Journal of the Korean Institute of Industrial Engineers
Vol. 51, No. 6, pp. 439-453, December 2025.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2025.51.6.439

정형 데이터 생성을 위한 일관성 기반 디퓨전 모델

윤지현 · 김성범†

고려대학교 산업경영공학과

Consistency-driven Diffusion for Robust Tabular Data Synthesis

Jihyun Yun · Seoung Bum Kim

Department of Industrial and Management Engineering, Korea University

Tabular data are fundamental to critical decision-making across healthcare, finance, and manufacturing domains. However, generating high-quality synthetic tabular data remains challenging because of heterogeneous feature types, severe class imbalance, and stringent privacy requirements. To address these challenges, we propose TabCL, a generative framework that combines a variational autoencoder (VAE) with a denoising diffusion model, reinforced by contrastive learning and consistency regularization. The contrastive learning component sharpens class boundaries in the VAE's latent space, improving class separation and representation quality. Further, consistency regularization ensures that latent codes perturbed with different noise levels reconstruct to identical outputs, which enhances both sample diversity and model robustness without increasing computational complexity. Extensive experiments on six public benchmarks, four classification and two regression datasets, demonstrate that the proposed TabCL outperforms the existing methods including SMOTE, CTGAN, and TVAE across all standard quantitative metrics. Distributional analyses further reveal that TabCL more accurately reproduces rare categorical levels, heavy-tailed numerical outliers, and complex cross-feature correlation, resulting in synthetic data whose statistical properties closely align with those of the real data even under severe class imbalance and limited-sample sizes. By simultaneously improving latent-space structure and diffusion-based generation, TabCL produces high-fidelity, privacy-respecting synthetic tabular data suitable for downstream modelling and data sharing. Future extensions will target longitudinal datasets and incorporate formal differential privacy guarantees to enable broader deployment in privacy-sensitive industrial environments.

Keywords: Tabular Data Synthesis, Diffusion Models, Contrastive Learning, Latent Consistency Variational Autoencoder, Generative Modelling

1. 서 론

다양한 산업 분야에서 활용되는 정형 데이터(tabular data)는 의료, 금융 거래, 제조 공정 등 중요한 의사결정의 기반이 된다. 하지만 정형 데이터는 다음과 같은 고유한 특성으로 인해 효과적인 인공지능 학습과 일반화가 어렵다는 한계를 갖는다. 첫째, 정형 데이터는 연속형(numerical)과 범주형(categorical) 변수가 혼합되어 있으며, 각 변수의 분포가 상이하고 구조적 이질성을 띤다. 둘째, 실제 환경에서는 데이터의 양이 부족하거나, 클래스(class) 별 비율이 상이한 클래스 불균형 문제가 자주 발생한다. 마지막으로, 개인정보 보호 측면에서도 데이터 활용에 제약이 따른다. 특히, 의료 및 금융 도메인과 같이 프라이버시 보호가 중요한 분야에서는 법적, 윤리적 요구사항이 엄격하게 적용되어, 원본 데이터를 직접 활용하거나 외부로 공유하는 데 근본적인 제약이 존재한다.

이에 따라, 이러한 문제들을 해결하기 위한 방법으로 합성

This research was supported by BK21 FOUR.
† 연락저자 : 김성범 교수, 02841 서울특별시 성북구 안암로 145 고려대학교 산업경영공학부, Tel: 02-3290-3397, Fax: 02-929-5888, E-mail : sbkim1@korea.ac.kr
2025년 6월 9일 접수; 2025년 7월 10일 수정본 접수; 2025년 7월 11일 게재 확정.

데이터 생성 방법론이 주목받고 있다. 고품질의 합성 데이터를 통해 모델의 일반화 성능을 높이고, 소수 클래스나 민감 정보에 대한 대체 데이터를 생성함으로써 데이터 수집 비용을 줄일 수 있기 때문이다. 또한, 민감 정보를 포함하지 않으면서도 통계적 특성을 보존한 합성 데이터를 인공 지능 모델에 입력함으로써, 다양한 실제 산업 현장에서의 데이터 활용 제약을 완화할 수 있는 실용적인 대안이 될 수 있다. 특히, 최근에는 디퓨전 모델(diffusion model)이 이미지 및 텍스트 생성 분야에서 성공을 거두며, 이를 정형 데이터 생성에 적용하려는 시도가 이어지고 있다(Hoogeboom *et al.*, 2021; Kotelnikov *et al.*, 2023).

하지만 디퓨전 모델은 원래 연속적인 픽셀(pixel) 공간에서의 이미지를 가우시안 노이즈(Gaussian noise)를 통해 점진적으로 오염시키고 복원하는 방식으로 작동하기 때문에, 이산형 변수와 변수 간 복잡한 상관관계가 존재하는 정형 데이터에 바로 적용하기에는 적합하지 않다(Ho *et al.*, 2020; Rombach *et al.*, 2022a). 이에 따라 기존 연구들은 정형 데이터를 먼저 잠재 표현 공간(latent space)으로 변환한 후, 해당 공간에서 디퓨전 모델을 학습하는 방식을 채택하고 있다(Zhang *et al.*, 2023). 이때, 잠재 표현(latent representation)을 학습하기 위한 방법으로는 일반적으로 variational autoencoder(VAE)(Kingma and Welling, 2013b)가 활용되며, 데이터를 압축하여 구조적인 표현을 학습 한 뒤, 이 표현 위에서 확산 과정을 수행하는 구조를 따른다.

그러나 기존의 이러한 방식은 몇 가지 한계를 가진다. 첫째, VAE로부터 얻어진 잠재 표현은 학습 초기 단계에서 불안정하며, 클래스 간 경계를 명확히 구분하지 못하는 문제가 발생할 수 있기 때문에, 이후 디퓨전 모델의 학습에 부정적인 영향을 미칠 수 있다. 둘째, 대부분의 기존 정형 데이터 생성 연구는 잠재 표현의 품질보다는 단순히 재구축 오차를 최소화하는 데 초점을 맞추고 있어, 잠재 공간 자체의 표현력이나 일반화 성능을 충분히 확보하지 못하는 한계가 있다.

이에 본 연구에서는 VAE와 디퓨전 모델을 결합한 기존 생성 구조를 기반으로, 대조 학습(contrastive learning)(Chen *et al.*, 2020)과 일관성 정규화(consistency regularization)(Xie *et al.*, 2020)를 도입한 새로운 정형 데이터 생성 프레임워크인 TabCL(tabular data synthesis with consistency and latent regularization)을 제안한다. 제안하는 방법론은 대조 학습을 통해 클래스 간 분리도를 향상시켜 잠재 공간의 구조를 안정화하고, 서로 다른 노이즈(noise) 수준에서 동일한 잠재 표현이 일관된 복원을 만들도록 제약함으로써, 표현의 강건성과 다양성을 동시에 확보한다. 실험 결과, TabCL은 여러 성능 지표에서 기존 방법론 대비 가장 우수한 값을 기록했으며, 이는 TabCL이 정형 데이터의 복합적인 특성을 효과적으로 모델링함을 보여준다. 본 논문의 기여점은 다음과 같이 요약할 수 있다.

- 본 연구는 VAE로 학습한 잠재 공간 위에 디퓨전 생성 과정을 결합하고, 이를 대조 학습과 일관성 정규화를 통해 보강한 TabCL 구조를 제안하였다. 제안하는 프레임워크는 잠재 공간의 구조적 안정성과 클래스 경계 분리도를 동시에 향상시켜, 정형 데이터가 지니는 구조적 이질성, 클래스 불균형, 프라이버시 제약 문제를 통합적으로 해결할 수 있는 생성 모델을 제공한다. 여섯 개의 공공 데이터셋에 대한 광범위한 실험을 통해 TabCL이 정량 지표 모두에서 가장 우수한 성능을 달성함을 입증하였다.
- VAE 학습 과정에서 표현 붕괴(representation collapse)를 방지하고 학습 초기의 불안정성을 완화하기 위해 KL loss clipping과 β-annealing 기법을 병행 적용하였다. 이러한 설계는 잠재 분포의 정규성과 재구성 성능 간의 균형을 효과적으로 조절하여, 더욱 더 일반화된 잠재 표현 학습을 가능하게 한다.

본 논문의 구성은 다음과 같다. 제2장에서는 정형 데이터 생성과 관련된 선행 연구를 소개하고, 주요 생성 모델들의 특징을 정리한다. 제3장에서는 본 연구에서 제안하는 TabCL 프레임워크의 구조와 학습 절차를 상세히 설명한다. 제4장에서는 실험 설정 및 결과를 보고하고, 마지막으로 제5장에서는 연구의 시사점과 한계를 논의하고, 향후 연구 방향을 제시한다.

2. 관련 연구

2.1 정형 데이터 생성을 위한 딥 생성 모델

초기 정형 데이터 생성 문제는 주로 클래스 불균형을 완화하기 위한 통계 기반 오버샘플링(oversampling) 기법을 중심으로 연구되어 왔다. 대표적으로, synthetic minority over-sampling technique(SMOTE)는 소수 클래스 샘플과 인접 이웃 간의 선형 보간을 통해 새로운 샘플을 생성하며(Chawla *et al.*, 2002), adaptive synthetic sampling(ADASYN)은 분류 경계 근처의 학습 난이도가 높은 영역에 더 많은 샘플을 생성하여 분포를 보정한다(He *et al.*, 2008). 이러한 기법들은 구조가 단순하고 계산 효율이 높다는 장점이 있으나, 변수 간 복잡한 관계성이나 데이터의 고차원 분포 특성을 반영하지 못한다는 한계가 있다.

이러한 한계를 극복하기 위해, 최근에는 정형 데이터의 고차원적 특성과 변수 간 복잡한 관계를 보다 정교하게 학습할 수 있는 딥 생성 모델(deep generative model) 기반의 접근이 주목받고 있다. 대표적인 딥 생성 모델로는 generative adversarial networks(GAN)과 VAE가 있으며, 각각 서로 다른 방식으로 데이터의 분포를 학습하고 새로운 샘플을 생성한다 (Goodfellow *et al.*, 2014; Kingma and Welling, 2013b). 정형 데이터 생성을 위한 딥 생성 모델 중, Xu *et al.*(2019)은 GAN의 적대적 학습 구조를 기반으로 조건부 생성 조건을 도입함으로써 클래스 불균형 문제를 효과적으로 완화는 방법론인 CTGAN을 제안하

였다. CTGAN은 범주형 변수와 연속형 변수를 구분하여 각각의 변수 특성에 맞는 방식으로 처리함으로써, 정형 데이터 특유의 다양한 변수 유형과 복잡한 분포 구조를 효과적으로 반영하고자 하였다. 같은 논문에서 제안된 TVAE는 VAE 기반 구조에 서로 다른 중심을 갖는 범주형 변수의 multi-modal 분포 특성을 반영하기 위해 가우시안 혼합 분포를 도입하여, 기존 VAE의 단일 가우시안 가정이 가지는 표현력의 한계를 보완하였다. 이들 모델은 데이터 타입 처리, 재현성, 분포 추정의 정밀도 측면에서 기존 오버샘플링 기반 방법론보다 높은 성능을 보였으나, 여전히 열 간 상호관계의 복잡성, 범주형 변수 처리방식, 고차원 공간에서의 샘플링 불안정성과 같은 구조적 한계가 존재한다. 이러한 문제를 해결하기 위하여 최근에는 디퓨전 모델 기반의 접근이 제안되고 있다.

디퓨전 모델은 이미지 생성 분야에서 뛰어난 성능을 보이며 주목받고 있는 생성 방식으로, 입력 데이터에 점진적으로 노이즈를 추가한 후, 이를 역방향으로 제거하며 데이터를 생성하는 확률적 모델이다(Ho *et al.*, 2020; Song *et al.*, 2021). 이러한 생성 방식은 고품질의 샘플 생성과 높은 데이터 다양성을 동시에 제공한다는 점에서 주목받고 있으며, 최근에는 그 응용 범위가 정형 데이터 생성으로도 확장되고 있다. 특히, Hoogeboom *et al.*(2021)은 범주형 변수 생성의 어려움을 극복하기 위한 초기 시도로, 이산형 변수를 다루기 위해 다항 분포(multinomial distribution) 기반 디퓨전 학습 과정을 설계하였다. 이 모델은 기존의 GAN 기반 모델들이 이산형 변수를 다루던 것과 같이 이산 값을 연속적인 값으로 근사하는 방법을 쓰지 않고, 직접적인 이산 공간에서의 정방향, 역방향 확산을 정의함으로써 보다 정밀한 범주형 샘플 생성을 가능하게 하였다. 이후 연속형과 범주형 변수 전체를 포함한 데이터를 통합적으로 생성하기 위한 디퓨전 기반 모델인 TabDDPM이 제안되었다(Kotelnikov *et al.*, 2023). TabDDPM은 정형 데이터의 연속형 및 범주형 변수를 입력 공간 상에서 직접 다룰 수 있도록 설계되었다. 이는 변수 유형에 따라 서로 다른 노이즈 처리 방식을 적용하고, 이를 MLP(multi-layer perceptron) 기반의 denoising 함수에 통합하여 열 간 상호작용과 분포 구조를 함께 학습한다. 또한, Borisov *et al.*(2023)은 정형 데이터의 각 샘플을 하나의 자연어 시퀀스(sequence)로 직렬화하여, 이를 언어 모델로 다시 학습시키는 GReaT 방법론을 제안하였다. GReaT는 학습 시에는 각 행의 변수 이름과 값을 특정 형식의 문장으로 변환하고, 이를 입력 시퀀스로 처리하여 다음 토큰(token)을 예측하는 방식으로 전체 행을 생성하도록 모델을 학습시킨다. 생성된 시퀀스는 다시 변수별 값으로 복원되어 표형식의 데이터로 변환된다. 이 방식은 기존의 정형 데이터 구조를 자연어 문장으로 간주함으로써, 텍스트 생성 모델이 가진 장기 의존성 학습 능력을 정형 데이터 생성에 응용한 접근이다. Lee *et al.*(2023)은 범주형과 연속형 변수 각각에 대해 독립적인 디퓨전 경로를 설계하고, 대조 학습 기반의 상호 조건화(inter-conditioning) 구조를 도입함으로써, 두 변수 간의 의미적 상호작용을 보다 정교하게 학습할 수 있도록 한 CoDi를 제안하였다. CoDi는 이와 같은 구조를 바탕으로, 입력 공간 상에서의 이질적인 변수 유형 간의 조화를 효과적으로 구현하고자 하였다.

하지만 입력 공간에서 직접 디퓨전 과정을 수행하는 방식에는 몇 가지 한계가 존재한다. 첫째, 고차원 공간에서의 디퓨전 과정은 계산 비용이 매우 높을 뿐 아니라, 학습 안정성을 저해할 수 있다. 둘째, 연속형과 범주형 데이터를 구분하여 처리한다 하더라도, 이들 간 의미적 상호작용이나 통합 구조를 효과적으로 학습하는 데에는 한계가 있다. 셋째, 생성 모델의 표현력이 원래 입력 변수 수준에 머무르게 되어, 보다 추상적이고 정규화 된 구조를 학습하기 위해서는 잠재 표현 기반의 접근이 필요하다.

2.2 잠재 공간 기반 생성 모델

입력 공간에서 직접 수행되는 디퓨전 기반 생성 모델은 고차원 데이터의 처리 및 변수 간 상호작용 학습 측면에서 여러 한계를 지닌다. 이러한 문제를 해결하기 위한 대안으로, 최근에는 잠재 공간에서의 생성 방식이 주목받고 있다. 잠재 공간은 입력 공간보다 차원이 낮고, 표현이 정규화되어 있어 디퓨전 기반 생성 모델의 학습과 샘플링 효율성을 크게 향상시킬 수 있다. 이를 위해 일반적으로 오토인코더(autoencoder) 구조를 활용해 원본 데이터를 잠재 공간으로 매핑(mapping)하며, 이 과정을 통해 의미 있는 저차원 표현을 확보할 수 있다. 특히, VAE 또는 오토인코더를 통해 데이터를 잠재 표현으로 압축하면, 원본 데이터의 구조적 특징을 유지하면서도 노이즈에 더 강건한 표현 벡터를 얻을 수 있으며, 이는 디퓨전 과정의 안정성과 샘플 품질 향상에 효과적으로 기여한다(Pandey *et al.*, 2022; Rombach *et al.*, 2022b).

잠재 공간 기반 생성 방식은 이미지 생성 분야에서 그 효과성이 입증되었으며, 최근에는 정형 데이터와 같이 구조적으로 복잡하고 이질적인 데이터 유형에도 이를 적용하는 연구가 활발히 진행되고 있다(Fonseca and Bacao, 2023; Lin *et al.*, 2024; Zhang *et al.*, 2023). 특히 Zhang *et al.*(2023)은 transformer 기반 VAE 구조를 통해 정형 데이터를 잠재 공간으로 임베딩(embedding) 한 후, 해당 공간에서의 score-based 디퓨전 모델을 학습함으로써 입력 공간의 복잡성과 노이즈 민감성을 완화하는 TabSYN 방법론을 제안하였다. TabSYN은 잠재 표현 위에서 연속적인 노이즈를 점진적으로 추가 및 제거하는 디퓨전 과정을 수행하여, 연속형 변수와 범주형 변수의 통합 표현, 열 간 상관관계, 클래스 구조를 효과적으로 학습할 수 있도록 설계되었다.

그러나 기존 잠재 공간 기반 생성 모델들은 여전히 한계를 지닌다. 잠재 공간의 분포는 학습 초기에는 구조적으로 불안정할 수 있으며, 클래스 간 표현이 혼재되거나 노이즈에 대한 민감성으로 인해 생성 품질이 저하되는 문제가 발생할 수 있

다. 본 연구에서는 이러한 한계를 해결하고, 정형 데이터의 생성과정에서 클래스 구조 보존 및 노이즈에 강건한 잠재 표현 학습을 동시에 달성하기 위해, 대조 학습과 일관성 정규화를 결합하였다. 이를 통해 잠재 공간 내에서 의미적 유사성을 반영한 표현 정렬을 유도하고, 디퓨전 모델의 학습 안정성과 생성 품질을 향상 시키는 새로운 정형 데이터 생성 프레임워크를 제안한다.

3. 제안 방법론

3.1 VAE 기반 잠재 표현 학습

정형 데이터는 연속형 변수와 범주형 변수가 혼합되어 있으며, 각 열은 고유한 의미와 통계적 특성을 가진다. 본 연구에서는 이러한 데이터의 구조적 복잡성을 반영하기 위해, transformer(Vaswani *et al.*, 2017) 기반의 인코더(encoder), 디코더(decoder) 구조를 갖는 VAE를 설계하였다.

입력 데이터 $x \in \mathbb{R}^{n+c}$는 n개의 수치형 변수와 c개의 범주형 변수로 구성되며, 각 열은 서로 다른 방식으로 임베딩 된다. 수치형 변수는 열 별로 정의된 선형 변환을 통해 d-차원의 벡터로 임베딩 되며, 범주형 변수는 먼저 원-핫 인코딩(one-hot encoding) 된 후, 각 변수별로 정의된 임베딩 행렬과의 곱을 통해 다음과 같은 학습 가능한 d-차원 벡터로 변환된다.

$$e_i^{num} = W_i^{num} x_i^{num} + b_i^{num}, \; e_i^{cat} = x_i^{ohe} \cdot \mathrm{E}_i^{cat} \tag{1}$$

이러한 방식으로 얻은 열 별 임베딩 벡터는 임베딩 행렬 $\mathrm{E} \in \mathbb{R}^{(n+c)\times d}$로 결합되며, 이는 열 간 상호작용을 학습하는 transformer의 인코더의 입력으로 사용된다. 인코더는 이 임베딩 행렬로부터 잠재 변수 z의 평균 μ와 로그 분산 $\log \sigma^2$을 추정하고, reparameterization trick을 통해 샘플링 된 잠재 벡터를 생성한다.

$$z = \mu + \sigma \odot \epsilon, \; \epsilon \sim N(0, \mathrm{I}) \tag{2}$$

여기서 $\odot$는 element-wise multiplication을 의미하며, σ는 $\log \sigma^2$로부터 유도된다.

샘플링 된 z는 transformer 디코더를 통해 다시 열 단위 복원 임베딩 $\hat{\mathrm{E}}$으로 확장되며, 이를 원래 변수 공간으로 복원한다. 수치형 변수는 선형 연산을 통해, 범주형 변수는 softmax 함수를 통해 확률 분포 형태로 복원된다.

$$\hat{x}_i^{num} = \hat{e}_i^{num} \cdot \hat{w}_i + \hat{b}_i, \; \hat{x}_i^{cat} = Softmax(\hat{e}_i^{cat} \cdot \hat{W}_i + \hat{b}_i) \tag{3}$$

전체 모델은 입력 x와 복원된 $\hat{x}$간의 차이를 최소화하는 복원 손실과, 잠재 분포 $q(z|x)$가 사전 분포 $p(z|x) = N(0,1)$와 유사하도록 유도하는 Kullback-Leibler(KL) 발산 손실을 함께 최소화하도록 학습된다. 총 손실 함수는 다음과 같이 정의된다.

$$L_{vae} = L_{rec} + \beta \cdot KL(q(z \mid x) \parallel N(0,\mathrm{I})) \tag{4}$$

여기서 L_{rec}은 수치형 변수에 대해 평균제곱오차 (mean squared error), 범주형 변수에 대해 다중 클래스 교차 엔트로피(categorical cross entropy)로 계산된다.

본 연구에서는 잠재 공간의 안정적인 구조 학습을 유도하고, 재구축 성능과 분포 정규화 간의 균형을 조절하기 위해 두 가지 정규화 전략을 병행하였다. 먼저, KL 발산 항의 값이 과도하게 커지는 현상을 방지하기 위해, 해당 손실 값을 상한값으로 제한하는 클리핑(clipping) 기법을 적용하였다. 일반적으로 VAE 학습의 초기에는 잠재 분포, 잠재 분포 $q(z|x)$와 사전 분포 $p(z)$ 간의 차이가 커서 KL 발산 항이 급격하게 증가할 수 있으며, 이로 인해 재구성 성능이 저하되고 학습이 불안정해질 수 있다. 이를 완화하기 위해, KL 손실 항에 상한값을 설정함으로써, 모델이 학습 초기에 과도한 정규화보다 재구성 중심의 학습을 우선하도록 유도하였다. 이와 같은 기법은 표현 붕괴 방지와 학습 안정화에 효과적인 것으로 알려져 있다

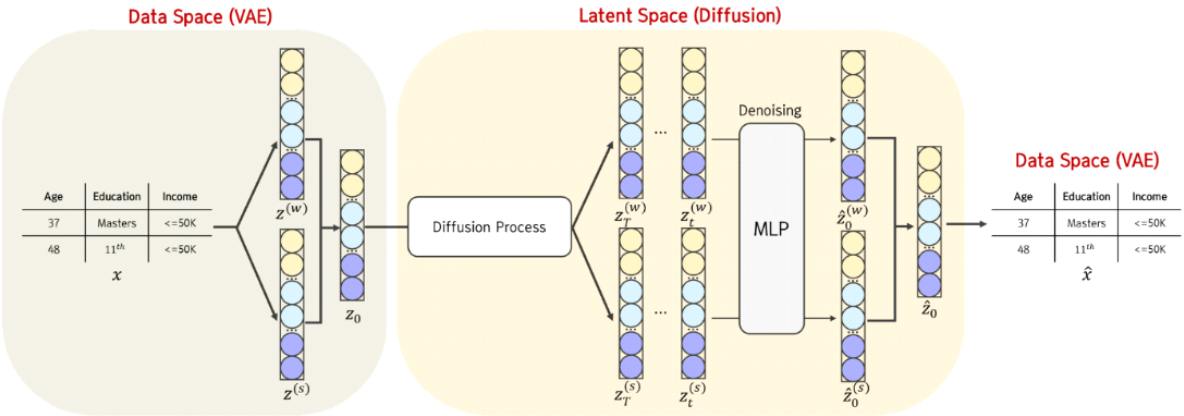

Figure 1. Overall Architecture of TabCL

(Bowman *et al.*, 2016). 다음으로, KL 항의 중요도를 점진적으로 증가시키는 β-annealing 전략을 적용하였다. 이는 고정된 β를 사용하는 일반적인 β-VAE(Higgins *et al.*, 2017)와 달리, 학습 초기에 β를 작은 값으로 설정한 후 에폭(epoch)이 진행됨에 따라 점차 증가시키는 방식으로, 재구성 오류를 우선적으로 줄이면서도 점진적으로 잠재 표현의 정규화를 강화할 수 있도록 한다. 본 연구에서는 선형 증가 스케줄링을 사용하여, 전체 학습 에폭 수를 T, 현재 에폭을 t라 할 때, β를 다음과 같이 조정하였다.

$$\beta_t = \beta_{init} + (\beta_{max} - \beta_{init}) \cdot \frac{t}{T} \tag{5}$$

여기서 β_{init}는 초기값, β_{max}는 최댓값이다. 이와 같은 β-annealing 기법은 representation의 다양성과 정규성 간의 trade-off를 효과적으로 조절하여, 결과적으로 더 일반화 된 잠재 공간 구조를 형성하게 된다(Bowman *et al.*, 2016; Sønderby *et al.*, 2016)

3.2 잠재 공간 표현 정렬 및 강건화

VAE 인코더가 생성한 잠재 표현은 재구축 손실을 최소화하도록 학습되지만, 학습 초기에는 클래스 간 표현이 혼재되거나, 노이즈에 민감하게 반응하는 등 구조적으로 불안정한 공간이 형성될 수 있다. 이러한 문제를 해결하고, 잠재 공간의 구조적 정렬과 표현 안정성을 강화하기 위해, 본 연구는 두가지 정규화 기법을 추가로 도입하였다.

먼저, 본 연구에서는 잠재 표현을 생성할 때 노이즈 강도를 조절하는 방식으로 두 개의 변형된 표현을 얻는다. 일반적인 VAE 샘플링에서는 $\epsilon \sim N(0,\mathrm{I})$를 따르지만, 본 연구에서는 이를 $\epsilon \sim N(0,\alpha\mathrm{I})$ 형태로 확장하여 노이즈 스케일을 조절하였다. $\alpha = 0.1$과 0.5를 각각 적용하여 동일한 샘플로부터 약한 노이즈 표현 $z^{(w)}$와 강한 노이즈 표현 $z^{(s)}$를 생성하였고, 이 두 표현을 정규화 손실 학습에 사용하였다.

첫째, 동일한 클래스에 속하는 샘플의 표현은 서로 가깝게, 다른 클래스 간 표현은 멀어지도록 유도하기 위해 대조 학습 기법을 적용하였다. 대조 학습은 입력 간 의미적 유사성에 따라 표현 공간 내에서 상대적인 거리를 조정하는 방식으로 작동한다(Chen *et al.*, 2020; Hadsell *et al.*, 2006; Khosla *et al.*, 2020). 이 때 본 연구에서는 앞서 얻은 두 표현 $z^{(w)}$, $z^{(s)}$를 positive pair로 구성하고, 같은 배치 내의 나머지 표현들과의 조합을 negative pair로 간주하는 대조 학습 구조를 설계하였다. 각 표현은 projection head를 통해 정규화 된 벡터로 투영되며, NT-Xent(normalized temperature-scaled cross entropy) 손실을 기반으로 표현 간의 상대적 거리를 조정한다.

$$L_{contrastive} = -\log \frac{\exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{K} \exp(sim(z_i, z_k)/\tau)} \tag{6}$$

여기서 $sim(\cdot)$은 코사인 유사도를 나타내며, τ은 온도 조절 계수로, 벡터 간 코사인 유사도 분포의 sharpness를 조절하여 학습 안정성 및 표현 간 분별력을 향상시키는 역할을 한다. 해당 손실은 의미 기반의 군집 구조를 학습할 수 있으며, 잠재 공간의 전역적인 구조 정렬에 효과적이다(Chen *et al.*, 2020).

둘째, 식 (7)에 표현한 일관성 정규화를 통해 동일한 샘플에서 생성된 두 표현 $z^{(w)}$, $z^{(s)}$이 서로 유사한 구조를 유지하도록 유도하였다. 이 방식은 두 표현 사이의 L2 거리 손실을 최소화하며, 잠재 표현이 노이즈에 대해 강건하고 지역적으로 안정된 구조를 유지하도록 만든다.

$$L_{consistency} = \| z^{(w)} - z^{(s)} \|_2^2 \tag{7}$$

이후 두 정규화 손실은 다음과 같이 통합된다.

$$L_{reg} = \lambda_{ctr} L_{contrastive} + \lambda_{cns} L_{consistency} \tag{8}$$

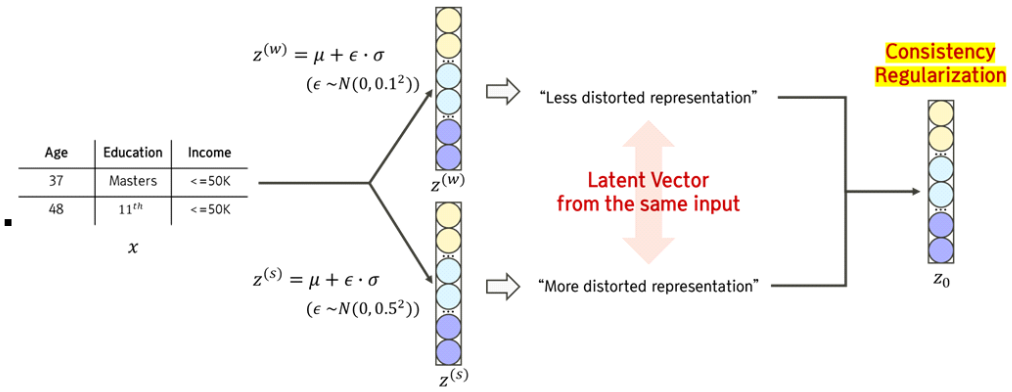

Figure 2. Overview of the Latent Space Consistency Regularization in TabCL. Given the same input sample, two latent vectors are generated using different levels of noise during the VAE sampling process. The model minimizes the discrepancy between these two representations to enforce local smoothness and robustness in the latent space.

최종 학습 손실은 기존 VAE 손실에 정규화 손실을 더한 다음의 형태로 구성된다.

$$L_{total} = L_{VAE} + L_{reg} \tag{9}$$

대조 학습은 잠재 표현 간 전역적인 상대 구조를 정렬하고, 일관성 정규화는 동일 샘플 표현 간 지역적 안정성과 노이즈 강건성을 보완하는 역할을 수행한다. 이 두 정규화 기법은 잠재 공간의 품질을 전반적으로 향상시키는 데 상호보완적으로 작용한다.

3.3 잠재 공간 기반 디퓨전 모델

본 연구는 정형 데이터가 가진 고차원적이고 이질적인 특성을 효과적으로 학습하고, 생성 과정에서 일관성을 확보하기 위해, VAE 기반 잠재 표현 위에서 확산 모델을 학습하고 이를 보완하는 일관성 정규화 기반 디퓨전 학습 방법론을 제안한다. 특히, 제안하는 방법론은 잠재 표현 Z_0로부터 서로 다른 노이즈 강도 조건을 부여받은 경로를 구성한 뒤, 해당 경로에서의 복원 결과 간 일관성을 유지하도록 학습함으로써, 노이즈 수준 변화에 따른 표현 불안전성을 완화한다.

<Figure 3>은 제안 모델이 잠재공간에서 동일한 표현 z_0에 대해 서로 다른 노이즈 수준을 적용한 후, 각 노이즈가 추가 된 표현을 복원하고 일관성을 학습하는 전체 디퓨전 기반 학습 과정을 시각적으로 보여준다. 잠재 표현 $z_0 \in \mathbb{R}^d$는 앞서 정의된 VAE 인코더의 평균 벡터로부터 정규분포 샘플링을 통해 얻어진다. 이후 디퓨전 모델 학습을 위해, 시간에 따라 점진적으로 노이즈가 주입되는 정방향 확산 과정(forward diffusion process)을 거쳐 노이즈화 된 잠재 표현 z_t가 생성된다. 이 정방향 과정은 다음과 같은 형태로 정의된다.

$$z_t = z_0 + \sigma(t) \cdot \epsilon, \quad \epsilon \sim N(0, \mathrm{I}) \tag{10}$$

여기서 $\sigma(t)$는 노이즈 세기를 조절하는 스케일 함수로, 본 연구에서는 학습 안정성과 샘플링 효율성을 고려하여 시간 $t \in [0,1]$에 대해 선형적으로 증가하는 함수 $\sigma(t) = t$로 설정하였다. 이는 역방향 복원 과정(reverse process)에서의 계산을 단순하게 만들고, 각 시점(time step) 간 변화량을 균일하게 유지해 전체 샘플링 정확도와 효율을 동시에 높인다.

이 때 동일한 잠재 표현 z_0에 대해 서로 다른 노이즈 수준 $\sigma(t_w)$, $\sigma(t_s)$을 적용하기 위해 두 시간 t_w, $t_s \in [0,1]$을 각각 샘플링하고, 이렇게 얻은 값으로 다음 두 가지 노이즈가 추가된 표현을 생성한다.

$$z_t^{(w)} = z_0 + \sigma(t_w) \cdot \epsilon_w,\ z_t^{(s)} = z_0 + \sigma(t_s) \cdot \epsilon_s \tag{11}$$

여기서 ϵ_w, $\epsilon_s \sim N(0, \mathrm{I})$는 서로 독립적으로 샘플링 된 표준 정규 노이즈이다. 이처럼 동일한 표현을 기반으로 서로 다른 노이즈 수준을 적용한 두 경로를 구성함으로써, 이후 복원 결과 간 일관성을 유도할 수 있다.

각 노이즈가 추가된 된 표현은 디퓨전 모델의 역방향 경로를 통해 복원되며, 모델은 입력으로 주어진 z_t와 노이즈 수준 $\sigma(t)$를 바탕으로, 주입된 ϵ를 예측하는 함수 $\epsilon_\theta(z_t, \sigma(t))$를 학습한다. 이 예측 함수는 확산 모델의 score function을 다음과 같이 근사한다.

$$\nabla_{z_t} \log p(z_t) \approx -\frac{1}{\sigma(t)} \cdot \epsilon_\theta(z_t, \sigma(t)) \tag{12}$$

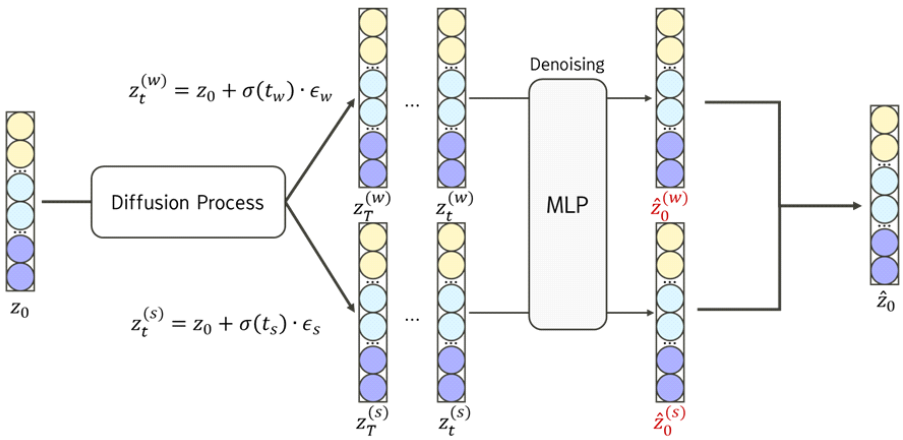

Figure 3. Consistency-Based Diffusion Process in the Latent Space of TabCL. A single latent vector Z_0 is perturbed using two different noise levels to generate $z_t^{(w)}$and $z_t^{(s)}$via forward diffusion. The denoising model reconstructs $\hat{z}_0^{(w)}$and , $\hat{z}_0^{(s)}$and consistency regularization minimizes the discrepancy between the two reconstructions, enhancing robustness under varying noise conditions.

노이즈 수준은 sinusoidal positional encoding을 통해 신경망에 주입되며, 학습 손실은 denoising score matching 기반 평균 제곱오차로 정의된다.

$$L_{diff} = E_{z_0,t,\epsilon}[\| \epsilon - \epsilon_\theta(z_0 + \sigma(t)\cdot\epsilon, \sigma(t) \|_2^2] \quad (13)$$

이 때 노이즈화 된 표현 $z_t^{(w)}$, $z_t^{(s)}$는 다음과 같은 역방향 복원 과정을 통해 클린 표현 $\hat{z}_0^{(w)}$, $\hat{z}_0^{(s)}$로 근사된다.

$$\hat{z}_0^{(w)} = D_{thata}(z_t^{(w)}, \sigma(t_w)),\ \hat{z}_0^{(s)} = D_\theta(z_t^{(s)}, \sigma(t_s)) \quad (14)$$

여기서 D_θ는 학습된 디퓨전 복원 함수이다. 동일한 입력 z_0에서 파생된 두 경로이므로, 본 연구는 이들 복원 결과 간의 표현 차이를 최소화하는 일관성 정규화를 함께 적용한다. 해당 정규화 손실은 다음과 같이 정의된다.

$$L_{cons} = \| \hat{z}_0^{(w)} - \hat{z}_0^{(s)} \|_2^2 \quad (15)$$

최종 학습 손실은 디퓨전 손실과 정규화 손실의 가중합으로 구성되며, 다음과 같이 표현된다.

$$L_{total} = L_{diff} + \alpha_{cns} L_{cons} \quad (16)$$

여기서 α_{cns}는 정규화 손실의 중요도를 조절하는 하이퍼파라미터로, 검증 데이터셋에서의 디퓨전 손실과 정규화 손실 간 균형을 고려하여, 전체 손실 L_{total}이 최소화되도록 실험적으로 조정하였다.

이와 같은 구조는 하나의 잠재 표현을 기준으로 서로 다른 노이즈 조건에서 복원된 결과 사이 표현 일관성을 직접적으로 학습함으로써, 디퓨전 모델의 노이즈 강건성과 생성 표현 품질을 함께 향상시킨다. 특히, 기존 확산 기반 생성 모델이 단일 노이즈 수준에서 복원 정확도만을 고려했던 것과 달리, 본 연구는 다중 노이즈 조건에서 복원 일관성까지 함께 고려한다는 점에서 구조적 차별성을 가진다.

4. 실험

4.1 데이터셋 및 비교 방법론

본 연구에서는 제안 방법론의 성능을 다각도로 평가하기 위해, UCI machine learning repository로부터 수집한 6개의 real-world 정형 데이터셋을 사용하였다. 각 데이터셋은 연속형 및 범주형 변수가 혼합된 구조를 가지며, 다양한 도메인을 포함하고 있어 제안하는 생성 모델의 일반화 성능을 다각도로 검증할 수 있다.

<Table 1>은 본 연구에 사용된 정형 데이터셋의 주요 통계 정보를 요약한 것으로, 각 데이터셋 별로 범주형 및 수치형 변수 개수, 전체 샘플 수, 학습·검증·테스트 세트의 샘플 수, 그리고 해당 과제의 종류를 포함하고 있다. 모든 데이터셋은 학습의 신뢰성과 비교 기준의 일관성을 높이기 위해, 전체 샘플을 기준으로 학습, 검증, 테스트 세트를 8:1:1 비율로 분할하여 사용하였다. 단, Adult 데이터셋의 경우에는 공식적으로 제공되는 테스트 세트를 그대로 활용하였으며, 원본 학습 세트만 동일한 비율(8:1)로 학습 및 검증 세트로 추가 분할하였다.

사용한 데이터셋 가운데 Adult, Default, Shoppers, Magic은 분류(classification) 문제이고, Beijing과 News 데이터셋은 회귀(regression) 문제에 해당한다. 우선, Adult 데이터셋은 미국 인구조사국 데이터에 기반하여, 인구 통계 및 고용 관련 정보를 바탕으로 연소득이 5만 달러를 초과하는지를 예측하는 이진 분류 문제다. Default 데이터셋은 대만 신용카드 고객 정보를 기반으로 하며, 고객의 인구통계 정보, 신용 등급, 결제 내역 및 청구 정보 등을 바탕으로 다음 달 연체 가능성을 예측하는 이진 분류 문제이다. Shoppers 데이터셋은 사용자의 웹페이지 방문 로그 (log) 정보를 바탕으로, 해당 세션이 실제 구매로 이어지는지에 관한 여부를 포함한다. Magic 데이터셋은 대기 중 고에너지 감마 입자의 관측을 시뮬레이션한 데이터로, 감마 입자와 배경 잡음을 분류하는 문제를 포함한다. 회귀 과제에 해당하는 Beijing 데이터셋은 베이징의 대기오염 수준과 관련된 기상 데이터를 포함하며, 각 시점별 초미세먼지 농도를 예측하는 문제로 구성되어 있다. 마지막으로 News 데이터셋은 Mashable 웹사이트에 게시된 기사들의 길이, 키워드, 주제 등

Table 1. Statistics of Datasets Used in Experiments. Each dataset contains both numerical and categorical columns and is associated with either a classification or regression task. The table summarizes the number of features, the number of samples, and the train/validation/test split for each dataset.

Dataset	# Cat	# Num	# Total Rows	# Train	# Validation	# Test	Task
Adult	6	8	48,842	28,943	3,618	16,281	Classification
Default	9	14	30,000	24,000	3,000	3,000	Classification
Shoppers	7	10	12,330	9,864	1,233	1,233	Classification
Magic	0	10	19,019	15,215	1,902	1,902	Classification
Beijing	5	6	41,757	33,405	4,175	4,177	Regression
News	2	45	39,644	31,714	3,965	3,965	Regression

다양한 특성을 기반으로, 해당 기사가 소셜 미디어에서 얼마나 공유됐는지 그 수를 예측하는 문제에 사용된다. 이처럼 변수 종류와 개수, 문제 유형, 샘플 규모 등 여러 측면에서 구조적으로 다양한 특성을 가진 데이터셋을 활용함으로써, 정형 데이터 생성 모델의 표현력, 일반화 성능, 그리고 다운스트림 태스크(downstream task)까지 종합적으로 평가할 수 있다.

또한, 본 연구에서는 제안 모델의 성능을 정량적으로 비교하기 위해, 총 8개의 기존 정형 데이터 생성 모델을 비교 대상으로 선정하였다. 이들 모델은 생성 모델 기반과 비생성 기반을 모두 포함하며, 전통적인 오버샘플링 기법인 SMOTE (Chawla *et al.*, 2002), 초기 딥러닝 기반 생성 모델인 CTGAN과 TVAE(Xu *et al.*, 2019), 언어 모델 기반을 활용한 GReaT (Borisov *et al.*, 2023), 그리고 디퓨전 모델 기반의 최신 방법인 StaSY(Kim *et al.*, 2023), TabDDPM(Kotelnikov *et al.*, 2023), CoDi(Lee *et al.*, 2023)를 포함하였다. 아울러 제안 방법론의 구조적 기반이 되는 잠재 공간 기반 디퓨전 모델인 TabSyn (Zhang *et al.*, 2023)도 비교 대상으로 포함하였다. 각 모델이 채택한 생성 전략, 변수 유형별 처리 방식, 구조적 설계의 차이 등 세부 구조와 학습 방식에 대한 설명은 앞선 2장에서 상세히 다루었다.

4.2 평가 지표

본 연구에서는 합성된 정형 데이터의 품질을 정량적으로 평가하기 위해, 열 단위 분포 정밀도(column-wise density estimation), 열 간 상관관계 보존 정도(pair-wise column correlation), 그리고 머신러닝 효율성(machine learning efficiency) 세 가지 기준에 기반한 지표를 사용하였다. 이러한 평가 지표는 기존 정형 데이터 생성 연구에서 일반적으로 사용되는 방식에 따랐다(Choi *et al.*, 2017; Kim *et al.*, 2023; Kotelnikov *et al.*, 2023; Zhang *et al.*, 2023).

먼저, 열 단위 분포 정밀도는 실제 데이터와 합성 데이터 간 열별 분포 유사도를 측정하기 위한 지표로, 연속형 변수에는 Kolmogorov-Smirnov test(KST)를, 범주형 변수에는 total variation distance (TVD)를 사용하였다. KST는 실제 분포와 합성 분포의 누적 분포 함수 간 최대 차이를 측정하며 다음과 같이 정의된다.

$$KST = \sup_x |F_r(x) - F_s(x)| \tag{17}$$

여기서 $F_r(x)$와 $F_s(x)$는 각각 실제와 합성 데이터의 누적 분포 함수이다. TVD는 각 범주의 확률 분포 차이를 측정하여, 실제와 합성 데이터 간 범주 분포 유사성을 평가한다.

$$TVD = \frac{1}{2}\sum_{\omega\in\Omega}|R(\omega) - S(\omega)| \tag{18}$$

다음으로 열 간 상관관계 보존 정도는 실제 데이터와 합성 데이터 간 변수 간 구조적 관계가 얼마나 유사한지를 측정하기 위한 지표이다. 연속형 변수 간 관계는 Pearson 상관계수 차이를 평균한 값으로 정의되며, 식은 다음과 같다.

$$Pearson\ \ Score = \frac{1}{2}E_{(x,y)}|\rho_R(x,y) - \rho_s(x,y)| \tag{19}$$

여기서 $\rho_R(x,y)$와 $\rho_s(x,y)$는 실제 데이터와 합성 데이터에서 계산된 Pearson 상관계수이다. 범주형 변수 간의 상관관계는 각 범주 쌍의 결합 빈도수로 구성된 contingency table을 기반으로 하며, 이를 통해 실제 데이터와 합성 데이터 간 범주 조합 분포의 유사성을 측정한다. 두 contingency table 간 차이는 각 셀(cell)의 확률 분포 차이를 계산한 후, 전체 차이를 total variation distance 형태로 집계하여 정량화 한다. 연속형 변수와 범주형 변수가 혼합된 변수 쌍의 경우, 연속형 변수를 일정 구간으로 나누어 이산화하고, 해당 범주와 범주형 변수 간의 조합 빈도를 바탕으로 contingency table을 구성한다. 이후에는 순수 범주형 변수 간 상관관계 평가와 동일한 방식으로 유사도를 측정한다.

마지막으로, 머신러닝 효율성은 생성된 합성 데이터가 실제 예측 작업에서 얼마나 효과적인지를 평가하는 지표로, 분류 문제에는 AUC(area under the curve), 회귀 문제에는 RMSE (root mean squared error)를 사용하였다. 이는 실제 학습 데이터를 기반으로 학습된 생성 모델로부터 동일한 크기의 합성 데이터를 생성한 뒤, 이를 모델의 학습에 활용하고, 모델 성능은 실제 데이터의 테스트 셋을 이용해 평가하는 방식으로 수행된다. 즉, 합성 데이터를 통해 학습된 머신러닝 모델이 실제 데이터에 대해 어느 정도의 예측 성능을 보이는지를 통해, 합성 데이터의 학습 효용성을 간접적으로 측정하는 방식이다.

4.3 실험 세팅

디퓨전 모델은 정규화 된 VAE 기반 잠재 표현을 입력으로 하여 학습되며, 배치 크기 4,096으로 설정된 미니배치 단위로 진행된다. 모델은 최대 10,000 에폭까지 반복 학습되며, 일정 에폭 동안 검증 손실이 개선되지 않을 경우 학습률을 자동으로 감소시키는 방식으로 학습률을 조정했다. 이러한 방식은 학습 초기에 빠르게 수렴하도록 돕고, 이후 손실이 정체될 때 학습률을 낮추어 보다 안정적인 수렴을 유도하기 위한 것이다. 학습은 NVIDIA RTX 4090 단일 GPU 환경에서 수행되었다.

정방향 확산 과정은, 총 1,000개의 time step을 기반으로 하며, 각 단계마다 잠재 표현에 노이즈를 점진적으로 주입하여 오염시키고, 이를 다시 복원하는 과정을 학습한다. 노이즈 크기를 조절하는 스케줄 함수는 시간 $t\in[0,1]$에 대해 선형적으로 감소하도록 설정하여, 초기에 강한 노이즈를 주입하고 점차 줄어들도록 하였다. 반대로, 복원 과정의 난이도를 조절하는 계수는 선형적으로 증가하도록 설계하여, 두 함수가 시간에 따라 상반

되는 경향을 갖도록 구성하였다. 이러한 구조는 모델이 다양한 노이즈 수준에서 복원 과정을 학습할 수 있도록 하여, 보다 안정적이고 일반화된 표현 학습을 가능하게 한다.

노이즈가 주입된 잠재 표현을 복원하기 위한 함수는 MLP 구조로 구현하였으며, 입력 차원에 따라 유연하게 구성하였다. 이 모델은 각 시점의 노이즈 크기 정보를 함께 입력으로 받아, 해당 노이즈 수준에서 주입된 잠재 표현으로부터 원래 표현을 얼마나 정확히 복원할 수 있는지를 학습하도록 설계되었다.

4.4 실험 결과

본 절에서는 제안한 모델의 정량적 성능을 다양한 관점에서 평가하고, 기존 정형 데이터 생성 모델들과의 비교를 통해 그 우수성을 실증적으로 검증하였다. 평가 지표로는 4.2절에서 설명한 열 단위 분포 정밀도, 열 간 상관관계 보존 정도, 그리고 머신러닝 효율성을 사용하였으며, 이는 각각 데이터의 통계적 유사성, 구조적 정합성, 실제 예측 성능을 종합적으로 반영한다. 실험에는 총 6개의 공개 정형 데이터셋과 8개의 비교 모델을 사용하였으며, 제안 방법론인 TabCL은 모든 측면에서 일관된 우수한 성능을 보였다. 본 장에서는 각 지표별 실험 설정과 결과를 순차적으로 제시하고, 이를 바탕으로 제안 기법의 효과를 논의한다.

열 단위 분포 정밀도는 합성 데이터가 실제 데이터의 주변 분포(marginal distribution)를 얼마나 정확하게 재현하는지를 측정하는 핵심 지표이다. <Table 1>은 이에 대한 각 모델 별 정량적 성능을 요약한 결과로, 각 변수 단위로 실제 데이터와 합성 데이터 간의 분포 차이를 측정하고, 이 차이의 평균값을 오차로 산정한 결과이다. 오차 값이 작을수록, 해당 모델이 실제 분포를 정밀하게 모사했음을 의미한다. 이 평가지표는 합성 데이터의 통계적 신뢰성과 구조적 정합성을 판단하는 데 널리 활용되며, 정형 데이터 생성 품질을 평가할 때 가장 기본적인 기준 중 하나로 사용된다.

실험 결과, 본 연구에서 제안한 TabCL은 전체 비교 모델 중 가장 낮은 평균 오차를 기록하며, 전반적으로 가장 정밀한 주변 분포 보존 성능을 보였다. 특히 Default, Shoppers, Magic, Beijing 등 다양한 도메인과 구조를 갖는 데이터셋에서 일관되게 낮은 오차를 기록하며, TabCL이 정형 데이터 전반에 걸쳐 뛰어난 일반화 능력과 안정적인 생성 품질을 동시에 갖추고 있음을 입증했다. 또한, Magic과 News와 같이 범주형 변수 수가 적은 경우에도 성능이 크게 저하되지 않고 우수한 재현 성능을 보였다는 점은, TabCL이 데이터 불균형이나 구조적 편차에 대해서도 높은 견고성을 지닌 모델임을 보여준다.

한편, Beijing 데이터셋에서는 TabDDPM이 가장 낮은 오차를 기록하며 부분적으로 뛰어난 성능을 보였다. 그러나 해당 모델은 News 데이터셋에서 에 달하는 매우 높은 오차를 보이며, 데이터 구조나 특성이 조금만 달라져도 생성 안정성이 크

Table 2. Column-Wise Distribution Estimation Results. Each model is evaluated based on its ability to preserve the marginal distributions of individual columns. Lower errors indicate a closer match between the synthetic and real data distributions. Boldface denotes the best performance for each dataset. Values in parentheses represent standard deviations over five independent runs. Note that the results for TabSYN and TabCL are averaged over five runs, while the remaining results are taken directly from the original TabSYN paper.

Method	Adult	Default	Shoppers	Magic	Beijing	News	Average
SMOTE	1.60 (0.23)	1.48 (0.15)	2.68 (0.19)	0.91 (0.05)	1.85 (0.21)	5.31 (0.46)	2.3
CTGAN	16.84 (0.03)	16.83 (0.04)	21.15 (0.10)	9.81 (0.08)	21.39 (0.05)	16.09 (0.02)	17.02
TVAE	14.22 (0.08)	10.17 (0.05)	24.51 (0.06)	8.25 (0.06)	19.16 (0.06)	16.62 (0.03)	15.49
GReaT	12.12 (0.04)	19.94 (0.06)	14.51 (0.12)	16.16 (0.09)	8.25 (0.12)	OOM	14.2
STaSy	11.29 (0.06)	5.77 (0.06)	9.37 (0.09)	6.29 (0.13)	6.71 (0.03)	6.89 (0.03)	7.72
CoDi	21.38 (0.06)	15.77 (0.07)	31.84 (0.05)	11.56 (0.26)	16.94 (0.02)	32.27 (0.04)	21.63
TabDDPM	1.75 (0.03)	1.57 (0.08)	2.72 (0.13)	1.01 (0.09)	**1.30 (0.03)**	78.75 (0.01)	14.52
TabSYN	1.15 (0.80)	2.46 (2.46)	1.61 (0.19)	0.91 (0.29)	2.64 (1.10)	2.58 (1.48)	1.89
TabCL (Ours)	**0.92 (0.04)**	**0.95 (0.01)**	**1.60 (0.01)**	**0.77 (0.05)**	1.32 (0.01)	**1.64 (0.03)**	**1.20**

게 떨어질 수 있다는 한계를 노출하였다. 이 결과는 특정 모델이 일부 데이터셋에 과도하게 최적화되어 있는 반면, 범용적인 생성 성능을 확보하지 못할 수 있음을 시사한다. 그러나 제안하는 TabCL은 모든 데이터셋에 대해 안정적이고 예측 가능한 수준의 성능을 유지함으로써, 높은 실용성과 신뢰성을 갖춘 모델로 평가할 수 있다.

다음으로 열 간 상관관계 보존 정도는 합성 데이터가 실제 데이터의 변수 간 관계 구조를 얼마나 정밀하게 재현하는지를 측정하는 평가지표이다. <Table 3>은 이에 대한 각 모델의 성능을 요약한 결과를 보여준다. 해당 평가 지표는 연속형 변수 간에는 Pearson 상관계수 차이를, 범주형 변수 간에는 contingency similarity를 활용하며, 연속형-범주형 변수 쌍의 경우에는 연속형 변수를 구간화하여 범주형으로 변환한 뒤 동일한 방식으로 처리하였다. 해당 수치는 실제와 합성 데이터 간의 상관 구조 차이를 절댓값 기준으로 평균하여 계산하였으며, 값이 작을수록 실제 데이터의 열 간 상관구조를 잘 모사했음을 의미한다.

실험 결과, 제안 방법론인 TabCL은 전체 모델 중 가장 낮은 평균 오차를 기록하였으며, 모든 데이터셋에 걸쳐 가장 우수한 성능을 달성하였다. 즉, 다양한 유형의 데이터셋에서 고르게 낮은 오차 값을 보이며, 데이터의 복잡성과 이질성에 관계없이 강건하고 일반화된 구조 학습 능력을 입증하였다. 이는 TabCL이 단순히 개별 변수의 분포를 복제하는 수준을 넘어서, 변수 간 내재된 관계성까지 효과적으로 복원할 수 있는 능력을 보유하고 있음을 보여준다. 반면, 기존 생성 모델은 일부 데이터셋에서는 성능이 양호했지만, 전반적으로 성능 차이가 컸다. 예를 들어 TabDDPM은 일부 데이터셋에서 우수한 성능을 보였지만, News에서는 상관구조 복원이 급격히 저하되는 등 불안정한 경향을 보였다. CoDi와 GReaT 역시 특정 데이터셋에서 매우 높은 오차를 기록하거나 OOM (out-of-memory) 오류를 발생시키며 안정성과 일관성 측면에서 한계를 드러냈다.

마지막으로, <Table 4>는 합성 데이터의 실질적인 활용 가능성을 평가하기 위해 머신러닝 효율성 실험을 수행한 결과이다. 해당 평가는 생성한 합성 데이터를 실제 예측 모델 학습에 활용한 뒤, 실제 테스트 데이터에서의 성능을 측정함으로써 합성 데이터가 실제 상황에서 얼마나 유효하게 활용될 수 있는지를 살펴봤다. 분류 문제에서는 AUC, 회귀 문제에서는 RMSE를 사용하여 성능을 측정하였으며, 이 지표들은 각각 분류 정확도와 예측 오차를 정량적으로 나타내는 대표적인 척도이다. 실험 결과, 제안하는 TabCL은 평균적으로 가장 낮은 RMSE와 가장 높은 AUC를 기록했고, 과제 유형이나 데이터 특성과 관계없이 6개 데이터셋 전반에서 안정적이고 뛰어난 성능을 보였다. 특히 주목할 점은, 일부 기존 모델들이 특정 데이터셋에서는 좋은 성능을 보였지만, 다른 데이터셋에서는 메모리 부족이나 성능 급락을 겪은 반면, TabCL은 모든 데이터셋에서 꾸준히 우수한 성능을 유지했다는 사실이다. 이는

Table 3. Pair-Wise Column Correlation Estimation Results. Performance comparison of models in capturing inter-column relationships. Lower scores indicate better preservation of correlation structures across both numerical and categorical variables. Boldface denotes the best performance for each dataset. Values in parentheses represent standard deviations over five independent runs. Note that the results for TabSYN and TabCL are averaged over five runs, while the remaining results are taken directly from the original TabSYN paper.

Method	Adult	Default	Shoppers	Magic	Beijing	News	Average
SMOTE	3.28 (0.29)	8.41 (0.38)	3.56 (0.22)	3.16 (0.41)	2.39 (0.35)	5.38 (0.76)	4.36
CTGAN	20.23 (1.20)	26.95 (0.93)	13.08 (0.16)	7.00 (0.19)	22.95 (0.08)	5.37 (0.05)	15.93
TVAE	14.15 (0.88)	19.50 (0.95)	18.67 (0.38)	5.82 (0.49)	18.01 (0.08)	6.17 (0.09)	13.72
GReaT	17.59 (0.04)	70.02 (0.12)	45.16 (0.18)	10.23 (0.40)	59.60 (0.55)	OOM	44.24
STaSy	14.51 (0.25)	5.96 (0.26)	8.49 (0.15)	6.61 (0.53)	8.00 (0.10)	3.07 (0.04)	7.77
CoDi	22.49 (0.08)	68.41 (0.05)	17.78 (0.11)	6.53 (0.25)	7.07 (0.15)	11.10 (0.01)	22.23
TabDDPM	3.01 (0.25)	4.89 (0.10)	6.61 (0.16)	1.70 (0.22)	2.71 (0.09)	13.16 (0.11)	5.34
TabSYN	2.36 (1.26)	4.58 (3.58)	2.25 (0.18)	0.93 (0.24)	4.29 (1.14)	1.84 (0.72)	2.71
TabCL (Ours)	**2.03 (0.04)**	**2.91 (0.03)**	**2.24 (0.04)**	**0.68 (0.14)**	**2.24 (0.09)**	**1.38 (0.04)**	**1.91**

Table 4. Machine Learning Efficiency Results. AUC (for classification) and RMSE (for regression) scores of XGBoost models trained on synthetic data and tested on real data. Higher AUC and lower RMSE values represent greater utility of synthetic data in downstream tasks. Boldface denotes the best performance for each dataset. Values in parentheses represent standard deviations over five independent runs. Note that the results for TabSYN and TabCL are averaged over five runs, while the remaining results are taken directly from the original TabSYN paper.

Method	Adult	Default	Shoppers	Magic	Beijing	News
	AUC ↑	AUC ↑	AUC ↑	AUC ↑	RMSE ↓	RMSE ↓
Real	.927 (.000)	.770 (.005)	.926 (.001)	.946 (.001)	.423 (.003)	.842 (.002)
SMOTE	.899 (.007)	.741 (.009)	.911 (.012)	.934 (.008)	.593 (.011)	.897 (.036)
CTGAN	.886 (.002)	.696 (.005)	.875 (.009)	.855 (.006)	.902 (.019)	.880 (.016)
TVAE	.878 (.004)	.724 (.005)	.871 (.006)	.887 (.003)	.770 (.011)	1.01 (.016)
GReaT	.913 (.003)	.755 (.006)	.902 (.005)	.888 (.008)	.653 (.013)	OOM
STaSy	.906 (.001)	.752 (.006)	.914 (.005)	.934 (.003)	.656 (.014)	.871 (.002)
CoDi	.871 (.006)	.525 (.006)	.865 (.006)	.932 (.003)	.818 (.021)	1.21 (.005)
TabDDPM	.907 (.001)	.758 (.004)	.918 (.005)	.935 (.003)	.592 (.011)	4.86 (3.04)
TabSYN	0.908 (.002)	0.762 (.001)	0.916 (.001)	**0.938 (.001)**	0.636 (.005)	0.882 (.022)
TabCL (Ours)	**0.911 (.001)**	**0.765 (.002)**	**0.919 (.003)**	**0.938 (.003)**	**0.531 (.003)**	**0.846 (.001)**

TabCL이 단순히 특정 분포에만 맞춰진 모델이 아니라, 다양한 데이터 분포와 과제 타입에 두루 적용할 수 있는 합성 데이터 생성 능력을 갖췄다는 점을 보여준다. 결과적으로, 위 세 가지 평가 항목을 통해 TabCL은 단순한 통계적 유사성을 넘어서 실제 예측 모델 학습에 효과적으로 활용될 수 있는 고품질 합성 데이터를 만들어냈으며, 데이터 부족, 클래스 불균형, 프라이버시 문제 등 다양한 현실 과제를 해결할 수 있는 실용적인 대안임을 입증했다.

4.5 구성 요소별 성능 기여도

본 절에서는 제안하는 TabCL 프레임워크의 핵심 구성 요소인 KL loss clipping과 β-annealing 기법의 효과를 정량적, 정성적으로 검증하기 위해 절제 시험을 수행하였다. 정량적 평가는 4.2절에서 설명한 세 가지 주요 지표를 기준으로, 각 구성 요소를 제거한 모델과 전체 구조를 적용한 모델 간의 성능을 비교하였다. 또한 정성적 평가는 잠재 공간의 구조적 특성과 클래스 분리도를 시각적으로 비교하기 위해, t-SNE 기반의 시각화 결과를 분석하는 방식으로 수행하였다. 이를 통해 정량적 지표뿐만 아니라, 잠재 표현의 품질 측면에서도 각 구성 요소의 기여도를 보다 입체적으로 분석하였다.

<Table 5>는 열 단위 분포 정밀도를 기준으로 잠재 공간의 분포 정렬 성능을 측정한 결과를 나타낸다. 첫 번째 행은 전체 구조를 적용한 TabCL의 성능이며, 두 번째와 세 번째 행은 각각 KL loss clipping과 β-annealing 기법을 제거한 실험 결과이다. 마지막 행은 두 기법을 모두 제거한 경우이다. 실험 결과, KL loss clipping 기법을 제거한 경우 모든 데이터셋에서 오차가 증가하였으며, β-annealing 기법을 제거한 경우에도 Default 데이터셋을 제외한 5개의 데이터셋에서 오차가 상승하였다. 특히 두 기법을 모두 제거한 경우에는 모든 데이터셋에서 뚜렷한 성능의 저하가 일어났으며 평균 오차 역시 크게 증가하였다. 이는 잠재 표현의 정규성과 분포 정렬이 붕괴될 경우, 변수별 통계적 특성이 효과적으로 보존되지 않음을 의미한다.

같은 방식으로, 열 간 상관관계를 비교한 결과를 <Table 6>에 제시하였다. KL loss clipping 기법을 제거한 경우 Default 데이터셋을 제외한 모든 데이터셋에서 오차가 증가했으며, 특히 Magic 데이터셋에서는 성능 저하의 폭이 상대적으로 크게 나타났다. 한편, β-annealing 기법을 제거한 경우에는 모든 데이터셋에서 일관된 성능 저하가 발생하였다. 더욱이, 두 기법을 모두 제거한 경우에는 모든 데이터셋에서 오차가 급격히 증가

하였으며, 특히 Default 데이터셋에서 가장 뚜렷한 성능 저하가 나타났다. 이러한 결과는 잠재 표현의 구조적 정합성이 무너질 경우, 정형 데이터의 핵심 특성인 변수 간 상관관계 정보가 효과적으로 보존되지 못함을 의미하며, KL loss clipping과 β-annealing 두 정규화 기법이 해당 구조를 유지하는 데 실질적인 기여를 했음을 보여준다.

마지막 정량적 실험 결과인 <Table 7>은 생성된 합성 데이터의 활용 가능성을 평가하기 위해, 머신러닝 효율성 지표를 기준으로 수행한 실험 결과를 보여준다. 실험 결과, KL loss clipping과 β-annealing 기법을 순차적으로 제거함에 따라 분류 및 회귀 성능이 전반적으로 저하되는 경향이 나타났으며, 특히 두 기법을 모두 제거한 경우 대부분의 데이터셋에서 큰 폭의 성능 저하가 발생하였다. 반면, KL loss clipping과 β-annealing을 모두 포함한 TabCL 전체 구조를 학습한 모델은 전반적으로 가장 안정적이고 우수한 성능을 유지하였다. 이러한 결과는 정규화 기법이 잠재 표현의 품질을 향상시킬 뿐만 아니라, 생성된 데이터의 실제 활용 가능성을 높이는 데에도 기여했음을 보여준다.

마지막으로, 잠재 표현의 구조적 정렬 상태 및 클래스 간 분리도를 시각적으로 검토하기 위해 수행한 t-SNE 기반 시각화를 수행하였으며, 그 결과를 <Figure 4>에 나타내었다. 해당 시각화는 transformer 인코더에서 전체 입력을 대표하도록 학습된 [CLS] 토큰의 잠재 벡터를 기반으로 하며, 이 벡터는 주로 다운스트림 작업에서 활용되는 핵심 정보로 간주된다. <Figure 4> (a)는 기존 방법론인 TabSYN의 잠재 공간 분포로, 클래스 간 색상 구분이 뚜렷하지 않고 서로 다른 클래스의 샘플들이 전체 공간에 무작위적으로 혼재된 형태를 보였다. 이는 TabSYN이 클래스 정보를 효과적으로 반영한 잠재 구조를 형성하지 못했음을 보여준다. 반면, <Figure 4>(b)는 제안한 TabCL의 잠재 공간 분포로, 표현들이 곡선 형태의 흐름을 따라 정렬되어 있으며, 동일한 클래스의 샘플들이 지역적으로 응집된 양상을 나타냈다. 특히 클래스 간 경계가 명확하게 구분되며, 표현 간 유사성이 잠재 공간 내 위치로 잘 반영되어 있는 것이 확인되었다. 이러한 결과는 제안하는 TabCL이 잠재 표현의 구조화와 클래스 분리에 효과적으로 작용했음을 보여준다.

Table 5. Ablation study of TabCL Based on Column-Wise Distribution Estimation Error. Each row shows the effect of removing a key component (KL loss clipping or -annealing) from the full model. Lower values indicate better alignment between synthetic and real column distributions. Boldface denotes the best performance for each dataset. Values in parentheses represent standard deviations over five independent runs.

Method	Adult	Default	Shoppers	Magic	Beijing	News	Average
TabCL	**0.92 (0.04)**	**0.95 (0.01)**	**1.60 (0.01)**	**0.77 (0.05)**	**1.32 (0.01)**	**1.64 (0.03)**	**1.20**
w/o KL clipping	0.96 (0.12)	1.1 (0.03)	1.75 (0.01)	1.47 (0.02)	1.40 (0.01)	1.83 (0.14)	1.42
w/o β-annealing	1.36 (0.01)	**0.95 (0.03)**	1.61 (0.01)	1.61 (0.04)	1.40 (0.01)	1.66 (0.03)	1.43
w/o KL clipping+ β-annealing	1.59 (0.11)	7.67 (0.04)	1.85 (0.01)	1.64 (0.02)	1.43 (0.01)	6.9 (0.24)	3.50

Table 6. Ablation study of TabCL Based on Pair-Wise Column Correlation Estimation Error. Each row shows the effect of removing a key component (KL loss clipping or β-annealing) from the full model. Lower values indicate better preservation of inter-column correlation structures across both numerical and categorical variables. Boldface denotes the best performance for each dataset. Values in parentheses represent standard deviations over five independent runs.

Method	Adult	Default	Shoppers	Magic	Beijing	News	Average
TabCL	**2.03 (0.04)**	2.91 (0.03)	**2.24 (0.04)**	**0.68 (0.14)**	**2.24 (0.09)**	**1.38 (0.04)**	**1.91**
w/o KL clipping	2.82 (0.01)	2.76 (0.02)	2.67 (0.01)	1.67 (0.12)	2.91 (0.24)	1.44 (0.06)	2.38
w/o β-annealing	2.39 (0.02)	3.03 (0.04)	2.32 (0.04)	3.58 (0.01)	2.85 (0.02)	1.39 (0.03)	2.59
w/o KL clipping+ β-annealing	3.96 (0.04)	8.74 (0.21)	2.9 2(0.04)	2.66 (0.01)	3.09 (0.03)	3.24 (0.04)	4.10

Table 7. Ablation study of TabCL Based on Machine Learning Efficiency. Each row shows the effect of removing a key component (KL loss clipping or β-annealing) from the full model. AUC (for classification) and RMSE (for regression) scores are computed using XGBoost models trained on synthetic data and tested on real data. Boldface denotes the best performance for each dataset. Values in parentheses represent standard deviations over five independent runs.

Method	Adult	Default	Shoppers	Magic	Beijing	News
	AUC ↑	AUC ↑	AUC ↑	AUC ↑	RMSE ↓	RMSE ↓
Real	.927 (.000)	.770 (.005)	.926 (.001)	.946 (.001)	.423 (.003)	.842 (.002)
TabCL	**0.911 (.001)**	**0.765 (.002)**	0.919 (.003)	**0.938 (.003)**	**0.531 (.003)**	**0.846 (.001)**
w/o KL clipping	0.909 (.002)	0.762 (.001)	0.922 (.001)	**0.938 (.004)**	0.583 (.002)	0.862 (.024)
w/o β-annealing	**0.911 (.001)**	0.757 (.002)	**0.927 (.001)**	0.934 (.001)	0.541 (.002)	0.851 (.011)
w/o KL clipping+ β-annealing	0.908 (.002)	0.752 (.001)	0.921 (.001)	0.937 (.004)	0.558 (.002)	0.862 (.024)

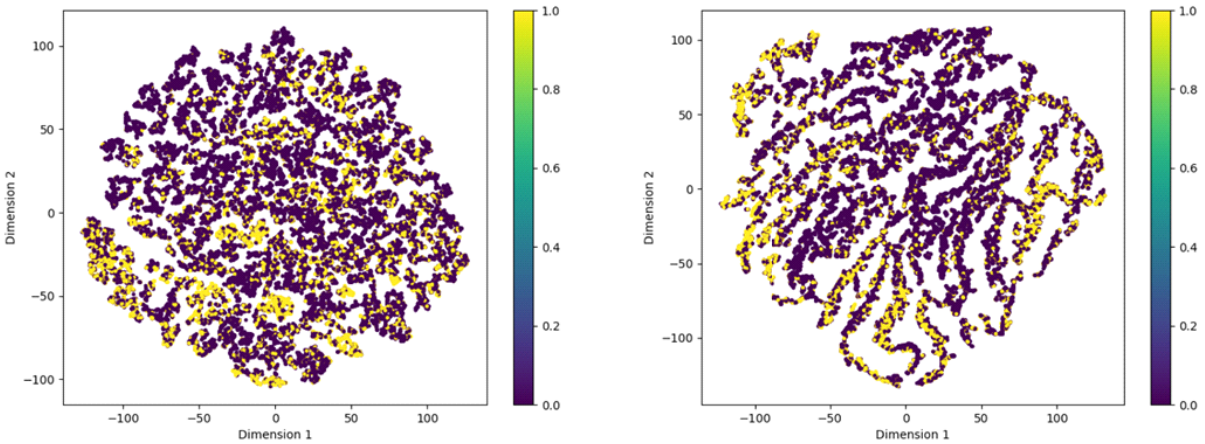

Figure 4. Comparison of Latent Representation Structures using t-SNE on the Adult Dataset. (a) latent representations from TabSYN (baseline) show dispersed and entangled clusters with weak class separation. (b) latent representations from TabCL (proposed) exhibit a clearer structure and stronger class-wise separation, indicating better representation alignment.

5. 결 론

본 연구는 정형 데이터 생성에서 핵심 과제로 꼽히는 변수 간 구조적 이질성, 클래스 불균형, 개인정보 문제를 해결하기 위해, VAE 기반 잠재 표현 위에서 디퓨전 모델을 학습하고 이를 대조 학습과 일관성 정규화로 보완하는 새로운 정형 데이터 생성 프레임워크인 TabCL을 제안하였다. 특히, 기존 VAE 기반 생성 모델의 잠재 표현이 학습 초기 단계에서 구조적으로 불안정하고 클래스 간 구분력이 낮아 디퓨전 모델 학습에 부정적인 영향을 미치는 문제를 극복하고자, 대조 학습을 통해 클래스 간 표현을 명확히 분리하고, 서로 다른 노이즈 조건에서도 표현의 일관성을 유지하는 정규화 전략을 병행하였다. 다양한 정형 데이터셋에 대한 실험 결과, 제안한 TabCL은 열 단위 분포 정밀도, 변수 간 상관관계 보존, 머신러닝 효율의 평가 지표에서 기존 생성 모델보다 일관되게 뛰어난 성능을 보였다. 이는 TabCL이 단변수 및 다변수 수준에서의 통계 구조를 정교하게 복원할 수 있음을 의미하며, 단순히 생성 품질이 우수한 것에 그치지 않고, 실제 예측 모델 학습에도 효과적으로 기여할 수 있는 합성 데이터를 제공한다는 점에서 중요한 의의를 가진다. 특히, TabCL은 연속형 및 범주형 변수가 혼합된 복잡한 정형 데이터 구조를 잠재 공간에서 효과적으로 모델링함으로써, 기존 생성 모델들이 간과하거나 제대로 학습하지 못했던 클래스 간 구조적 구분성과 데이터 내 변수 간 상호작용을 균형 있게 재현하였다. 이러한 구조적 표현력은 실제 분류 및 회귀 과제에서의 성능 향상으로도 이어져, 합성 데이터의 실용성과 신뢰도를 동시에 입증하였다. 또한, 다양한 도메인에 걸쳐 구성된 6개의 공개 데이터셋에 대한 일관된 우수 성능은 TabCL의 도메인 일반화 능력과 모델의 확장 가능성을

보여주는 근거로 작용하며, 향후 다양한 산업 환경에서의 적용 가능성을 뒷받침한다. 향후 연구에서는 산업 현장에서 수집한 대규모·장기 시계열 기반 데이터셋을 활용해 TabCL의 확장성과 일반화 가능성을 보다 심층적으로 검증할 예정이다. 또한, 프라이버시가 중요한 영역이나 소수 클래스 기반 의사결정 등 다양한 응용 시나리오에 TabCL을 적용하여, 실제 활용 사례 중심의 기여를 강화해 나갈 계획이다.

참고문헌

Borisov, V., Sessler, K., Leemann, T., Pawelczyk, M., and Kasneci, G. (2023), Language Models are Realistic Tabular Data Generators, *The Eleventh International Conference on Learning Representations*.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R., and Bengio, S. (2016), Generating Sentences from a Continuous Space, *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 10.

Chawla, N., Bowyer, K., Hall, L. O., and Kegelmeyer, W. P. (2002), SMOTE: Synthetic minority over-sampling technique, Jair.Org, 16, 321-357. http://www.jair.org/index.php/jair/article/view/10302.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020), A simple framework for contrastive learning of visual representations, *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119, 1597-1607. http://proceedings.mlr.press/v119/chen20j.html.

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., and Sun, J. (2017), Generating multi-label discrete patient records using generative adversarial networks, *Machine Learning for Healthcare Conference*, 286-305.

Fonseca, J. and Bacao, F. (2023), Tabular and latent space synthetic data generation: A literature review, *Journal of Big Data*, **10**(1), 1-37. https://doi.org/10.1186/S40537-023-00792-7/FIGURES/5

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014), Generative Adversarial Networks. http://arxiv.org/abs/1406.2661.

Hadsell, R., Chopra, S., and LeCun, Y. (2006), Dimensionality reduction by learning an invariant mapping, *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06),* https://ieeexplore.ieee.org/abstract/document/1640964/.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008), ADASYN: Adaptive synthetic sampling approach for imbalanced learning, *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322-1328.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017), beta-vae: Learning basic visual concepts with a constrained variational framework, *International Conference on Learning Representations*.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models, *Advances in Neural Information Processing Systems*, **33**, 6840-6851.

Hoogeboom, E., Nielsen, D., Jaini, P., Forré, P., and Welling, M. (2021), Argmax flows and multinomial diffusion: Learning categorical distributions, *Advances in Neural Information Processing Systems*, **34**, 12454-12465.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Research, G., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020), Supervised contrastive learning, *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada, https://proceedings.neurips.cc/paperfiles/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

Kim, J., Lee, C., and Park, N. (2023), Stasy: Score-Based Tabular Data Synthesis, *11th International Conference on Learning Representations*, ICLR 2023.

Kingma, D. P. and Welling, M. (2013b), Auto-Encoding Variational Bayes. http://arxiv.org/abs/1312.6114

Kotelnikov, A., Baranchuk, D., Rubachev, I., and Babenko, A. (2023), TabDDPM: Modelling Tabular Data with Diffusion Models, *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202:17564-17579.

Lee, C., Kim, J., and Park, N. (2023), Codi: Co-evolving contrastive diffusion models for mixed-type tabular synthesis, *International Conference on Machine Learning*, 18940-18956.

Lin, X., Xu, C., Yang, M., and Cheng, G. (2024), CTSyn: A Foundational Model for Cross Tabular Data Generation, *The Thirteenth International Conference on Learning Representations*. http://arxiv.org/abs/2406.04619.

Pandey, K., Mukherjee, A., Rai, P., and Kumar, A. (2022), DiffuseVAE: Efficient, Controllable and High-Fidelity Generation from Low-Dimensional Latents. https://arxiv.org/pdf/2201.00308

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022a), High-resolution image synthesis with latent diffusion models, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684-10695.

Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., and Winther, O. (2016), Ladder variational autoencoders, *Advances in Neural Information Processing Systems*, 29.

Song, J., Meng, C., and Ermon, S. (2021), Denoising Diffusion Implicit Models, *9th International Conference on Learning Representations*. https://arxiv.org/pdf/2010.02502.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017), Attention is All You Need, *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, http://arxiv.org/abs/1706.03762.

Xie, Q., Dai, Z., Hovy, E., Luong, T., and Le, Q. (2020), Unsupervised data augmentation for consistency training, *Advances in Neural Information Processing Systems*, **33**, 6256-6268.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019), Modeling tabular data using conditional gan, *Advances in Neural Information Processing Systems 32 (NeurIPS 2019),* https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html.

Zhang, H., Zhang, J., Srinivasan, B., Shen, Z., Qin, X., Faloutso, C., Rangwala, H., and Karypis, G. (2023), Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space, *12th International Conference on Learning Representations*, ICLR 2024. https://arxiv.org/pdf/2310.09656.

저자소개

윤지현 : 고려대학교 산업경영공학부에서 2023년 학사 학위를

취득하고, 고려대학교 산업경영공학과에서 2025년 석사 학위를 취득하였다. 연구 분야는 Tabular Data Sythesis, Generative Models이다.

김성범 : 고려대학교 산업경영공학부 교수로 2009년부터 재직하고 있으며, 인공지능공학연구소 소장, 기업산학연협력센터 센터장, 한국데이터마이닝학회 회장을 역임했다. 미국 University of Texas at Arlington 산업공학과에서 교수를 역임하였으며, 한양대학교 산업공학과에서 학사학위를 미국 Georgia Institute of Technology에서 산업시스템공학 석사 및 박사학위를 취득하였다. 인공지능, 머신러닝, 최적화 방법론을 개발하고 이를 다양한 공학, 자연과학, 사회과학 분야에 응용하는 연구를 수행하고 있다.