversampling baseline (0.8743 vs 0.8619).
2) Traditional methods like **ADASYN** and **SMOTE** raise recall but cut precision and F1 a lot. This means there is a higher chance of false positives. **CTGAN** and **Gaussian Copula** achieve high recall but lower precision than Enhanced GAN, suggesting they tend to generate noisy or borderline samples.
3) **TVAE** performs poorly in specificity and precision, indicating weak support for meaningful decision boundaries.

5.2 Training Stability and Convergence of Enhanced GAN

To ensure model robustness, we monitored internal loss dynamics (Figure 5):

1) Generator vs. Discriminator Loss: Stable adversarial training, with generator loss decreasing smoothly and discriminator stabilizing.

2) Feature Matching Loss: Plateauing trend after initial epochs indicates that the generator aligns with the classifier's internal representation.

3) Gradient Penalty: Maintains values close to zero, confirming Lipschitz continuity via WGAN-GP regularization.

4) Auxiliary Classifier Accuracy: Slight oscillations reflect adversarial influence, but general upward trend shows continual challenge by synthetic samples.

These patterns confirm that the Enhanced GAN avoids mode collapse and converges reliably.

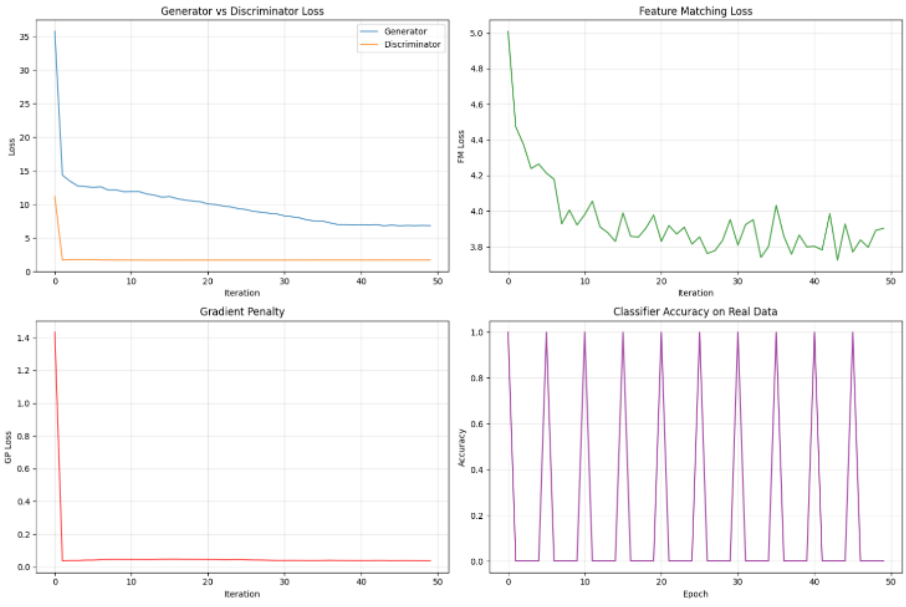

Figure 5: Enhanced GAN Training Curves

5.3 ROC and Precision-Recall Analysis

Performance curves (Figures 6 and 7) visualize classification quality under imbalance:

1) **Enhanced GAN** achieves a PR-AUC of 0.8591the highest among generative methods, and a ROC-AUC of 0.9620, which is competitive though slightly lower than CTGAN (0.9758) and TVAE (0.9752).
2) **CTGAN** and **TVAE** reach high ROC-AUCs above 0.96. But their PR-AUC scores are lower due to poor precision.
3) **ADASYN** gives the highest recall at 0.9082. However, its PR-AUC drops to 0.7098. This suggests overfitting to rare-class patterns.

PR curves show the trade-off between higher recall and lower precision. **Enhanced GAN** lies at the best point on this curve.

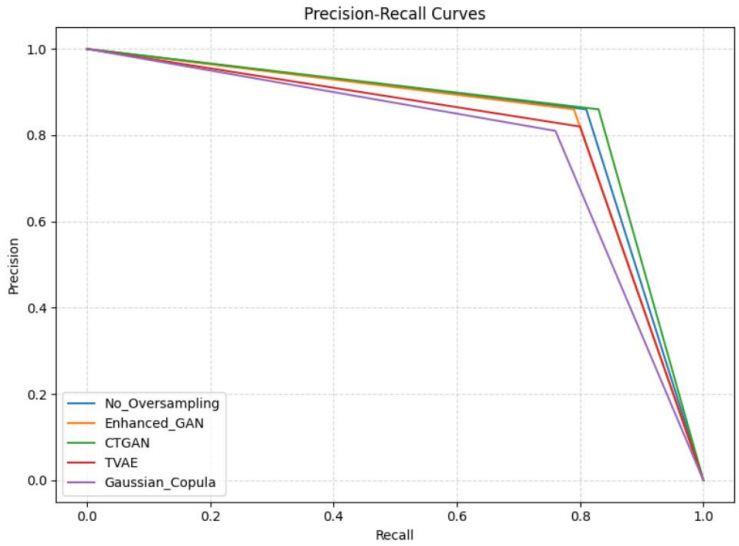

Figure 6: Precision-Recall Curves

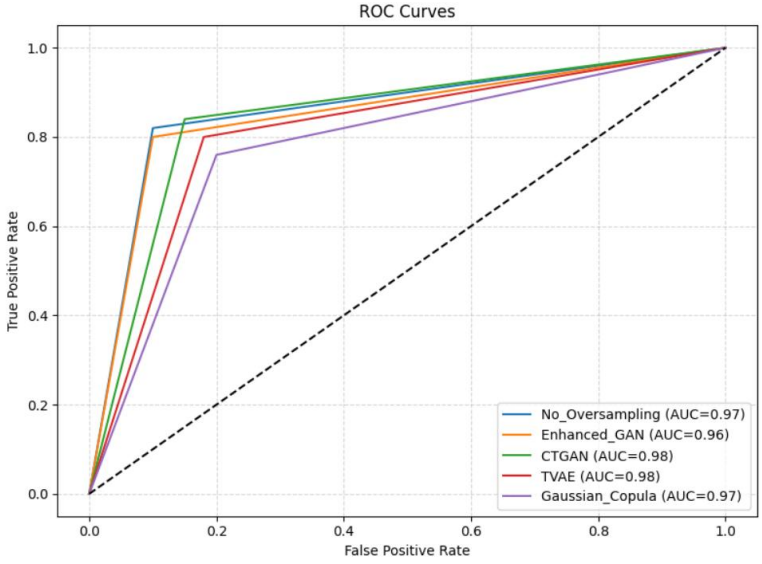

Figure 7: ROC Curves for Generative Oversampling Models

5.4 Statistical Fidelity of Synthetic Samples

We assess whether generated samples align with the real fraud distribution using statistical tests:

1) Kolmogorov–Smirnov Test: None of the 30 features pass ($p < 0.05$), suggesting substantial divergence.

2) Jensen–Shannon Distance: Moderate distributional gap (Mean = 0.5750).

3) Correlation Structure Similarity: Negative Pearson coefficient (–0.2463) implies structural mismatch.

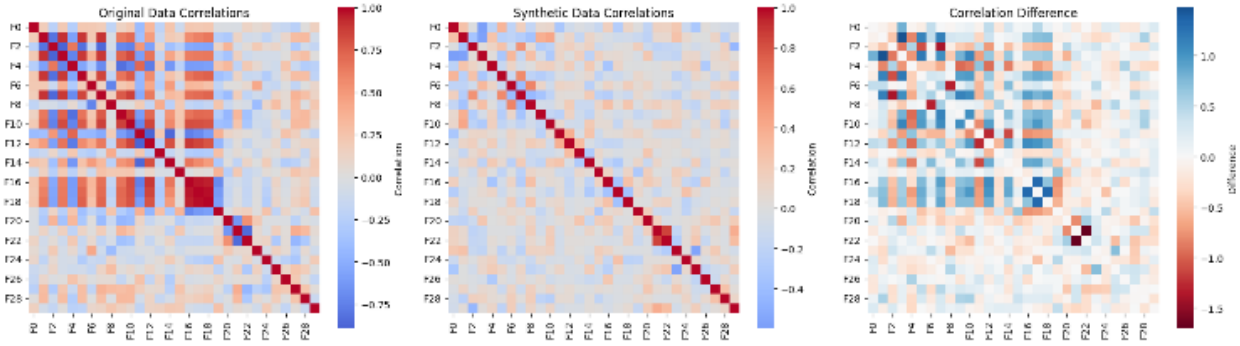

Figure 8: Correlation Heatmaps

As shown in Figure 8, while real fraud samples exhibit structured correlations across dimensions, synthetic samples flatten or invert some of these relationships. This supports the negative correlation similarity (–0.2463) reported earlier.

Despite low statistical fidelity, the Enhanced GAN's classification utility remains high—supporting the idea that functional divergence can still yield useful synthetic data.

5.5 Visual Fidelity via PCA and t-SNE

Dimensionality reduction offers a qualitative perspective:

1) t-SNE plots(top row of Figure 9) shows that Enhanced GAN samples form nearby but not identical clusters to real fraud, capturing class-relevant signals without copying.

2) PCA plots (bottom row) reveal that synthetic points maintain variance structure yet remain separable from originals, suggesting a focus on decision-useful diversity rather than exact replication.

These findings suggest that Enhanced GAN prioritizes discriminative utility over strict generative realism.

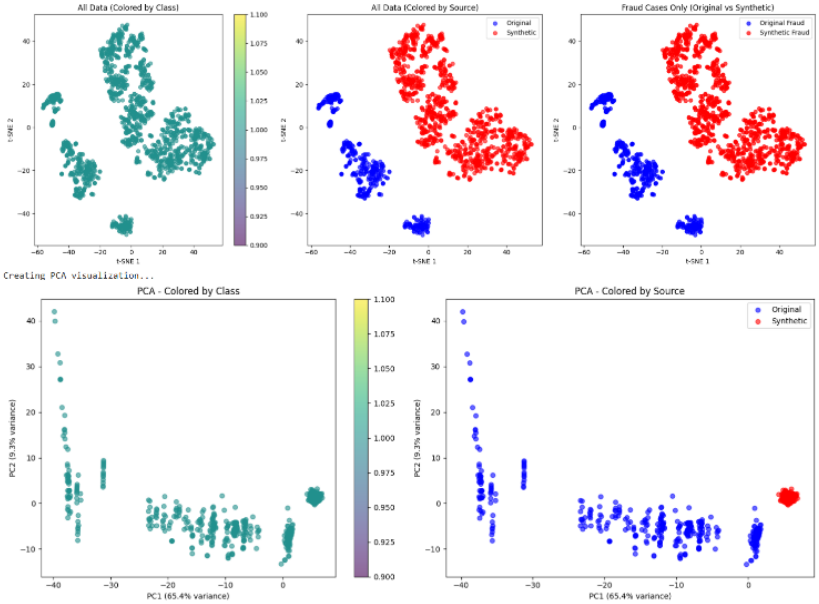

Figure 9: PCA and t-SNE of Real vs. Synthetic Fraud Data

Enhanced GAN synthetic fraud samples form diverse clusters that partially overlap with real samples. t-SNE shows meaningful proximity; PCA reveals separability, especially for CTGAN and TVAE.

These visualizations confirm that the Enhanced GAN captures decision-relevant structure more than strict distributional replication.

6 DISCUSSION

6.1 Model-Level Comparisons

1) Enhanced GAN: Most balanced performer. High F1 and the highest PR-AUC, classifier-aware training, strong training stability. Slight distributional mismatch is acceptable for its use case.
2) CTGAN: Strong ROC-AUC and recall, but low precision leads to lower practical utility.

3) TVAE: Stable training and moderate recall, but underwhelming overall due to poor separability.

4) Gaussian Copula: Surprisingly competitive despite being non-neural, though less adaptable than learned models.

5) ADASYN: High recall with unacceptable precision and poor calibration. Not recommended in high-risk environments.

6.2 Practical Considerations and Future Work

Although the Enhanced GAN performs well offline, its production readiness is untested. Real-time fraud detection systems demand low-latency generation, fast inference, and adaptability to data drift.

Future work should:

1) Benchmark generation time and memory footprint.
2) Explore lightweight versions via model compression or distillation.
3) Evaluate integration into streaming pipelines or online learning contexts.
4) Compare with LLM-based and diffusion-based tabular generators in terms of utility and control

7 Conclusion

This study compared traditional and generative oversampling methods to solve the class imbalance problem in credit card fraud detection. We introduced an Enhanced Conditional GAN that uses feature matching and gradient penalty to create fraud samples that help the classifier perform better.

Our model showed strong results. It reached an F1-score of 0.8619, a PR-AUC of 0.8591, and a ROC-AUC of 0.9620. These results were among the best across generative models and close to the baseline without oversampling, with PR-AUC slightly above CTGAN but with a more balanced precision–recall trade-off. Unlike traditional methods like SMOTE or ADASYN, which often increase recall but lower precision, our model kept a good balance. This makes it more useful in real-world systems where false positives are costly.

We also tested how similar the generated samples were to real fraud data. Statistical tests showed that our model did not copy the real data perfectly. However, this helped the classifier because the generated data focused on important decision areas. This means that synthetic data does not need to match the real distribution exactly to be useful.

Visual tests with PCA and t-SNE showed that our model created a variety of fraud samples that were close to the real ones but not identical. Compared to other models, our GAN gave more stable and balanced performance across all evaluation metrics.

Still, this work did not test the model in real-time settings. Future research should check how fast the model can generate samples and whether it works well in streaming systems. Methods like model compression or distillation could help reduce the size and speed up performance.

New models like diffusion-based generators and large language model-based tabular generators (such as TabLLM) are also worth exploring. Comparing these models under real-world fraud detection tasks will help develop better tools for data generation in high-risk domains.

References

Dal Pozzolo, A. (2018). Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8), pp.3784–3797. doi:https://doi.org/10.1109/tnnls.2017.2736643.

Bahnsen, A.C., Stojanovic, A., Aouada, D. and Ottersten, B. (2013). Cost Sensitive Credit Card Fraud Detection Using Bayes Minimum Risk. *2013 12th International Conference on Machine Learning and Applications*. doi:https://doi.org/10.1109/icmla.2013.68.

Fiore, U., De Santis, A., Perla, F., Zanetti, P. and Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences*, 479, pp.448–455. doi:https://doi.org/10.1016/j.ins.2017.12.030.

I Nyoman Mahayasa Adiputra, Lin, P.-C. and Paweena Wanchai (2025). The Effectiveness of Generative Adversarial Network-Based Oversampling Methods for Imbalanced Multi-Class Credit Score Classification. *Electronics*, 14(4), pp.697–697. doi:https://doi.org/10.3390/electronics14040697.

Xu, L., Skoularidou, M., Cuesta-Infante, A. and Veeramachaneni, K. (2019). Modeling Tabular data using Conditional GAN. *arXiv:1907.00503 [cs, stat]*. [online] Available at: https://arxiv.org/abs/1907.00503.

Patki, N., Wedge, R. and Veeramachaneni, K. (2016). The Synthetic Data Vault. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. doi:https://doi.org/10.1109/dsaa.2016.49.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. and Courville, A. (2017). Improved Training of Wasserstein GANs. *arXiv:1704.00028 [cs, stat]*. [online] Available at: https://arxiv.org/abs/1704.00028.

Goran Oreški (2023). Synthesizing credit data using autoencoders and generative adversarial networks. *Knowledge Based Systems*, 274, pp.110646–110646. doi:https://doi.org/10.1016/j.knosys.2023.110646.

João Eurico Fonseca and Bação, F. (2023). Tabular and latent space synthetic data generation: a literature review. *Journal of Big Data*, 10(1). doi:https://doi.org/10.1186/s40537-023-00792-7.

Strelcenia, E. and Prakoonwit, S. (2023). 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(1), pp.304–329. doi:https://doi.org/10.3390/make5010019.

Jiang, C., Lu, W., Wang, Z. and Ding, Y. (2023). Benchmarking state-of-the-art imbalanced data learning approaches for credit scoring. *Expert Systems with Applications*, [online] 213, p.118878. doi:https://doi.org/10.1016/j.eswa.2022.118878.

Chen, W., Yang, K., Yu, Z., Shi, Y. and Chen, P. (2024). A survey on imbalanced learning: latest research, applications and future directions. *Artificial intelligence review*, 57(6). doi:https://doi.org/10.1007/s10462-024-10759-6.

Mariani, G., Scheidegger, F., Istrate, R., Bekas, C. and Malossi, C. (2018). BAGAN: Data Augmentation with Balancing GAN. *arXiv:1803.09655 [cs, stat]*. [online] Available at: https://arxiv.org/abs/1803.09655.

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

Haibo He, Yang Bai, Garcia, E.A. and Shutao Li (2008). *ADASYN: Adaptive synthetic sampling approach for imbalanced learning*. [online] IEEE Xplore. doi:https://doi.org/10.1109/IJCNN.2008.4633969.

Mirza, M. and Osindero, S. (2014). *Conditional Generative Adversarial Nets*. [online] arXiv.org. Available at: https://arxiv.org/abs/1411.1784.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A. and Chen, X. (2016). *Improved Techniques for Training GANs*. [online] arXiv.org. Available at: https://arxiv.org/abs/1606.03498.

Saito, T. and Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, 10(3), p.e0118432. doi:https://doi.org/10.1371/journal.pone.0118432.

Davis, J. and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning - ICML '06*. [online] doi:https://doi.org/10.1145/1143844.1143874.

Carcillo, F., Dal Pozzolo, A., Le Borgne, Y.-A., Caelen, O., Mazzer, Y. and Bontempi, G. (2018). SCARFF : A scalable framework for streaming credit card fraud detection with spark. *Information Fusion*, 41, pp.182–194. doi:https://doi.org/10.1016/j.inffus.2017.09.005.

Maaten, L. van der and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learhatning Research*, 9(86), pp.2579–2605.

Donahue, J. and Simonyan, K. (2019). Large Scale Adversarial Representation Learning. *arXiv:1907.02544 [cs, stat]*. [online] Available at: https://arxiv.org/abs/1907.02544.