핵심 기술 연구 설계서 Synthetic Generator

□ 연구 개요

○ 배경

- AGI(aritifical generative intelligence) 모델 개발을 위해서는 학습을 위한 **다양한 데이터가 필수**적이지만, 이들 데이터를 충분히 확보하는 것은, 매우 어려운 일이다[1].
- **합성데이터는 이들 인공지능 학습의 대안**으로 대두되고 있으며, 현재 다양한 분야에서 모델 학습 및 활용 분야에서 좋은 성능을 보여주고 있다[2~5].

○ 관련 연구

- **TVAEs**[6]
  · Triplet-based Variational Autoencoder는 VAE에 Triplet Loss을 결합하여 데이터의 잠재 공간 표현을 개선하는 모델이며, 특히 민감도가 낮은 테이블 데이터에서 원본의 특성을 더 잘 유지하는 합성데이터를 생성하는 데 효과적임.
- **CT-GAN**[6]
  · Generative adversarial network 방법론을 기반으로 합성데이터를 생성하는 알고리즘으로, 이산형 속성과 연속형 속성을 처리할 수 있는 모델이다. 특히 conditional vector를 사용함으로써 원본 데이터의 특성을 유지할 수 있는 장점이 있음
- **CTAB-GAN**[7]
  · 덜 민감한 데이터에서 혼합형 인코더를 사용하여 범주형, 연속형 변수들의 결측치를 효과적으로 처리할 수 있으며, CT-GAN에 비해 소수 클래스를 효율적으로 처리 가능
- **CTAB-GAN+**[8]
  · CTAB-GAN에서 처리하기 어려운 민감한 개인정보 보호를 보장하면서 고품질의 합성데이터를 생성하기위해 제안된 방법으로 Differential privacy 방법론을 적용하여 데이터의 보안성 강화
- **Tab-DDPM**[9]
  · 이미지 합성데이터 생성에 강력한 알고리즘인 Diffusion 방법론을 정형 테이블 데이터에 사용될 수 있도록 변형한 알고리즘으로, 연속형과 범주형의 변수를 각각의 루틴으로 학습 및 데이터를 생산

❍ **기존 연구의 한계점 및 개선 사항**

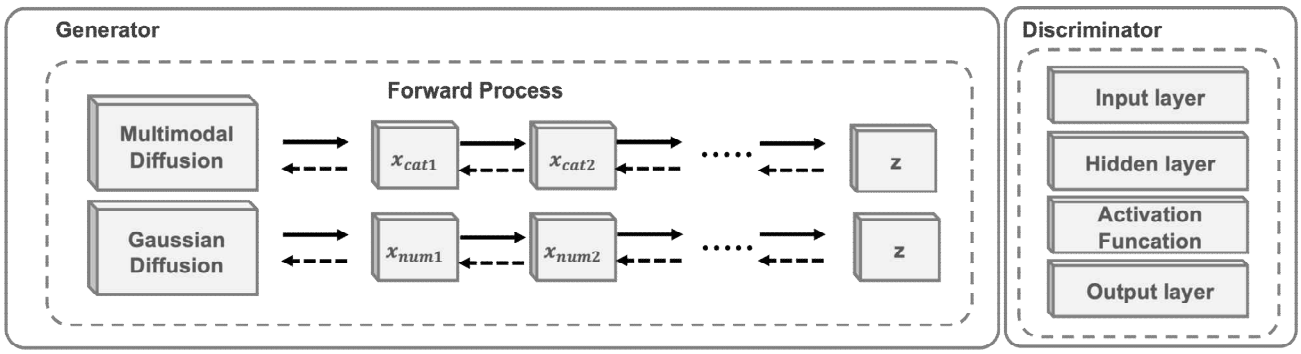
- 기존의 연구들에서는 원본 데이터의 특성을 반영하기 위한 연구들이 대부분 진행되었으며, 이는 학습의 다양성에 있어서 크게 이바지하지 못함
- 이에 Tab-DDPM은 합성데이터의 다양성 확보하기 위해 제안되었지만, 연속형 속성의 Gaussian Diffusion 단독 사용으로 인한 연속형 속성 확장의 한계성 보유
- 특히, ADMET에서 사용되는 주요 Smiles의 RDKit에 포함된 다양한 형태의 수치를 반영하기 위한 합성데이터 생성 방법론 필요

❍ **연구 목표**

- 이에 본 연구에서 수행하고자 하는 Synthetic generator는 Smiles의 RDKit에 포함된 다양한 연속형 변수를 학습할 수 있는 multi-distribution을 제안할 뿐만 아니라, 합성 데이터의 다양성을 위한 Diffusion 기반의 Tabualr Synthetic Generation based on Multi-distribution을 연구 개발한다.

□ 주요 설계

❍ Synthetic Generator 설계도



❍ 수도 코드 (Pesudo code)

```
Input: ADMET Training data TD
Output: ADMET Synthetic Generator SG
function synthetic generator (TD):
  Initialize Generative Model SG
  Initialize Diffusion Encoder
    Initialize Diffusion Decoder
  for number of training iteration do
      for k step do
          Sample mini-batch of noise samples from noise prior
          Sample mini-batch of examples from data generating multi-distribution
          Update the discriminator by ascending its stochastic gradient
      end for
      Sample mini-batch of noise samples from noise prior
      Update SG model parameters by descending its stocahstic gradient
  end for
  return SG
```

□ 2단계 개발 일정

구분	내용	1	2	3	4	5	6	7	8	9	10	11	12
1	Synthetic Generator 구현(Toxicity)												
2	Synthetic Generator 소스코드 검증												
3	Synthetic Generator Toxicity 생산												
4	Synthetic Generator 공개												

□ 참고문헌

1. Foundation model for generalist medical artificial intelligence, Nature, 2023
2. Synthetic data as a proxy for real-world electronic health records in the patient length of stay prediction, Sustainability, 13690, 2023
3. Interpretable data-driven approach based on feature selection methods and GAN -based models for cardiovascular risk prediction in diabetic patients, IEEE Access, 84292-84305, 2024
4. Enhanced diabetes detection and blood glucose prediction using Tiny ML integrated E-nose and breath analysis: A novel approach combining synthetic and real-world data, Bioengineering, 1065, 2024
5. Synthesis of hybrid data consisting of chest radiographs and tabular clinical records using dual generative models for COVID-19 positive cases, J.Imaging Inform, 1217-1227, 2024
6. Modeling tabular data using conditional gan. Neural Inf Process, 2019
7. Ctab-gan: Effective table data synthesizing, Asian Conference on Machine Learning, 2021
8. Ctab-gan+: Enhancing tabular data synthesis.Front, Big Data, 1296508, 2023
9. Tabddpm: Modelling tabular data with diffusion models in International Conference on Machine Learning, 2023