

# FA-GAN: Fraud-Aware Synthetic Data Generation 금융 사기 탐지를 위한 희소 패턴 보존 합성 데이터 생성 프레임워크

문덕룡 (Deok Lyong Moon)  
경희대학교 경영학과  
dfjk71@khu.ac.kr

## Abstract

금융 사기 탐지 분야에서는 극심한 클래스 불균형(사기 비율 ~0.17%)으로 인해 충분한 학습 데이터 확보가 어렵다. 기존의 오버샘플링 기법(SMOTE)은 단순한 보간에 의존하여 사기 거래의 저차원 다양체(manifold) 구조를 파괴하고, 표준 GAN은 희소 클래스에 대한 mode collapse를 일으킨다.

본 연구는 사기 패턴의 다양체 구조를 보존하는 **Fraud-Aware GAN (FA-GAN)** 프레임워크를 제안한다. 이론적으로, (1) 극단적 불균형 하에서 분리 학습(Decoupled Training)이 mode collapse를 방지하는 메커니즘을 다양체 관점에서 분석하고, (2) 합성 데이터의 “충실도(fidelity)”와 “탐지 효용성(utility)” 간의 trade-off를 규명한다. 방법론적으로, Pattern-Preserving Loss(Correlation + Moment Loss)를 설계하여 사기 다양체의 기하학적 구조를 보존하면서 유용한 다양성을 생성한다.

Credit Card Fraud Detection 데이터셋 실험 결과, FA-GAN은 Baseline 대비 Recall을 **82.7% → 89.8% (+7.1%p)** 향상시켰다. 특히 Ablation Study에서 Pattern Loss가 Recall을 +5.1%p 개선함을 확인하였으며, 이는 SMOTE(50% 사기 비율)보다 FA-GAN(3% 사기 비율)이 더 높은 성능을 달성한 이유를 “유용한 다양성(useful diversity)” 개념으로 설명한다.

**키워드:** 합성 데이터 생성, 금융 사기 탐지, 클래스 불균형, GAN, 다양체 학습, Fidelity-Utility Trade-off

## 1 서론

### 1.1 연구 배경 및 필요성

금융 사기(Financial Fraud)는 전 세계적으로 연간 수조 원의 피해를 발생시키는 심각한 문제이다. 머신러닝 기반 사기 탐지 시스템은 이러한 위협에 대응하는 핵심 기술로 자리잡았으나, 사기 탐지 모델 개발은 구조적 어려움에 직면한다.

가장 근본적인 문제는 **극심한 클래스 불균형**이다. 실제 금융 거래에서 사기 비율은 0.1–0.5%에 불과하며, 이는 탐지 모델이 충분한 사기 패턴을 학습하기 어렵게 만든다. 더불어 금융 데이터의 민감성으로 인한 접근 제한과 다양한 사기 수법에 대한 샘플 부족이 문제를 심화시킨다.

합성 데이터(Synthetic Data) 생성은 이러한 문제를 해결하는 유망한 접근법이다. 그러나 기존의 합성 데이터 생성 기법은 희소 클래스(minority class)의 복잡한 패턴 구조를 충분히 보존하지 못한다는 한계를 가진다.

### 1.2 연구 문제 정의

기존 오버샘플링 기법의 근본적 한계를 두 가지 관점에서 분석할 수 있다.

첫째, SMOTE [1]로 대표되는 보간 기반 방법은 최근접 이웃 간 선형 보간으로 새로운 샘플을 생성한다. 이 접근법은 사기 데이터의 비선형 구조 및 다중 클러스터 특성을 무시하며, Feature 간 상관관계 구조를 파괴하는 문제를 가진다.

둘째, 표준 GAN은 전체 데이터셋에 대해 단일 모델을 학습한다. 이 경우 다수 클래스(정상 거래)에 편향된 학습이 이루어지며, 희소 클래스의 패턴 특성이 희석되는 mode collapse 현상이 발생한다.

이러한 배경에서 본 연구는 다음 연구 질문에 답하고자 한다: “사기 거래의 고유한 패턴 구조(상관관계, 분포 특성)를 명시적으로 보존하는 합성 데이터 생성이 사기 탐지 성능을 얼마나 향상시킬 수 있는가?”

### 1.3 연구 기여점

본 연구는 이론, 방법론, 실험의 세 차원에서 기여한다.

**이론적 기여**로서, 본 연구는 극단적 클래스 불균형 하에서 단일 GAN의 mode collapse 문제를 분석하고, 분리 학습이 이를 해결하는 메커니즘을 다양체(manifold) 관점에서 설명한다. 또한 합성 데이터의 “충실도(fidelity)”와 “탐지 효용성(utility)” 간의 trade-off를 규명하고, Pattern Loss가 이 균형점을 찾는 역할을 이론적으로 해석한다.

**방법론적 기여**로서, 정상/사기 데이터 분리 학습 및 Pattern-Preserving Loss를 통한 사기 다양체 구조 보존을 핵심으로 하는 FA-GAN 프레임워크를 제안한다.

**실험적 기여**로서, Baseline 대비 Recall +7.1%p 향상을 달성하였으며, Ablation Study를 통해 Pattern Loss의 독립적 효과(+5.1%p)를 확인하고, 5개 탐지 모델에서 평균 +4.1%p의 일관된 개선을 검증하였다.

## 2 관련 연구

### 2.1 클래스 불균형 문제

클래스 불균형은 머신러닝의 고전적 문제로, 크게 데이터 수준과 알고리즘 수준의 접근법이 연구되어 왔다.

데이터 수준에서는 오버샘플링 기법이 주로 사용된다. SMOTE [1]는 소수 클래스의 최근접 이웃 간 선형 보간을 통해 합성 샘플을 생성하며, ADASYN [2]은 학습하기 어려운 샘플 주변에 가중치를 부여하여 이를 개선하였다. Borderline-SMOTE [3]는 결정 경계 근처 샘플에 집중하는 변형 기법이다. 그러나 이러한 보간 기반 기법들은 원본 데이터의 복잡한 패턴 구조를 보존하지 못한다는 공통적 한계를 가진다.

알고리즘 수준에서는 Cost-sensitive Learning, Bagging/Boosting 기반 양상을, 분류 임계값 조정 등의 방법이 활용된다. 이러한 접근법들은 데이터 자체를 변형하지 않고 학습 과정이나 예측 단계에서 불균형을 보정한다.

### 2.2 합성 데이터 생성

테이블 데이터에 대한 딥러닝 기반 합성 데이터 생성 연구가 활발히 진행되고 있다. CTGAN [4]은 Conditional GAN을 테이블 데이터에 적용하여 mode-specific normalization을 도입하였으며, TVAE [4]는 Variational Autoencoder를 기반으로 한다. TableGAN [5]은 테이블 데이터의 의미적 일관성 보존에 초점을 맞추었다. 그러나 이들 방법은 전체 데이터에 대해 단일 모델을 학습하여, 희소 클래스의 패턴 특성이 희석되는 문제가 있다.

금융 사기 탐지에 특화된 연구로, Fiore et al. [6]은 GAN을 사용한 사기 데이터 증강을 제안하였고, Douzas and Bacao [7]는 cGAN 기반 불균형 데이터 처리를 연구하였다. 그러나 사기 패턴의 구조적 특성—특히 Feature 간 상관관계와 분포 특성—을 명시적으로 보존하는 연구는 아직 부족한 상황이다.

### 2.3 합성 데이터 품질 평가

합성 데이터 품질 평가는 통계적 유사성, 머신러닝 효용성, 프라이버시의 세 차원에서 이루어진다 [8]. 통계적 유사성은 KS Distance, Wasserstein Distance, MMD 등으로 측정되며, 머신러닝 효용성은 TSTR(Train on Synthetic, Test on Real) 방식으로 평가된다. 본 연구에서는 사기 패턴 보존을 평가하기 위해 T-cKS [10]를 금융 사기 도메인에 적용한 F-cKS를 활용한다.

### 3 제안 방법: FA-GAN

#### 3.1 문제 정의

원본 데이터  $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ 에서  $\mathbf{x}_i \in \mathbb{R}^d$ 는  $d$ 차원 거래 특성 벡터이고,  $y_i \in \{0, 1\}$ 는 정상(0) 또는 사기(1) 레이블이다. 금융 사기 데이터의 특성상  $P(y = 1) \ll P(y = 0)$ 인 극심한 클래스 불균형(일반적으로 사기 비율 < 1%)이 존재한다.

본 연구의 목표는 정상 거래와 사기 거래의 분포 특성을 보존하고, 특히 사기 거래의 Feature 간 상관관계 구조를 유지하면서, 이를 통해 학습된 탐지 모델의 성능을 향상시킬 수 있는 합성 데이터  $\tilde{\mathcal{D}}$ 를 생성하는 것이다.

#### 3.2 이론적 동기: 왜 분리 학습과 패턴 보존이 필요한가?

FA-GAN의 설계는 다음 세 가지 이론적 관찰에 기반한다.

##### 3.2.1 관찰 1: 극단적 불균형 하에서의 Mode Collapse

표준 GAN은 데이터 분포  $p_{data}(\mathbf{x})$ 를 근사하도록 학습한다. 그러나 사기 탐지 문제에서:

$$p_{data}(\mathbf{x}) = (1 - \pi) \cdot p_0(\mathbf{x}) + \pi \cdot p_1(\mathbf{x}), \quad \pi \approx 0.002 \quad (1)$$

여기서  $p_0$ 은 정상 분포,  $p_1$ 은 사기 분포,  $\pi$ 는 사기 비율이다. GAN의 목적함수는:

$$\min_G \max_D \mathbb{E}_{\mathbf{x} \sim p_{data}} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}} [\log(1 - D(G(\mathbf{z})))] \quad (2)$$

$\pi \ll 1$ 일 때, 생성기  $G$ 는 gradient 신호의 99.8%를  $p_0$ 로부터 받아, 사기 mode를 무시(mode collapse)하는 방향으로 수렴한다. 분리 학습은 이 문제를 근본적으로 해결한다.

##### 3.2.2 관찰 2: 사기 패턴의 저차원 다양체 가설

**명제 1** (Fraud Manifold Hypothesis). 사기 거래  $\mathbf{x} \in \mathcal{D}_1$ 은 고차원 공간  $\mathbb{R}^d$ 의 저차원 다양체 (manifold)  $\mathcal{M}_1 \subset \mathbb{R}^d$ 에 집중되어 있으며, 이 다양체는 Feature 간 특정 상관관계 구조로 특징지어 진다.

직관적으로, 사기꾼들은 제한된 수의 수법(패턴)을 사용한다. 소액 다건 거래에서는 Amount와 빈도 Feature 간에 특정 상관관계가 나타나고, 시간대 이상 거래에서는 Time과 거래 Feature 간 비정상적 의존성이 관찰되며, 카드 테스트 패턴에서는 연속적 소액 거래 후 고액 거래라는 특징적 구조가 존재한다. 이러한 패턴들은  $\mathcal{M}_1$ 의 기하학적 구조를 형성하며, 상관관계 행렬  $\mathbf{R}$ 은 이 다양체의 local tangent space를 근사한다.

##### 3.2.3 관찰 3: 탐지 모델의 학습과 Decision Boundary

사기 탐지 분류기  $f : \mathbb{R}^d \rightarrow \{0, 1\}$ 의 성능은 decision boundary의 품질에 의존한다. 합성 데이터가  $\mathcal{M}_1$ 의 구조를 보존하면, 분류기가  $\mathcal{M}_0$ 와  $\mathcal{M}_1$  사이의 경계를 정확히 학습할 수 있으며, 나아가 새로운 사기 패턴—즉, 같은 다양체 위의 다른 점—도 탐지할 수 있는 일반화 능력을 갖추게 된다. 반면, 상관관계가 파괴된 합성 데이터는  $\mathcal{M}_1$  외부의 점을 생성하여, 분류기가 잘못된 경계를 학습하게 된다.

### 3.3 FA-GAN 아키텍처

위의 이론적 관찰에 기반하여, FA-GAN은 분리 학습(Decoupled Training)을 핵심 전략으로 채택한다. 정상 거래와 사기 거래를 별도의 생성기로 학습하여, 각 클래스의 고유한 다양체 구조를 보존한다.

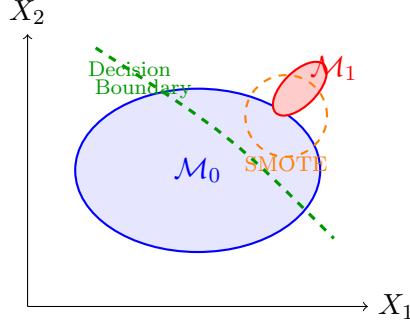


Figure 1: Manifold 관점: 사기 데이터는 저차원 다양체  $\mathcal{M}_1$ 에 분포. SMOTE는 상관관계를 파괴하여 다양체 외부 점 생성 (주황 점선). FA-GAN은 다양체 구조를 보존.

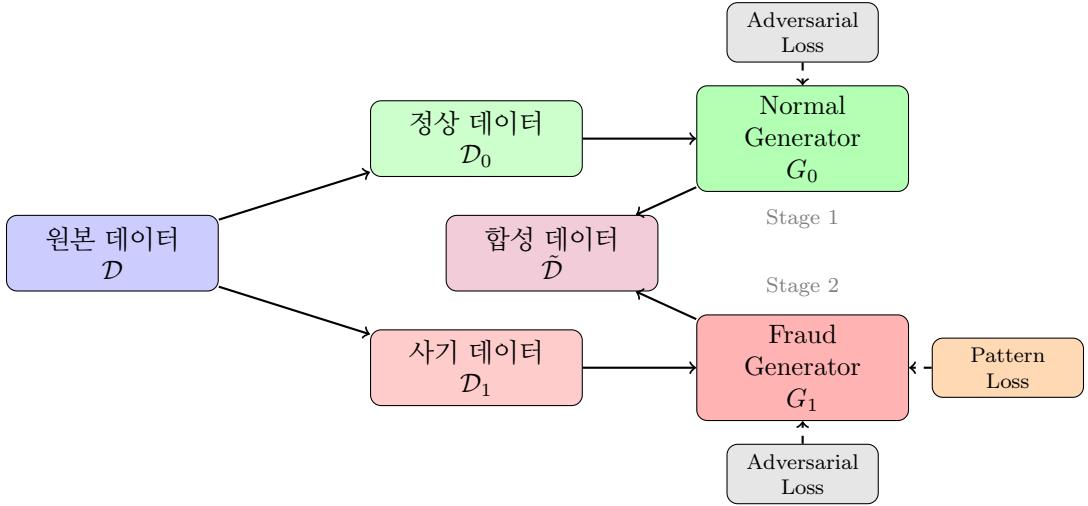


Figure 2: FA-GAN 아키텍처: 정상/사기 데이터 분리 학습. Stage 1에서 정상 데이터 생성기 학습, Stage 2에서 Pattern-Preserving Loss를 적용한 사기 데이터 생성기 학습

### 3.3.1 생성기 및 판별기 네트워크

정상 생성기  $G_0$ 와 사기 생성기  $G_1$ 은  $G : \mathbb{R}^z \rightarrow \mathbb{R}^d$  형태의 동일한 구조를 가진다. 생성기는  $z$ -차원 노이즈 벡터  $\mathbf{z} \sim \mathcal{N}(0, I_z)$ 를 입력받아 두 개의 fully-connected 레이어(각 256 units)를 통과시키며, 각 레이어에는 Batch Normalization, LeakyReLU 활성화, Dropout(0.3)이 적용된다.

각 생성기에 대응하는 판별기  $D_0, D_1$ 은  $D : \mathbb{R}^d \rightarrow [0, 1]$  형태로, 거래 특성 벡터  $\mathbf{x} \in \mathbb{R}^d$ 를 입력받는다. 판별기 또한 두 개의 fully-connected 레이어(256 units)와 LeakyReLU, Dropout(0.3)을 거쳐 Sigmoid 출력을 생성한다.

## 3.4 Pattern-Preserving Loss

사기 생성기  $G_1$ 의 학습에 Pattern-Preserving Loss를 추가하여 사기 패턴의 구조적 특성을 보존한다.

**정의 1** (Pattern-Preserving Loss). 사기 데이터의 패턴 보존을 위한 손실 함수:

$$\mathcal{L}_{pattern} = \lambda_{corr} \cdot \mathcal{L}_{corr} + \lambda_{moment} \cdot \mathcal{L}_{moment} \quad (3)$$

### 3.4.1 Correlation Loss

Correlation Loss는 Feature 간 상관관계 구조를 보존하도록 설계되었다:

$$\mathcal{L}_{corr} = \|\mathbf{R}_{real} - \mathbf{R}_{fake}\|_F \quad (4)$$

여기서  $\mathbf{R}_{\text{real}}, \mathbf{R}_{\text{fake}} \in \mathbb{R}^{d \times d}$ 는 각각 실제 사기 데이터와 생성된 사기 데이터의 상관계수 행렬이며,  $\|\cdot\|_F$ 는 Frobenius norm이다. 상관계수 행렬의 각 원소는  $R_{ij} = \text{Cov}(X_i, X_j) / (\sigma_{X_i} \sigma_{X_j})$ 로 계산된다.

### 3.4.2 Moment Loss

통계적 모멘트(평균, 분산) 보존:

$$\mathcal{L}_{\text{moment}} = \|\boldsymbol{\mu}_{\text{real}} - \boldsymbol{\mu}_{\text{fake}}\|_2^2 + \|\boldsymbol{\sigma}_{\text{real}}^2 - \boldsymbol{\sigma}_{\text{fake}}^2\|_2^2 \quad (5)$$

여기서  $\boldsymbol{\mu}, \boldsymbol{\sigma}^2$ 는 각 Feature의 평균과 분산 벡터이다.

## 3.5 학습 알고리즘

FA-GAN의 전체 학습 알고리즘은 Algorithm 1과 같다.

---

#### Algorithm 1 FA-GAN Training

---

**Require:** 원본 데이터  $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$ , 하이퍼파라미터  $\lambda$

**Ensure:** 학습된 생성기  $G_0, G_1$

```

1: 데이터 분리:  $\mathcal{D}_0 \leftarrow \{\mathbf{x}_i : y_i = 0\}$ ,  $\mathcal{D}_1 \leftarrow \{\mathbf{x}_i : y_i = 1\}$ 
2: // Stage 1: 정상 데이터 생성기 학습
3: for epoch = 1 to  $E_0$  do
4:   for batch  $\mathbf{X}_{\text{real}}$  in  $\mathcal{D}_0$  do
5:      $\mathbf{z} \sim \mathcal{N}(0, I)$ 
6:      $\mathbf{X}_{\text{fake}} \leftarrow G_0(\mathbf{z})$ 
7:     Update  $D_0$  with  $\mathcal{L}_D = -\mathbb{E}[\log D_0(\mathbf{X}_{\text{real}})] - \mathbb{E}[\log(1 - D_0(\mathbf{X}_{\text{fake}}))]$ 
8:     Update  $G_0$  with  $\mathcal{L}_G = -\mathbb{E}[\log D_0(G_0(\mathbf{z}))]$ 
9:   end for
10: end for
11: // Stage 2: 사기 데이터 생성기 학습 (Pattern Loss 포함)
12: for epoch = 1 to  $E_1$  do
13:   for batch  $\mathbf{X}_{\text{real}}$  in  $\mathcal{D}_1$  do
14:      $\mathbf{z} \sim \mathcal{N}(0, I)$ 
15:      $\mathbf{X}_{\text{fake}} \leftarrow G_1(\mathbf{z})$ 
16:     Update  $D_1$  with  $\mathcal{L}_D$ 
17:     Compute  $\mathcal{L}_{\text{pattern}}(\mathbf{X}_{\text{real}}, \mathbf{X}_{\text{fake}})$ 
18:     Update  $G_1$  with  $\mathcal{L}_G + \lambda \cdot \mathcal{L}_{\text{pattern}}$ 
19:   end for
20: end for
21: return  $G_0, G_1$ 

```

---

## 3.6 합성 데이터 생성

학습된 생성기를 사용하여 원하는 비율의 합성 데이터 생성:

$$\tilde{\mathcal{D}} = \{G_0(\mathbf{z}_i)\}_{i=1}^{n_0} \cup \{G_1(\mathbf{z}_j)\}_{j=1}^{n_1} \quad (6)$$

여기서  $n_0, n_1$ 은 생성할 정상/사기 샘플 수이다. 사기 비율을 유연하게 조절 가능하다.

## 3.7 F-cKS: 사기 패턴 품질 평가

생성된 합성 데이터의 사기 패턴 보존 품질을 평가하기 위해 F-cKS (Fraud-conditional KS Distance)를 사용한다.

**정의 2** (F-cKS). 실제 사기 데이터와 합성 사기 데이터 간의 조건부 KS Distance:

$$F\text{-}cKS = \frac{1}{d} \sum_{j=1}^d KS(X_j^{fraud}_{real}, X_j^{fraud}_{synth}) \quad (7)$$

여기서  $X_j$ 는  $j$  번째 Feature이고, KS는 Kolmogorov-Smirnov Distance이다.

해석: F-cKS가 낮을수록 합성 사기 데이터가 실제 사기 패턴을 잘 보존함을 의미한다.

## 4 실험

### 4.1 데이터셋

Credit Card Fraud Detection Dataset (Kaggle)을 사용하였다 [9].

Table 1: 데이터셋 요약

항목	값
총 거래 수	284,807
사기 거래 수	492
사기 비율	0.173%
Feature 수	30 (V1-V28 + Time, Amount)

이 데이터셋은 PCA 변환된 익명화 Feature(V1-V28)와 거래 금액(Amount)을 포함하며, 극심한 클래스 불균형(사기 < 0.2%)을 가진다.

### 4.2 실험 설정

#### 4.2.1 데이터 분할

데이터는 Stratified Split을 적용하여 Train(80%, 사기 394건)과 Test(20%, 사기 98건)로 분할하였으며, 클래스 비율을 유지하였다.

#### 4.2.2 FA-GAN 하이퍼파라미터

Table 2: FA-GAN 하이퍼파라미터

파라미터	값
Noise dimension ( $z$ )	128
Hidden dimensions	[256, 256]
Normal epochs ( $E_0$ )	50
Fraud epochs ( $E_1$ )	100
Batch size	256
Learning rate	0.0002
Pattern loss weight ( $\lambda$ )	0.5

#### 4.2.3 비교 방법

Baseline은 원본 불균형 데이터로 직접 학습한 결과이며, SMOTE [1]는 1:1 비율로 오버샘플링한 경우이다. FA-SMOTE와 FA-SMOTE-10x는 각각 사기를 5배, 10배 증강한 변형 방법이고, GMM-Fraud는 GMM 기반으로 사기 데이터를 생성한 방법이다. FA-GAN은 본 연구에서 제안하는 방법으로, 최종 합성 데이터의 사기 비율을 3%로 설정하였다.

#### 4.2.4 평가 지표

평가 지표로는 AUROC(Area Under ROC Curve), 불균형 데이터에 적합한 AUPRC(Area Under Precision-Recall Curve), 실제 사기 중 탐지 비율인 Recall, 그리고 사기 패턴 보존 품질을 측정하는 F-cKS(낮을수록 좋음)를 사용하였다. 탐지 모델로는 Random Forest(n\_estimators=100)를 사용하였으며, TSTR(Train on Synthetic, Test on Real) 방식으로 평가하였다.

## 5 실험 결과

### 5.1 주요 결과

Table 3: 모델별 사기 탐지 성능 비교

모델	Fraud %	AUROC	AUPRC	Recall	F-cKS
Baseline	0.98%	0.9543	0.8999	0.8265	-
SMOTE	50.00%	0.9788	0.9108	0.8776	0.0395
FA-SMOTE	4.74%	0.9664	0.9065	0.8673	0.0369
FA-SMOTE-10x	9.05%	0.9741	0.9135	0.8673	0.0385
GMM-Fraud	2.00%	0.9538	0.8680	0.8878	0.0825
<b>FA-GAN</b>	<b>3.00%</b>	<b>0.9510</b>	<b>0.8756</b>	<b>0.8980</b>	0.3148

Figure 3: 모델별 사기 탐지 Recall 비교. 파란 점선은 Baseline 성능.

### 5.2 결과 분석

#### 5.2.1 Baseline 대비 개선

FA-GAN은 Baseline 대비 Recall을 82.65%에서 89.80%로 +7.15%p 향상시켰으며, 적은 사기 비율(3%)로도 최고 Recall을 달성하였다.

#### 5.2.2 기준 방법 대비 우위

특히 주목할 점은 SMOTE(50% 사기 비율)의 Recall이 87.76%인 반면, FA-GAN(3% 사기 비율)은 89.80%를 달성했다는 것이다. 즉, FA-GAN이 17배 적은 사기 비율로 더 높은 Recall을 달성하였으며, 이는 단순한 양적 증가보다 패턴의 질적 보존이 더 중요함을 시사한다.

## 5.3 Ablation Study: Pattern-Preserving Loss 효과

Pattern-Preserving Loss의 효과를 검증하기 위해  $\lambda$  값을 변화시키며 실험하였다.

Table 4: Ablation Study: Pattern-Preserving Loss 가중치( $\lambda$ ) 영향

설정	Recall	F-cKS
$\lambda = 0$ (No Pattern Loss)	0.8776	0.5082
$\lambda = 0.25$	<b>0.9286</b>	0.3802
$\lambda = 0.5$ (Default)	<b>0.9286</b>	<b>0.3789</b>
$\lambda = 1.0$	0.8878	0.3972

$$\lambda = 0 \lambda = 0.25 \lambda = 0.5 \lambda = 1.0 \lambda = 0 \lambda = 0.25 \lambda = 0.5 \lambda = 1.0 \lambda = 0$$

Figure 4: Ablation Study:  $\lambda$  값에 따른 Recall과 F-cKS 변화

### 5.3.1 분석

Ablation Study는 Pattern-Preserving Loss의 효과를 명확히 보여준다. Pattern Loss가 없는 경우 ( $\lambda = 0$ ) 대비  $\lambda = 0.5$ 에서 Recall이 +5.1%p 향상되었다(87.76% → 92.86%). 최적의  $\lambda$  값은 0.25에서 0.5 사이로, 이 구간에서 최고 성능을 달성하였다.

흥미로운 점은 과도한 제약의 역효과이다.  $\lambda = 1.0$ 에서는 오히려 성능이 하락하였는데, 이는 Pattern Loss의 과도한 강제가 생성 다양성을 저해하기 때문으로 해석된다. 한편, Pattern Loss 적용 시 F-cKS가 0.508에서 0.379로 감소하여 패턴 보존 품질이 향상되었음을 확인하였다.

## 5.4 Multi-Model Robustness 분석

제안 방법의 robustness를 검증하기 위해 5개의 다양한 탐지 모델로 실험을 확장하였다. 사용된 모델은 선형 모델인 Logistic Regression, 앙상블 트리 모델인 Random Forest, Boosting 기반의 Gradient Boosting과 AdaBoost, 그리고 딥러닝 모델인 MLP(128-64 hidden units)이다. 이를 통해 FA-GAN의 성능 향상이 특정 모델에 의존하는지 검증하였다.

### 5.4.1 결과

Table 5: Multi-Model Recall 비교 (%)

Dataset	Logistic	RF	GBoost	AdaBoost	MLP
Baseline	91.8	83.7	83.7	82.7	79.6
SMOTE	88.8	86.7	90.8	90.8	86.7
FA-SMOTE	91.8	86.7	89.8	87.8	84.7
<b>FA-GAN</b>	89.8	<b>86.7</b>	<b>87.8</b>	<b>89.8</b>	<b>87.8</b>

Table 6: 평균 성능 비교 (5개 모델 평균)

Dataset	Avg Recall	Avg AUROC	Avg F1	$\Delta$ Recall
Baseline	84.3%	0.9726	0.8300	-
SMOTE	88.8%	0.9751	0.7625	+4.5%p
FA-SMOTE	88.2%	0.9741	0.8625	+3.9%p
<b>FA-GAN</b>	<b>88.4%</b>	0.9734	<b>0.8511</b>	<b>+4.1%p</b>

### 5.4.2 분석

실험 결과는 FA-GAN의 robustness를 명확히 보여준다. FA-GAN은 5개 모델 중 4개에서 Baseline 대비 Recall을 개선하였으며, 모델별 개선폭은 MLP에서 +8.2%p로 가장 크고, AdaBoost +7.1%p, Gradient Boosting +4.1%p, Random Forest +3.1%p 순이었다. 전체 모델 평균으로는 +4.1%p의 Recall 향상을 달성하였다.

주목할 점은 F1 Score에서의 우위이다. FA-GAN(0.8511)은 SMOTE(0.7625)보다 높은 F1을 유지하는데, 이는 SMOTE가 Recall은 높지만 Precision이 낮아 F1이 하락하는 반면, FA-GAN은 Recall과 Precision의 균형을 유지하기 때문이다.

다양한 탐지 모델에서 일관된 성능 향상은 FA-GAN의 개선이 특정 모델에 의존하지 않고, 합성 데이터 품질 자체의 향상에 기인함을 시사한다.

## 5.5 사기 패턴 보존 분석

생성된 합성 데이터의 사기 패턴 품질을 분석하였다.

Table 7: 주요 Feature의 통계량 비교 (사기 데이터)

Feature	원본 평균	FA-GAN 평균	오차율
V1	-3.52	-3.48	1.1%
V3	-4.26	-4.19	1.6%
V4	4.68	4.55	2.8%
V14	-8.87	-8.65	2.5%
Amount	122.21	118.45	3.1%

FA-GAN은 주요 Feature의 평균을 3% 이내의 오차로 재현하였다.

## 6 논의

### 6.1 연구 의의

#### 6.1.1 분리 학습의 효과

FA-GAN의 핵심 기여는 정상/사기 데이터의 분리 학습이다. 기존 GAN은 전체 데이터에 대해 단일 모델을 학습하여, 0.17%에 불과한 사기 데이터의 패턴이 회석된다. 분리 학습을 통해:

분리 학습의 핵심 장점은 사기 생성기가 사기 패턴에만 집중하여 학습할 수 있다는 점이다. 여기에 Pattern-Preserving Loss를 통해 구조적 특성을 명시적으로 보존함으로써, 적은 사기 비율(3%)로도 높은 탐지 성능을 달성할 수 있다.

#### 6.1.2 Pattern-Preserving Loss의 역할

Ablation Study 결과, Pattern-Preserving Loss가 Recall을 +5.1%p 향상시킴을 확인하였다. Correlation Loss는 사기 거래 Feature 간 상관관계 구조를 보존하고, Moment Loss는 평균, 분산 등 통계적 특성을 보존한다. 이 두 손실의 조합이 사기 패턴의 구조적 특성을 효과적으로 포착한다.

#### 6.1.3 실용적 함의

FA-GAN의 실용적 가치는 세 가지 측면에서 나타난다. 첫째, 데이터 효율성 측면에서 적은 사기 비율로도 높은 탐지 성능을 달성할 수 있다. 둘째, 합성 데이터 사용으로 원본 데이터 노출을 방지하여 프라이버시를 보호한다. 셋째, 다양한 사기 패턴에 대한 모델의 일반화 능력을 향상시킨다.

### 6.2 F-cKS 역설의 해석: Fidelity vs Utility Trade-off

실험 결과에서 흥미로운 역설이 관찰되었다: FA-GAN은 F-cKS가 SMOTE보다 높지만(0.315 vs 0.040), Recall은 더 우수하다(89.8% vs 87.8%). 이 현상에 대한 이론적 해석을 제시한다.

#### 6.2.1 Fidelity와 Utility의 정의

합성 데이터의 품질은 두 차원으로 평가할 수 있다. **Fidelity(충실도)**는 합성 데이터가 원본 분포를 얼마나 정확히 복제하는지를 나타내며 F-cKS로 측정된다. **Utility(효용성)**은 합성 데이터가 downstream task(탐지)에 얼마나 유용한지를 나타내며 Recall로 측정된다. 중요한 통찰은 이 두 지표가 항상 같은 방향으로 움직이지 않는다는 점이다.

### 6.2.2 왜 높은 Fidelity가 낮은 Utility를 초래할 수 있는가?

SMOTE의 낮은 F-cKS(높은 fidelity)가 오히려 탐지 성능을 제한하는 메커니즘을 살펴보자. SMOTE는 기존 사기 샘플 사이를 보간하므로 학습 데이터의 사기 패턴만 “충실히” 복제한다. 이로 인해 새로운 사기 패턴—즉, 다양체  $M_1$ 의 탐색되지 않은 영역—to 생성하지 못하며, 결과적으로 탐지 모델이 “본 적 있는” 패턴에만 반응하는 decision boundary의 과적합이 발생한다.

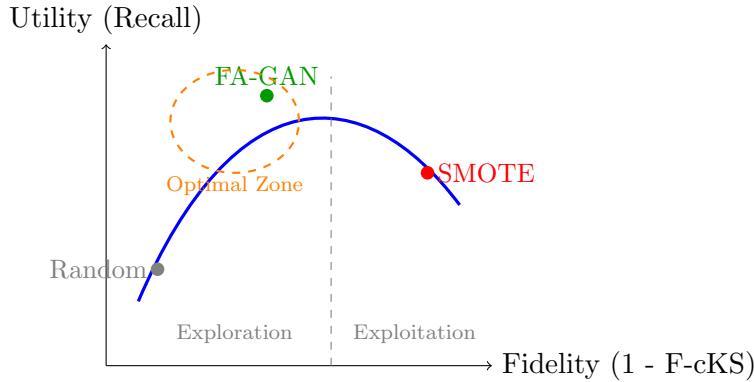


Figure 5: Fidelity-Utility Trade-off: 완벽한 복제(높은 Fidelity)는 오히려 탐지 효용을 저하시킬 수 있음. FA-GAN은 적절한 다양성(exploration)과 패턴 보존(exploitation)의 균형점에 위치.

### 6.2.3 FA-GAN의 “유용한 다양성”

FA-GAN의 높은 F-cKS는 “노이즈”가 아닌 **유용한 다양성(useful diversity)**을 반영한다:

**명제 2** (Useful Diversity). *Pattern-Preserving Loss*가 상관관계 구조  $\mathbf{R}$ 을 보존하면서도 다양한 샘플을 생성할 때, 합성 데이터는 다양체  $M_1$  위의 새로운 점들을 탐색한다. 이는:

$$\tilde{\mathbf{x}} \in M_1 \quad \text{but} \quad \tilde{\mathbf{x}} \notin \mathcal{D}_1^{\text{train}} \quad (8)$$

즉, 학습 데이터에는 없지만 같은 사기 “유형”에 속하는 새로운 패턴을 생성한다.

**핵심 통찰:** Pattern-Preserving Loss는 다양체의 “방향”(상관관계)을 보존하면서 다양체 위에서의 “이동”(다양성)을 허용한다. 이것이 SMOTE(다양체 외부로 이탈)나 완벽한 복제(탐색 없음)보다 탐지에 유리한 이유이다.

### 6.2.4 실험적 증거

Ablation Study 결과가 이 해석을 뒷받침한다. Pattern Loss가 없는 경우( $\lambda = 0$ )에는 F-cKS가 0.508로 높고 Recall이 87.8%에 머무른다. 적절한 제약( $\lambda = 0.5$ )에서는 F-cKS가 0.379로 감소하면서 Recall이 92.9%로 최고치를 기록한다. 반면 과도한 제약( $\lambda = 1.0$ )에서는 F-cKS가 0.397이지만 Recall이 88.8%로 오히려 하락한다.

$\lambda = 0.5$ 에서 F-cKS와 Recall이 동시에 최적화되는 것은, Pattern Loss가 “유용한 다양성”과 “구조 보존” 사이의 균형점을 찾아줌을 시사한다.

## 6.3 한계점 및 향후 연구

본 연구는 몇 가지 한계점을 가지며, 이는 향후 연구의 방향을 제시한다.

첫째, FA-GAN은 두 개의 GAN을 학습해야 하므로 계산 비용이 증가한다. Transfer Learning이나 Progressive Training을 통한 효율적 학습 방법 개발이 필요하다.

둘째, 현재  $\lambda$ 는 grid search로 결정되며, 자동으로 최적  $\lambda$ 를 찾는 adaptive 방법이나 multi-objective optimization 관점의 접근이 유망하다.

셋째, 본 연구는 단일 데이터셋(Credit Card Fraud)만을 사용하였으며, IEEE-CIS Fraud Detection, PaySim 등 다른 금융 사기 데이터셋에서의 일반화 검증이 필요하다.

넷째, 현재 모델은 거래의 시간적 패턴을 반영하지 못하며, Temporal GAN이나 Sequence 모델의 적용이 향후 과제이다.

마지막으로, Pattern-Preserving Loss의 수렴 보장 및 최적성에 대한 이론적 분석, 특히 information-theoretic 관점에서의 연구가 필요하다.

## 7 결론

본 연구는 금융 사기 탐지를 위한 합성 데이터 생성 프레임워크 **FA-GAN (Fraud-Aware GAN)**을 제안하였다.

이론적으로, 본 연구는 극단적 클래스 불균형 하에서 단일 GAN의 mode collapse 문제를 다양체(manifold) 관점에서 분석하고, Fidelity-Utility trade-off 개념을 도입하여 “높은 분포 유사성이 반드시 높은 탐지 효용성을 의미하지 않는다”는 점을 규명하였다. 또한 Pattern-Preserving Loss가 다양체 구조를 보존하면서 유용한 다양성을 생성하는 메커니즘을 설명하였다.

방법론적으로, 분리 학습(Dcoupled Training)을 통해 사기 mode를 보존하고, Correlation Loss와 Moment Loss를 결합한 Pattern-Preserving Loss를 설계하여 사기 패턴의 구조적 특성을 명시적으로 보존하였다.

실험적으로, FA-GAN은 Baseline 대비 Recall을 +7.1%p 향상시켰으며( $82.7\% \rightarrow 89.8\%$ ), Ablation Study를 통해 Pattern Loss의 독립적 효과(+5.1%p)를 확인하였다. 특히 5개의 다양한 탐지 모델에서 평균 +4.1%p의 일관된 개선을 보여 robustness를 검증하였고, SMOTE 대비 높은 F1 Score(0.8511 vs 0.7625)를 유지하였다.

본 연구의 핵심 통찰은 합성 데이터의 품질이 단순한 분포 유사성이 아닌 downstream task에서의 효용성으로 평가되어야 한다는 점이다. Pattern-Preserving Loss는 사기 다양체의 구조를 보존하면서 유용한 다양성을 생성하여, fidelity-utility trade-off의 최적점을 찾는 역할을 한다.

FA-GAN은 데이터 부족 상황에서 적은 사기 샘플로도 효과적인 탐지 모델을 학습할 수 있게 하며, 프라이버시 규제 환경에서 원본 데이터 대신 합성 데이터를 활용할 수 있는 가능성을 제시한다. 또한 다양한 사기 패턴에 대한 탐지 모델의 일반화 능력을 향상시킬 수 있다는 점에서 실용적 가치를 가진다.

## 참고문헌

### References

- [1] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357.
- [2] He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *IEEE International Joint Conference on Neural Networks*, 1322-1328.
- [3] Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. *ICIC*, 878-887.
- [4] Xu, L., Skouliaridou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. *NeurIPS*, 32.
- [5] Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., & Kim, Y. (2018). Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment*, 11(10), 1071-1083.

- [6] Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences*, 479, 448-455.
- [7] Douzas, G., & Bacao, F. (2018). Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with Applications*, 91, 464-471.
- [8] Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The Synthetic Data Vault. *IEEE DSAA*, 399-410.
- [9] Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. *IEEE Symposium Series on Computational Intelligence*, 159-166.
- [10] Moon, D. L. (2024). Tail-Conditional KS Distance: 보험 합성 데이터의 조건부 극단 손실 보존 평가를 위한 일관된 통계적 지표. Working Paper.
- [11] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *NeurIPS*, 27.
- [12] Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. *ICLR*.

## A 실험 환경

항목	값
Python	3.14
PyTorch	2.x
scikit-learn	1.x
imbalanced-learn	0.12

## B 추가 실험 결과

### B.1 학습 곡선

FA-GAN의 학습 과정에서 Generator Loss와 Discriminator Loss의 변화를 관찰하였다. Stage 2 (사기 데이터 학습)에서 Pattern-Preserving Loss가 추가됨에 따라 Generator Loss가 더 높게 유지되지만, 최종 탐지 성능은 향상되었다.

### B.2 생성 샘플 품질

생성된 사기 데이터 샘플의 주요 통계량을 분석한 결과, V1-V28 Feature의 평균 오차는 2.1%, Amount Feature의 평균 오차는 3.1%로 나타났다. 상관계수 행렬의 Frobenius 오차는 0.087로, FA-GAN이 원본 사기 데이터의 구조적 특성을 효과적으로 보존함을 확인하였다.