공학석사 학위논문

환자 데이터 생성을 위한 지역 차분 프라이버시가 적용된 적대적 생성 네트워크

LDP-GAN : Generative Adversarial Network with Local Differential Privacy for Patient Data Synthesis

울산대학교 대학원
의 과 학 과
권 한 슬

환자 데이터 생성을 위한 지역 차분 프라이버시가 적용된 적대적 생성 네트워크

지도교수 김 영 학

이 논문을 공학석사학위 논문으로 제출함

2022 년 8 월

울산대학교 대 학 원
의 과 학 과
권 한 슬

권한슬의 공학석사학위 논문을 인준함

심사위원 이 계 화 (인)

심사위원 김 영 학 (인)

심사위원 전 태 준 (인)

울 산 대 학 교 대 학 원

2022 년 8 월

국문 요약

전자의무기록(EMR)은 환자의 건강 상태, 진료결과, 처방 정보 등을 담은 의료 데이터의 일종이다. 환자에 대한 많은 정보를 담고 있어 다양하게 활용될 수 있으며 여러 방면에서 의료의 질을 향상시킬 수 있는 잠재력을 가지고 있다. 특히 최근 큰 발전을 이룬 기계학습(Machine learning)이 의료분야에도 도입됨에 따라 전자의무기록도 활용 도가 높아지고 있다. 그러나 전자의무기록은 환자의 민감한 개인정보를 다수 포함하고 있어 수집, 활용 및 공유가 까다롭다. 이러한 특성은 전자의무기록에 관한 연구를 어렵게 하며 활용도를 떨어뜨린다. 이런 경우 생성모델이 한가지 해결책이 될 수 있다. 생성모델은 실제 데이터를 모방해서 이와 유사한 가짜 데이터를 생성하는 모델을 말한다. 이 생성모델에서 생성된 가짜 데이터를 활용하면 개인정보에 관한 제약을 피할 수 있다.

생성모델에는 다양한 종류가 있지만 최근에는 딥러닝(Deep learning)을 활용한 생성모델이 가장 주목받고 있다. 딥러닝 생성모델은 이미지 분야에서 많은 발전을 이뤘고 사람의 눈으로는 진위를 판별하기 어려운 고해상도의 이미지도 생성할 수 있게 됐다. 딥러닝 생성모델은 의료 데이터에도 적용되었고 임상적으로 유의미한 데이터를 생성할 수 있었다. 딥러닝 생성모델이 좋은 성능을 보이기는 하지만 개인정보 완전하게 해결해 주지는 않는다. 몇몇 연구에서 딥러닝 모델에 대한 공격에 관한 내용이 다루어졌고 모델의 출력 값을 바탕으로 학습데이터를 유추할 수 있음이 밝혀졌다. 이는 딥러닝 생성모델을 사용하는 경우에도 여전히 프라이버시에 대한 위험이 있으며 개인정보 보호 목적을 위해 사용하는 경우라면 모델에 대한 보호가 필요함을 의미한다.

본 연구에서는 멤버십 추론 공격(Membership inference attack)으로부터 안전한 딥러닝 생성모델을 개발하는 것을 목표로 한다. 이 목표를 위해 딥러닝 생성모델 중 하나인 적대적 생성모델 신경망(GAN)을 사용했다. GAN의 한 종류인 WGAN-GP를 기본 모델로서 사용했고 프라이버시 보호를 위해 차분 프라이버시(Differential Privacy)를 접목했다. 차분 프라이버시에서는 수학적으로 디자인된 잡음을 통해 프라이버시를 보호하며 잡음의 강도에 관련된 파라미터인 ε을 사용해 효용성(Utility)과 프라이버시(Privacy) 보호 수준 사이의 Trade-off관계를 조절한다. 이 연구에서는 차분 프라이버시 중에서도 지역 차분 프라이버시를 채택하여 교란된 데이터로만 모델을 학습하는 방식을 개발했다. 교란된 데이터로만 학습을 수행하기 때문에 모델에 대한 공격으로부터 원본 데이터를 강력하게 보호할 수 있다.

이런 방식으로 학습된 모델의 성능은 효용성 측면과 프라이버시 측면으로 나누어서 평가되었다. ε에 따라 두 평가지표 모두 유의미한 변화를 보였으며 두 지표사이의 Trade-off 관계를 적절히 조절하여 최적의 모델을 얻는 것이 가능함을 보였다. 이 실험 결과는 적절한 잡음을 가하면 모델에 대한 공격으로부터 학습 데이터를 보호할 수 있음을 의미한다. 이 연구의 결과를 통해 전자의무기록의 개인정보 문제로 인해 생기는 제약을 어느정도 해결할 수 있을 것으로 예상된다.

중심 단어: EMR, deep-learning, Differential Privacy, GAN, Trade-off

약어 목록

EMR	전자의무기록(Electronic Medical Records)
GAN	적대적 생성 신경망(Generative Adversarial Network)

차례

그림 목차

표 목차

수식 목차

서 론

최근 데이터 분석이 많은 분야에서 중요해짐에 따라 데이터를 저장하기 위한 데이터베이스와 데이터를 분석하기 위한 분석기술들이 급속도로 발전하고 있다. 기업과 기관들은 방대한 데이터를 저장하고 활용하기 시작했고 이로 인해 기존에는 불가능했던 분석들이 가능해졌다. AI가 가장 대표적인 분야라고 할 수 있는데 그중 특히 딥러닝(Deep learning)이 빅데이터의 혜택을 많이 받았다. 딥러닝은 분류 문제, 예측 문제, 이미지 처리, 자연어처리, 강화학습 등 다양한 분야에 적용 가능하며 데이터에서 스스로 규칙을 학습한다. 의료 분야는 딥러닝이 적용되기 좋은 분야 중 하나이다. 환자는 늘어나는데 반해 의료 자원은 한정적이어서 이를 효율화 하기위한 AI의 도입이 반드시 필요하기도 하고 의료기기의 발전으로 의료데이터의 양과 퀄리티가 증가하고 있기 때문이다. 수많은 연구에서 딥러닝이 특정 문제를 해결함에 있어선 전문의 수준의 판단이 가능함을 보였고 점점 영역을 확장하고 있다. 하지만 딥러닝이 효과적으로 의료영역에서 적용되기 위해서 해결해야 할 문제가 몇가지 남아있다. 현재 가장 중요한 문제는 환자 데이터의 개인정보 보안에 관한 문제로 이는 의료 데이터를 다루는데 있어서 가장 복잡한 문제이다[1]. 개인의 의료정보는 프라이버시가 굉장히 민감해서 데이터를 수집하고 활용하는데 많은 법적, 제도적 제약이 따른다. 데이터를 수집하려면 반드시 환자의 동의를 구해야 하고 사용할 때는 일정수준 이상의 보안이 확보되어야 한다. 이러한 법적, 제도적 조치를 뒷받침하는 여러 privacy 모델들이 연구됐는데 유명한 privacy 모델로는 k-익명성 [2], l-다양성 [3], t-접근성 [4]으로 대표되는 비식별화 조치가 있다. 이 모델에서는 식별자를 제거하고 준식별자를 변형, 조작하여 프라이버시를 보호한다. 다만 K-익명성은 Homogeneity attack, Background knowledge attack에 취약하고[3], l-다양성은 Skewness attack, Similarity attack 등에 취약하다[4].

이런 모든 사항을 전부 만족시키더라도 데이터 활용에 제약이 걸리는 경우가 많다. 의료기관간 데이터를 공유하는 경우가 대표적이다. 개인정보 보안에 관한 모든 조건을 충족해도 기관의 규정으로 환자데이터를 외부로 반출 못하게 막는 경우가 많다. 이러한 조치때문에 다기관의 데이터를 동시에 활용하는 일은 대부분의 경우에서 불가능하거나, 복잡하고 시간이 많이 소모된다. 그러나 한 기관의 데이터로만 수행된 연구는 그지역의 특성에만 맞는 결과일수도 있으며 보편적인 연구를 수행하기 위해선 여러 기관의 데이터를 함께 연구하는 것이 반드시 필요하다. 생성모델은 의료데이터를 활용하는데 있어서 발생하는 이런 종류의 문제를 해결할 수 있는 좋은 대안이 될 수 있다. 생성모델에서 생성된 데이터는 실제로 존재하지 않는 가상의 환자의 데이터로 개인정보에 관한 많은 제약으로부터 자유롭다. 또한 합성 데이터(Synthetic data)는 환자로부터 수집 절차를 거쳐 데이터베이스에 저장된 정식 데이터가 아닌 단순한 함수의 출력 값이므로 기관에 완전히 종속되지 않을 수 있고 여러 규정에서 상대적으로 자유롭다. 따라서 합성 데이터를 활용하는 것은 의료 데이터가 가지는 여러 제약을 벗어나 데이터의 공유를 원활하게 할 수 있다.

최근에 가장 주목받는 생성모델은 딥러닝에 기반한 방식인데 우리의 연구에서도 이 방식을 사용한다. 개인정보 보안 문제는 데이터 자체에 대한 위협도 존재하지만 딥러닝 모델에 대한 위협도 존재한다. 몇몇 연구결과에 따르면 모델의 출력 값을 바탕으로 모델의 학습에 사용된 데이터를 복원하거나[4] 데이터가 훈련에 사용된 여부를 유추할 수 있음이 밝혀졌다[6]. 이런 방식으로 모델에 학습된 데이터를 유추하는 행위를 Membership inference attack이라고 한다. Membership inference attack은 훈련에 사용된 데이터의 프라이버시에 위협이 될 수 있어 의료 데이터 등을 학습에 사용할 때는 적절한 대비가 필요하다. 위에서 언급한 딥러닝 생성모델도 이런 Attack에서 자유로울 수 없다. 특히 생성모델은 실제와 유사한 데이터를 직접 생성하므로 훈련에 사용된 데이터가 재구축(Reconstruct)되지 않도록 해야 한

다. 순수한 딥러닝 생성모델에선 반복적인 시도를 통해서 훈련 데이터를 복원하거나 특정 데이터가 훈련에 사용된 여부를 쉽게 구별할 수 있다. 따라서 생성모델을 설계함에 있어서 이러한 공격으로부터 데이터를 보호할 수 있는 조치가 필요하다. 기존의 프라이버시 생성모델들은 학습과정에서 Gradient에 잡음(Noise)을 가하거나 분포를 교란하는 방식으로 프라이버시를 보호했다. 그러나 이러한 방식에서는 데이터의 세밀한 특성까지 고려하기 어려워 항목(Feature)별로 서로 다른 특성을 가지는 EMR에 적절하지 않을 수 있다.

우리는 이 연구에서 이러한 개인정보와 데이터 특성을 효율적으로 고려하는 생성모델인 LDP-GAN(Local Differential Privacy – Generative Adversarial Network)을 제안한다. GAN(Generative Adversarial Network)[7]은 강력한 성능을 제공하는 딥러닝 생성모델이며 우리가 제안하는 LDP-GAN도 이 GAN을 기반으로 한다. GAN은 생성자(Generator)와 판별자(Discriminator), 두 모델로 이루어져 있는데 두 모델을 적대적으로 학습시켜 정교한 생성자를 얻는 것을 목표로 한다. 이 연구에서는 GAN에 Wasserstein distance를 접목시킨 WGAN-GP(Wasserstein Generative Adversarial Network – Gradient Penalty)[8]를 사용한다. 또한 훈련 데이터의 프라이버시를 보호하기 위한 장치로 차분 프라이버시(Differential Privacy)[9]라는 개념을 사용한다. 차분 프라이버시에선 데이터 베이스에서 쿼리에 대한 응답에 수학적으로 정교하게 디자인된 잡음을 가해 개인정보를 차등적으로 보호하는 기법을 말한다. 데이터베이스에서 나온 개념인 차분 프라이버시는 Abadi et el.[10]의 연구에서 딥러닝에 적용될 수 있음이 보였고, Xie et al. [11]의 연구에서 GAN에 적용됐다. 기본적으로 딥러닝에 차분 프라이버시가 적용될 때는 Gradient에 잡음을 적용하는 전역 차분 프라이버시(Global Differential Privacy)의 형태를 띄는데 우리가 제안하는 모델에서는 데이터 자체에 잡음을 가하는 지역 차분 프라이버시(Local Differential Privacy)를 사용했고, 이에 따라 LDP-GAN이라 명명했다. 데이터에 직접에 잡음을 가하는 방식은 데이터의 세세한 특성을 고려한 잡음을 디자인할 수 있다는 장점을 갖는다. LDP-GAN에서 모델은 Ambient-GAN[12]의 구조를 통해 교란된 데이터로부터 실제 데이터 분포를 유추하는 방식으로 학습한다. 이 학습과정에선 실제 데이터를 사용하지 않고 오직 교란된 데이터로만 학습하므로 실제 데이터를 재구축하거나 데이터가 훈련에 사용된 여부를 쉽게 알 수 없다. 이러한 학습 매커니즘(Mechanism)에 의해 훈련에 사용된 데이터의 프라이버시를 강력하게 보호하면서 실제 데이터의 임상적 특성을 잘 반영하는 가상의 데이터를 얻게 된다. 실험 결과는 우리의 이론이 실제와 일치함을 증명했고 기존 모델 대비 몇몇 부분에서 뛰어남을 보였다.

용어 정의

Gradient : 딥러닝(Deep-learning) 학습 과정에서 손실함수에 대한 기울기. 딥러닝의 학습은 손실함수의 기울기가 작아지는 방향으로 가중치를 조절하는 것을 말한다.

합성 데이터(Synthetic data) : 생성모델에서 생성된 가짜 데이터.

과적합(Overfitting) : 기계학습에서 훈련 데이터를 과도하게 학습하는 것을 뜻한다. 보통 기계학습 모델에서 실제 성능을 떨어뜨려서 문제가 된다. 이 연구에서 과적합은 훈련 데이터의 보안성에 관한 문제를 야기한다고 판단함.

잡음(Noise) : 원래의미는 “필요한 신호에 섞여 신호를 바꾸어 버리는 전기적인 장애 또는 잘못된 부호.”등을 의미하나 이 연구에서 잡음은 특정분포에서 추출된 의미 없는 숫자들을 뜻함.

차분 프라이버시(Differential Privacy, DP) : 프라이버시를 정량적으로 모델화하여 프라이버시 보호정도를 측정할 수 있는 기술.

전역 차분 프라이버시(Global Differential Privacy, GDP) : 차분 프라이버시의 한 종류로 쿼리에 대한 응답에 잡음을 가해 정보를 보호. 이 연구에서는 딥러닝 학습과정에서 Gradient에 잡음을 가하는 방식을 전역 차분 프라이버시로 분류.

지역 차분 프라이버시(Local Differential Privacy, LDP) : 차분 프라이버시의 한 종류로 데이터 수집과정에서 데이터에 잡음을 가해 정보를 보호. 이 연구에서는 학습 데이터에 잡음을 가하여 학습하는 방식을 지역 차분 프라이버시로 분류.

매커니즘(Mechanism) : 차분 프라이버시의 매커니즘. 이 연구에서는 데이터나 학습과정을 교란하는 방식을 뜻함. 잡음을 가하여 데이터를 교란할 때는 잡음을 추출하는 분포를 의미함.

효용성(Utility) : 생성된 데이터의 데이터로써 가치. 이 연구에서는 실제 데이터와 유사한 정도로 효용성을 판단함.

판별자 스코어 : 데이터를 판별자에 입력으로 넣었을 때의 출력 값.

Attack : 훈련 데이터를 추론하려는 악의적인 시도. 사용하는 정보의 양에 따라 Black-box Attack, White-box Attack등으로 분류.

관련 연구

1. 비 차분 프라이버시 모델

Mukherjee et al. [15]의 연구에서는 데이터셋을 여러 개로 나누어 각각의 독립적인 GAN을 학습시킨다. 각각의 GAN은 서로 다른 분포를 학습하게 되는데 판별자 D_p는 이 분포들을 식별하도록 학습된다. 각 생성자에서 생성된 데이터들은 D_p에 입력으로 들어가고 D_p는 생성된 데이터가 몇 번째 생성자에서 생성된 건지를 판단한다. 이후 생성자를 학습할 때는 해당 생성자의 레이블이 아닌 잘못된 레이블을 D_p에 넣고 학습시킨다. 예를 들어 1번 생성자를 학습시킬 때는 D_p에 2번 레이블을 주고 학습을 시키면 1번 GAN 모델은 1번 데이터셋의 분포를 학습하지만 D_p에 의해 2번 분포를 학습하도록 조정 받는다. D_p에 의해 1번 GAN은 1번 데이터셋과 2번 데이터셋의 분포가 타협된 중간 분포를 습하여 1번 데이터셋에 과적합(Overfitting)하는 것을 막는다. 보통 효용성과 프라이버시는 Trade-off 관계에 있는데 Mukherjee et al.의 연구에선 Lambda 값을 조절하여 이 관계를 조절한다. 이 모델은 학습 시 데이터를 여러 개로 나누어야 하는데 충분한 양의 데이터가 없어서 데이터를 나누기 어려운 경우에는 적절하지 못하다. 의료 데이터의 경우 특정 질병에 대한 데이터는 희소해서 데이터셋 전체에서 오직 몇 건만 존재하는 경우가 많은데 이러한 경우 데이터를 나눠서 학습하면 기대하는 결과를 얻기 힘들 수 있다.

Liu et al. [16]의 연구는 추가적인 모듈을 사용하여 Facial 이미지에 대한 프라이버시를 보장한다. Generator와 Discriminator외에 Verificator, Regulator를 추가적으로 사용하여 훈련 데이터와 같은 데이터가 생성되는 것을 방지한다. Generator는 실제 이미지 X를 입력받고 $\hat{X}$을 아웃풋으로 생성한다. Discriminator는 $\hat{X}$이 실제 이미지 분포에 머물도록 하며 Verificator와 Regulator는 X와 $\hat{X}$ 이 서로 비슷하지만 다르도록 한다. Verificator는 Siamese neural networks[17]구조로 Contrastive loss[18]로 Pre-trained 되어 X와 $\hat{X}$이 최대한 멀어지도록 한다. Regulator는 SSIM[19]를 통해 X와 $\hat{X}$이 픽셀 레벨에서 같아지도록 한다. 이 세가지 모듈을 통해 생성자는 입력 이미지와 비슷하지만 식별은 불가능한 이미지를 생성하게 된다. 이러한 방식은 이미지에 특화되어 EMR같은 Tabular data에는 적용되기 힘들다. 또한 입력으로 실제 데이터가 들어가야 하므로 다양한 데이터를 무한정 생성하기 어렵고 프라이버시 수준을 수치화 하거나 조절하기 어렵다. 차분 프라이버시를 사용하는 경우는 이런 부분을 수치화해서 조절할 수 있으며 다음 Subsection에서는 이런 모델들에 관해 설명한다.

2. 차분 프라이버시 모델

Song et al. [20]의 연구에선 딥러닝에 차분 프라이버시를 적용할 수 있음을 보였다. 대표적인 딥러닝 학습과정에서 Gradient에 잡음(Noise)을 가해 학습에 사용된 데이터의 프라이버시를 보호한다. Xu et al. [21] 은 이러한 개념을 GAN 모델에 확장 적용했다. 학습과정에서 판별자의 Gradient에 적절한 잡음을 가해 Differentially private한 생성모델을 학습시킨다. 이때 잡음의 강도를 조절해 효용성과 프라이버시 사이의 관계를 조절한다. Beaulieu-Jones et al. [22]의 연구에서도 비슷한 방법을 택한다. 이 연구에서는 AC-GAN[23]에서 판별자의 Gradient에 잡음을 가해 Clinical data의 프라이버시를 보존했다.

이런 연구에서 보이듯 대부분의 차분 프라이버시를 활용한 딥러닝 연구에는 Gradient에 잡음을 가하는 전역 차분 프라이버시(Global Differential Privacy)의 형태를 사용한다. 우리

는 이번 연구에서 데이터 자체에 잡음을 가하는 지역 차분 프라이버시(Local Differential Privacy)를 활용해 프라이버시를 보호하는 생성모델을 제안한다. 지역 차분 프라이버시는 학습 과정에서 Gradient에 잡음을 가하는 형식이기 때문에 데이터의 세세한 특성을 고려하기 힘들다. 따라서 본 연구에서는 EMR의 데이터의 특성을 고려하여 데이터에 잡음을 가하는 방식의 생성모델을 제안한다.

방 법

이 Section에서는 LDP-GAN의 Method와 여기에 사용된 몇 가지 개념들에 대해 설명한다. LDP-GAN에는 GAN에 Wasserstein distance를 접목한 WGAN-GP모델을 기본으로 사용한다. WGAN-GP는 기본 GAN이 가지고 있던 Mode collapse같은 문제들을 해결했으며 생성데이터의 품질을 향상시켰다. 이 연구에서도 생성되는 EMR데이터의 효용성을 보장하기 위해 WGAN-GP를 사용했다. 프라이버시를 보호하기 위한 장치로는 차분 프라이버시의 한 종류인 지역 차분 프라이버시를 사용해 데이터에 잡음을 가하고 이 교란된 데이터로부터 GAN 모델을 학습했다. 이 데이터는 ε 값에 따라 ε-differential privacy를 만족한다.

1. GAN/WGAN

Generative Adversarial Networks(GAN)는 이름에서 보이듯 적대적으로 학습하는 비지도 학습의 일종이다. GAN의 목표는 데이터를 학습해 실제 데이터의 분포를 파악하여 실제데이터와 비슷한 데이터를 생성하는 것이다. GAN은 판별자와 생성자를 경쟁적으로 학습시켜 이 목표를 달성한다. 판별자는 실제 데이터와 합성 데이터를 입력으로 받아 두 데이터를 구분하도록 학습한다. 이 이진분류 문제를 학습하기 위해 판별자는 Binary cross entropy를 손실함수로 가지는데, 수식으로 다음과 같이 나타낼 수 있다.

$$\underset{D}{\boldsymbol{maximize}}\ (E_{x\sim p_x}[\log(D(x)] + E_{z\sim p_z}[\log(1-D(G(z))])$$

[수식1 - 판별자의 목적함수]

생성자는 잡음을 입력으로 받아 데이터를 생성하여 이 생성된 데이터를 판별자가 실제 데이터라고 판단하도록 학습된다. 생성자의 학습을 수식으로 나타내면 다음과 같다.

$$\underset{G}{\boldsymbol{minimize}}\left(E_{z\sim p_z}\left[\log\left(1-D(G(z))\right)\right]\right)$$

[수식2 - 생성자의 목적함수]

학습과정이 반복될수록 판별자는 점점 정교해지는 생성데이터를 구별하기 위해 더욱 성능이 좋아져야 하고 생성자는 성능이 좋아지는 판별자를 속이기 위해 더욱 실제와 같은 데이터를 생성해야 한다. 이런 경쟁적 학습과정이 끝나면 뛰어난 성능의 판별자도 속일 수 있는 정교한 생성자를 얻게 된다. GAN의 학습과정은 다음과 같이 수식으로 나타낼 수 있다.

$$\underset{G}{\boldsymbol{min}}\ \underset{D}{\boldsymbol{max}}(E_{x\sim p_x}[\log(D(x)] + E_{z\sim p_z}[\log(1-D(G(z))])$$

[수식3 - GAN의 학습과정]

[수식1]에서 이상적인 판별자는 다음과 같이 나타낼 수 있다.

$$D^*(x) = \frac{p_{data}}{p_{data} + p_{gen}}$$

[수식4 - 최적의 판별자]

이 최적의 판별기를 [수식1]에 대입하면 판별기의 목적은 Jensen-Shannon Divergence을 최대화하는 것과 같아진다. 이 Jensen-Shannon Divergence는 실제 데이터의 분포 P_{data}와 합성 데이터의 분포 P_g이 겹치지 않을 때 기울기가 소실되어 학습이 진행되지 않는다. 그래서 WGAN에선 이를 개선하여 Jensen-Shannon Divergence대신 EMD(Earth Mover Distance) 개념을 도입했다. 두 분포 P_r, P_g의 EMD는 다음과 같이 나타낼 수 있다.

$$EMD(P_r, P_g) = \inf_{\gamma \in \Pi(P_r, P_g)} E_{(x,y)\sim\gamma}\left[||x - y||\right]$$

[수식5 - Earth Mover Distance]

EMD는 두 분포를 같게 하기위해 얼마나 많은 질량을 얼마나 멀리 옮겨야 되는지에 대한 개념이다. 하지만 질량을 옮기는 경우의 수가 너무 많아 바로 사용할 수 없어서 Kantorovich-Rubinstein이론을 이용한다. Kantorovich-Rubinstein 이론을 사용하면 f가 K-Lipschitz 함수일 때 EMD를 다음과 같이 바꿔서 사용할 수 있다.

$$W(P_r, P_\theta) = \sup_{||f||_{L\leq 1}} E_{x\sim P_r}[f(x)] - E_{x\sim P_\theta}[f(x)]$$

[수식 6 - Wasserstein Distance]

이때 f는 판별자와 같은 역할을 하는 Critic이며, WGAN논문에서는 f를 K-Lipschitz 함수로 만들기 위해 가중치를 Clipping하였다. 이 가중치를 Clipping하는 방식을 WGAN-GP에선 Gradient에 Penalty를 주는 방식(Gradient-Penalty, GP)으로 개선했고, 더 좋은 성능을 보였다. 이후 WGAN-GP는 GAN연구의 주류로 자리잡았고 우리의 연구에서도 이 방식을 채택한다.

2. 차분 프라이버시

차분 프라이버시는 잡음을 사용해 정보를 보호하는 프라이버시 모델이다. 차분 프라이버시가 사용되는 상황은 다음과 같다. 데이터 분석가가 데이터베이스에 쿼리를 날리고, 데이터베이스는 쿼리에 대한 응답을 분석가에게 반환한다. 쿼리는 평균이나 총합 등을 묻는 통계적인 질문을 의미한다. 이때 쿼리를 반복적으로 날리면 데이터베이스에 저장된 데이터를 특정하여 유추할 수 있게 된다. 이를 보호하기 위해 차분 프라이버시가 적용될 수 있는데, 차분 프라이버시는 응답에 대해 수학적으로 디자인된 잡음을 섞어 반환한다. 단 하나의 데이터 포인트에 대해서만 다른 두 데이터베이스인 D와 D' 에 대해 차분 프라이버시 공식은 다

음과 같이 나타낼 수 있다.

$$\Pr[A(D) \in S] \leq \exp(\varepsilon) \times \Pr[A(D') \in S] + \delta$$

[수식7 - (ε, δ) - Differential Privacy]

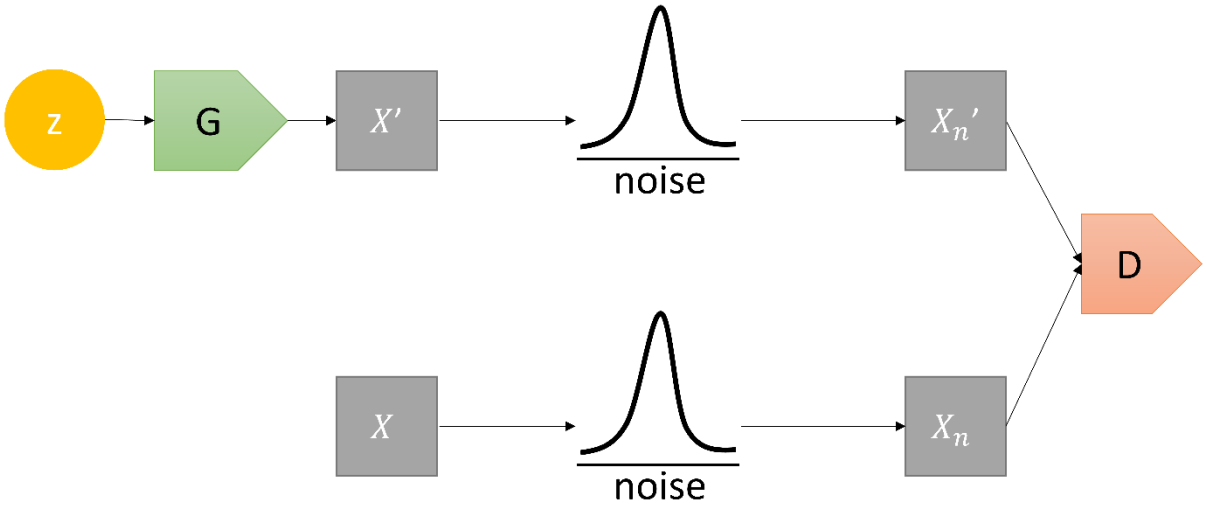
이때 D와 D' 은 하나의 데이터 포인트에서만 차이가 있는 데이터베이스들이며 이를 인접데이터 베이스라고 한다. A는 차분 프라이버시 매커니즘(Mechanism)을 나타내며 S는 Output들의 Subset이라고 볼 수 있다. 위 수식을 만족할 때 매커니즘 Ap는 D와 D' 에 대해서 (ε, δ)-differential privacy를 만족한다고 한다. ε은 Privacy budget을 의미하며 ε 이 큰 경우는 작은 잡음이 가해지고 효용성은 커지는 반면 프라이버시 보호 수준은 감소하게 된다. Δ는 프라이버시가 우연히 유출되는 확률을 의미하는데 보통 데이터베이스 사이즈가 커질수록 데이터가 유출될 확률도 커지므로 보통 δ는 1/d 보다 작게 설정한다. δ가 0일때를 특히 ε-differential privacy라고 하며 log를 사용해 다음과같이 다시 나타낼 수 있다.

$$\left| \log\left(\frac{\Pr[A(D) \in S]}{\Pr[A(D') \in S]}\right) \right| \leq \varepsilon$$

[수식8 - ε - Differential Privacy]

위에서 언급한 차분 프라이버시의 개념은 가장 기본적인 차분 프라이버시의 개념 인 동시에 전역 차분 프라이버시(Global Differential Privacy, GDP)에 해당한다. 지역 차분 프라이버시(Local Differential Privacy, LDP)는 데이터 자체에 잡음을 가하는 방식으로 보통 데이터 수집 주체에 대한 신뢰가 없을 때 사용한다. 우리가 제안하는 방식에서도 LDP를 활용하며 LDP의 대표적인 매커니즘인 Randomized Response[24]와 Laplace mechanism 등을 적용한다. Randomized Response는 가장 대표적인 LDP 매커니즘으로 보통 Binary data에 대해 p의 확률로 진실로 응답하고 1-p의 확률로 반대로 응답하는 방식이다. 확률 p가 다음과 같을 때 Randomized Response mechanism은 ε-differential privacy를 만족한다. Laplace mechanism은 Laplace 분포에서 잡음을 추출하여 데이터나 응답에 가하는 방식으로 LDP와 GDP모두에 적용가능한 매커니즘이다.

3. LDP-GAN



[그림1 - LDP-GAN의 구조도]

LDP-GAN은 GAN을 Membership inference attack으로부터 보호하는 것을 목표로 한다. 이 목적을 위해 LDP-GAN은 컨트롤 가능한 잡음을 원본 데이터에 가하고 이 데이터로부터 학습을 시작한다. 이 교란된 데이터로 일반적인 GAN을 학습시키면 생성되는 데이터도 교란된 상태로 생성되는데 이는 우리가 지향하는 바가 아니다. LDP-GAN에선 교란된 데이터로부터 명확한 데이터의 분포를 학습하기 위해 Ambient-GAN의 구조를 사용했다[12]. Ambient-GAN은 훼손된 Real-world 이미지로부터 깨끗한 이미지를 생성하기 위해 고안되었으나 우리의 연구에서는 실제 데이터를 학습과정에서 숨겨 프라이버시를 보호하기 위해 사용됐다. Ambient-GAN은 이미지가 변형되는 패턴을 알고 있을 때 사용가능한데 논문에서는 Black-pixel, Extraction-patch, Pad-rotate Project등의 방법으로 데이터가 변형되는 방식을 통제했다[12]. 이런 방법은 이미지의 일정 영역에 구멍을 뚫거나 픽셀단위로 변형을 가하는 등 이미지에 최적화된 기법이다. 이러한 픽셀 단위로 변형을 가하는 기법들은 EMR같은 데이터에는 적절하지 않다. 0~255사이의 제한된 값을 가지며 결측값이 없는 이미지 데이터와 달리 EMR은 결측값이 많고, 평균, 분산, 범위 같은 통계량들이 항목별로 큰 차이를 보인다. 또한 EMR은 상관관계(Correlation) 같은 구조적 특성도 중요한데 EMR 데이터에 일관되게 픽셀단위로 변형을 가하는 것은 이런 데이터 특성들을 반영하지 못한다. 그리고 이러한 기법들은 프라이버시를 위해 고안된 방법이 아니어서 프라이버시 수준을 산정하기 어렵다. 따라서 우리가 제안하는 방법은 기존의 Ambient-GAN의 Measurement 방식을 지역 차분 프라이버시로 대체하여 EMR 데이터의 항목별 특성을 고려하면서 프라이버시 수준을 수학적으로 수치화 하여 파라미터 선정이나 매커니즘간 비교를 용이하게 했다.

LDP-GAN의 학습과정은 [그림1 - LDP-GAN의 구조도]와 같다. LDP-GAN을 학습시키기 위해 실제 Raw 데이터에 ε-differential privacy를 만족하는 잡음을 가해 교란된 데이터셋 X_n을 만들어 학습에 사용했다. 생성자는 합성 데이터 X'을 생성하고 X'에 X_n에 가해진 것과 같은 잡음을 더해 X_n'을 만들어준다. 판별자는 X'과 X_n'을 구분하도록 학습된다. 반면 생성자는 X_n'으로 판별자를 속이도록 학습된다. 생성자가 X_n'으로 판별자를 속이기 위해선 X'은 원본 데이터 X와 비슷하게 생성되어야 한다. 이러한 구조는 특별한 특징을 갖는다. 판별자는 교란된 데이터의 분포를 파악하는 반면 생성자는 잡음 너머에 있는 실제 데이터에 가까운 분포를 추론하게 된다. 판별자는 교란된 데이터만 학습하므로 실제 데이터

에 과적합하지 않고 이에 따라 생성자도 정확한 실제데이터의 분포를 파악하기 어려워진다. 다만 추론을 통해 생성자는 판별자보다 실제데이터에 가까운 분포를 학습하게 된다. 학습과정에서 실제 데이터는 학습에 사용되지 않으므로 공격에 대해 안정성을 확보할 수 있고, 특히 판별자는 교란된 데이터만 학습하므로 판별자에서 원본 데이터가 유출될 위험이 적어진다. 반면에 생성자는 실제와 유사한 분포를 추론해 효용성 높은 데이터를 생성한다. LDP-GAN은 [수식 9 - LDP-GAN의 의사코드]의 절차를 통해 학습된다.

Algorithm 1 LDP-GAN

Requires : a_D, learning rate of discriminator; a_G, learning rate of discriminator; M, total number of training data; m, batch size; n_G, number of generator iteration; n_D, number of discriminator iteration per generator iteration; w_G , weights of generator ; w_D , weights of discriminator ; D_l, dimension of latent space; T, train data; H, Holdout data; F_N, noise function; WL, WGAN-GP loss; G, generator; D, Discriminator;

N <- F_N(T)
Initialize w_D, w_G
for *t1* = 1,….., n_G do
  for *t2* = 1, … n_D. do:
    x ~ Sample(N, m)
    z ~ uniform(-1,1, shape = (m, D_l))
    $\hat{x}$ <- F_N(G(z))
    grad <- $\nabla_{w_D}[WL(x,\hat{x})]$
    w_D<- $w_D + a_D \cdot Adam(w_D, grad)$
  end for
  z ~ uniform(-1,1, shape = (m, D_l))
: $\hat{x}$ <- F_N(G(z))
  grad <- ∇_{w_G} D($\hat{x}$)
  w_G<- $w_G + a_G \cdot Adam(w_G, grad)$
end for
return D, G

[수식 9 - LDP-GAN의 의사 코드]

실 험

1. 실험 설정

이 Section에서는 실험에 사용한 설정에 관해 설명한다. 첫번째 Subsection에서 사용한 데이터셋의 Demographic과 특징에 관해 설명한다. 두번째 Subsection에서는 이 실험에서 학습시킨 모델의 종류에 관해 다루고 세번째 Subsection에서는 각 모델의 학습에 사용된 소프트웨어, 파라미터(Parameter)등을 설명한다.

1.1 데이터셋

	AMC (N = 572,811)
Gender ([F,M])	[257160, 315651]
Age (Year)	56.32 ± 14.72
Systolic blood pressure (mmHg)	123.06 ± 12.61 (N = 461,693)
Diastolic blood pressure (mmHg)	74.29 ± 7.94 (N = 461,693)
BMI (kg/m2)	24.11 ± 3.50 (N = 457,621)
CV/CS Encounter(N)	Visits total (N = 571,163)
0	250,160
1	68,037
2	78,406
≥ 3	174,560
Echocardiography	428,004 (74.71%)
Pulmonary function	265,817 (46.40%)
Thallium SPECT	156,615 (27.34%)
Treadmill	68,203 (11.90%)
CT	79,064 (13.80%)
Holter monitoring	46,636 (8.14%)
Six-minute walk test	8,871 (1.54%)
Cardiac rehabilitation	1,990 (0.34%)
Pediatric echocardiography	1,720 (0.30%)

* BMI : Body Mass Index

* CV/CS : Cardiology or Cardiothoracic Surgery Department

* SPECT : Single Photon Emission Computed Tomography

* CT : Computed Tomography

[표1 - 데이터의 Demographics.]

이 연구에서는 Asan Medical Center-Heart Registry[25] 데이터셋이 사용되었고 [표1 - 데이터의 Demographics]는 이 데이터셋의 Demographics를 보여준다. 연구에는 Asan Medical Center-Heart Registry 데이터셋에서 추출된 총 6625명의 환자의 493개의 항목(Feature)이 사용되었다. 이중 6000개의 데이터는 훈련에, 나머지 625개의 데이터는 Holdout 데이터셋으로 분류되어 테스트 등에 사용한다. 데이터 셋은 Basic information, Diagnosis, Lab test, Physical information같은 카테고리의 항목들로 구성된다. 항목 493개는 Binary 51개, Count 60개, Continuous 382개로 이루어져 있다. 결측률은 항목에 따라 0%에서 97.5%사이에 분포하며 항목 평균 43.7%의 결측률을 가진다.

1.2 학습

	레이어종류(크기)	Dropout	활성화 함수	배치 정규화
Layer 1	NN(480	0.3	LeakyReLU(a=0.3)	-
Layer 2	NN(360	0.3	LeakyReLU(a=0.3)	-
Layer 3	NN(240)	0.3	LeakyReLU(a=0.3)	-
Layer 4	NN(120)	0.3	LeakyReLU(a=0.3)	-
Layer 5	NN(60)	0.3	LeakyReLU(a=0.3)	-
Layer 6	NN(30)	0.3	LeakyReLU(a=0.3)	-
Layer 7	Flatten	-	-	-
Layer 8	NN(1)	-	-	-

* NN : Neural Network

[표2 - Critic 모델 구조 요약.]

	레이어종류(크기)	Dropout	활성화 함수	배치 정규화
Layer 1	NN(120)	-	LeakyReLU(a=0.3)	0
Layer 2	NN(240)	-	LeakyReLU(a=0.3)	0
Layer 3	NN(360)	-	LeakyReLU(a=0.3)	0
Layer 4	NN(480)	-	LeakyReLU(a=0.3)	0
Layer 5	NN(493)	-	-	-

* NN : Neural Network

[표3 - Generator 모델 구조 요약.]

분류	파라미터	수치
학습전체	학습률	0.00005
학습전체	최적화 함수	Adam(b1=0, b2=0.9)
학습전체	배치	32
학습전체	스텝수	100000
학습전체	Decay	0.005 per 5000step

생성자	잠재 벡터 차원	100
생성자	잠재 벡터 범위	[-1,1]
판별자	스텝당 학습 횟수	5
손실함수	Lambda	10

[표4 - 학습 파라미터 요약]

두 모델 DP-GAN[21], LDP-GAN을 학습시켜 비교한다. 두 모델 모두 WGAN-GP를 기반으로 사용하며 판별자(Critics)의 구조는 [표2- critics 모델 구조 요약], 생성자(Generator)의 구조는 [표3 - generator 모델 구조 요약]와 같다. 그 외의 세부 학습 파라미터는 [표4 - 학습 파라미터 요약]을 따르며 e을 0.1, 0.25, 0.5, 0.75, 1, 2.5, 5, 7.5, 10, 20 총 10가지로 나누어 학습시킨다.

1.3 소프트웨어

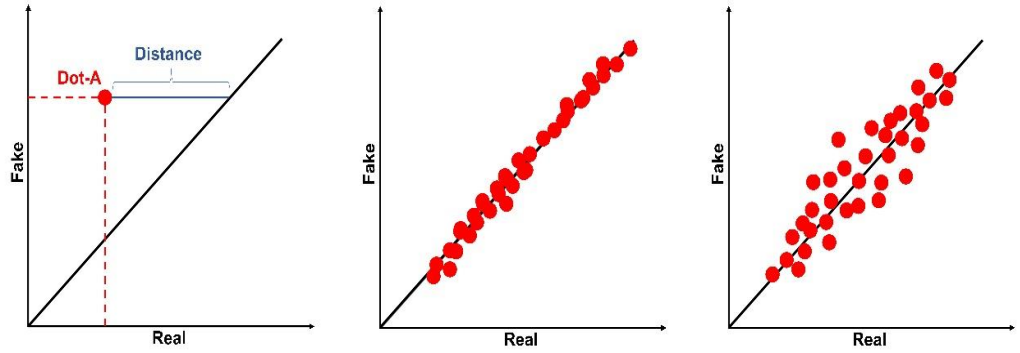
학습과 검증에 사용된 모든 소프트웨어는 Python과 그 라이브러리(Library)를 기반으로 한다. 학습은 Python 3.8.10버전과 Python의 딥러닝 프레임워크(Deep learning Framework)인 Tensorflow 2.6 버전을 기반으로 수행된다. 효용성 평가에서 Dimension Wise Prediction(DWP)를 계산할 때는 파이썬 라이브러리 Scikit-learn 0.24.2 버전의 Decision Tree[26]를 사용한다. Correlation matrix는 Python 라이브러리인 Pandas 1.3.0 통해 구한다. Partial Black-box Attack의 Nelder-mead 알고리즘 [22]과 White-box Gradient Attack의 L-BFGS 알고리즘[28]은 파이썬 라이브러리인 Tensorflow_probability 0.14.1 버전을 사용한다. Partial Black-box Attack의 Nelder-mead 최적화에는 max _iteration 1000을 적용했고, White-box Gradient Attack의 L-BFGS 최적화에는 max_iteration 150을 적용한다.

2. 효용성

이 연구에서 모델들은 효용성과 프라이버시, 두가지 측면에서 평가되었다. 여기서 효용성은 생성된 데이터가 실제 임상 데이터와 특성이 유사한 정도를 나타내는 지표로, 합성 데이터의 가치를 나타낸다. 생성 EMR 데이터의 효용성은 특정 테스트에서 실제 데이터와 유사한 정도로 나타낸다. 이 연구에서 유사도는 산점도 형태로 나타내는 방법과 상관관계를 통해 나타내는 방법을 사용한다.

2.1 산점도

산점도는 항목의 특성에 맞는 스코어를 계산해서 좌표 평면상에 점으로 나타내며 한 개의 점은 한 개의 항목을 나타낸다. 스코어는 항목의 특성이나 파악하려는 정보에 따라 Dimension Wise Statistics(DWS), Dimension Wise Average(DWA), Dimension Wise Prediction(DWP) 등을 적절히 채택하여 사용할 수 있다. 이때 항목-A에 해당하는 점을 Dot-A라고 한다면 실제 데이터의 스코어는 Dot-A의 X좌표, 합성 데이터의 스코어는 Dot-A의 Y좌표가 된다[그림2 - (a)]. 이때 효용성이 가장 높은 경우는 합성 데이터와 실제 데이터의 스코어가 같은 경우이다. 이 경우 Dot-A의 X좌표와 Y좌표가 같아지고 점은 Y=X그래프 위에 찍히게 된다. X좌표와 Y좌표가 다를수록 두 데이터간 차이가 크다는 의미가 되고 합성 데이터가 실제 데이터의 특성을 잘 파악하지 못했음을 의미한다. 이 경우 점은 Y=X직선에서 멀리 찍히게 된다. 즉 점과 Y=X직선 사이의 거리는 실제 데이터와 합성 데이터가 다른 정도를 낸다. 모든 항목에 대해 점을 표시하고 점들과 그래프 사이의 거리를 측정해 효용성을 수치화 할 수 있다. [그림2 - (b)]는 대부분의 점들이 Y=X직선 와 가깝게 분포하여 직선과 평균적



(a) (b) (c)

[그림2 – 실제 데이터와 합성 데이터의 유사도를 비교하는 산점도의 예시. (a) A 항목을 좌표 평면에 나타낸 그림. Distance는 가짜 데이터가 실제 데이터와 다른 정도를 나타낸다. (b) 가능한 모든 항목을 점으로 나타낸 모습. 가짜 데이터가 실제 데이터의 분포를 잘 파악함. (c) 가능한 모든 항목을 점으로 나타낸 모습. 가짜 데이터가 실제 데이터의 분포와 큰 차이를 보임]

인 거리도 작고 분산도 낮은 모습을 보인다. 반면 [그림2 -(c)]는 점들의 분산이 높고 직선과 거리가 먼 점들이 많다. 이 두경우를 비교하면 [그림2-(b)]의 데이터가 [그림2-(c)]의 데이터보다 실제 데이터의 분포를 더 잘 파악했다고 할 수 있다. DWS, DWA, DWP 같은 스코어들은 다음과 같이 산출하여 사용할 수 있다.

2.1.1 Dimension Wise Statistics

Dimension Wise Statistics(DWS)는 진단 같은 binary한 항목들의 positive한 데이터의 비율을 스코어로써 사용한다. 특정 질환이 실제 데이터와 합성 데이터에서 비슷한 빈도로 발생된다면 합성 데이터가 실제 데이터의 분포를 잘 모방했다고 판단하는 근거가 될 수 있다. DWS는 binary 항목에 대해 다음의 공식으로 계산할 수 있다.

$$DWS = \frac{Number\ of\ positive\ patient}{Total\ number\ of\ patient}$$

[수식 10 - Dimension Wise Statistics의 공식]

2.1.2 Dimension Wise Average

Dimension Wise Average(DWA)는 count 항목들의 평균을 스코어로 사용한다. DWS와 마찬가지로 생성모델이 실제 데이터의 분포를 잘 파악했는지 판단하는 지표로 사용된다. Count 항목에 대해 DWA는 다음의 공식으로 계산할 수 있다.

$$DWA = \frac{Column\ sum}{Total\ number\ of\ patient}$$

[수식 11 - Dimension Wise Average의 공식]

2.1.3 Dimension Wise Prediction

Dimension Wise Prediction(DWP)는 각 항목의 예측성능을 스코어로 사용한다. 나머지 항목을 입력으로 사용해 Target 항목을 예측하는 기계학습 모델을 만들어 모델의 예측 성능을 비교하는데 항목의 데이터 타입에 따라 Mean Squared Error, Accuracy, ROC-AUC(Receiver Operating Characteristic - Area Under the Curve)[29] 스코어 등을 사용할 수 있다. 이 테스트는 합성 데이터가 실제 데이터의 내부적 구조를 얼마나 잘 반영했는지를 나타낸다.

2.2 상관관계

합성 데이터의 유효성을 비교하는 다른 방법으로는 데이터의 Correlation matrix를 이용하는 방법이 있다. Correlation matrix는 데이터 항목간 상관관계를 나타내는데 1에 가까울수록 큰 양의 상관관계, -1에 가까울수록 큰 음의 상관관계를 나타내며 0에 가까울 경우 약한 상관관계를 의미한다. 합성 데이터와 실제 데이터의 Correlation matrix의 원소간 차이를 통해 유효성을 평가할 수 있다. 합성 데이터가 실제 데이터와 유사할수록 두 데이터의 상관관계는 비슷해 지고 차이가 작아진다. 이 지표는 다음 수식으로 계산하여 얻는다. 여기서 CM은 Correlation matrix, 차는 Element-wise subtraction을 의미한다.

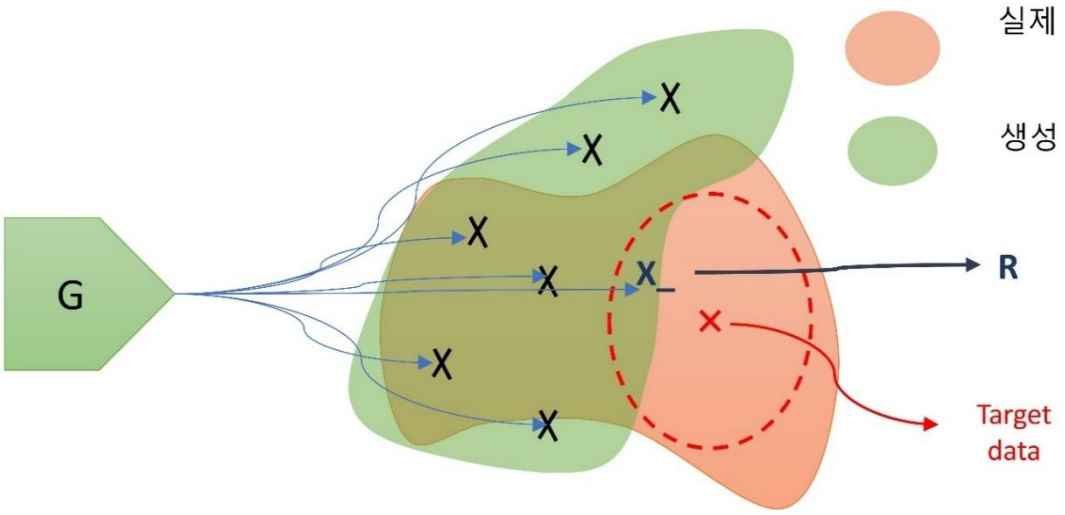
$$Corr = mean(|CM_{real} - CM_{fake}|)$$

[수식 12 - Correlation matrix의 차이를 구하는 공식]

3. 프라이버시

프라이버시는 모델이 정보 유출을 방지하는 능력을 나타내는 지표로써 여러 종류의 Attack을 모델에 가하여 평가한다. 이때 Attacker는 특정 데이터가 모델의 훈련에 사용된 여부를 판단하는데 이러한 Attack을 Membership inference attack이라고 한다. Membership inference attack은 딥러닝 모델에 대해 보편적으로 적용되는 방법으로써 Attacker는 훈련 데이터를 복원하려는 시도를 하며 복원된 데이터와 실제 데이터의 Similarity를 기준으로 해당 데이터가 훈련에 사용됐는지 추론한다. 데이터 중 하나를 Target으로 하고 Target과 합성 데이터 간 Distance를 줄이는 최적화 알고리즘을 통해 생성자에서 Target과 비슷한 데이터가 생성되도록 유도할 수 있다. 이때 이 Distance가 Threshold보다 낮다면 Attacker는 생성자를 통해 Target 데이터가 생성될 수 있다고 추론한다. 이 추론이 정확할수록 모델의 프라이버시 안정성은 낮다고 할 수 있다. Membership inference attack의 시나리오는 Attacker가 접근할 수 있는 정보의 양에 따라 Black-box Attack과 White-box Attack으로 나눌 수 있다. Black-box Attack은 Attacker가 제한된 정보에만 접근이 가능한 상황에서의 Attack이고 White-box Attack은 상대적으로 많은 정보에 접근이 가능한 상황에서의 Attack이다. 이 연구에서는 Black-box Attack을 세분화해서 Full Black-box Attack(FBA), Partial Black-box Attack(PBA)로 나누고 White-box Attack은 정보의 종류에 따라 White-box Gradient Attack(WGA), White-box Discriminator Attack(WDA)으로 나누어 모델을 평가한다[30].

3.1 Full black box attack



[그림 3 – Full Black-box Attack의 도식. 생성자에서 무작위로 다량의 데이터를 생성하여 Target data와 유사한 데이터가 생성되는지 평가한다.]

Full Black-box-Attack은 Attacker가 접근할 수 있는 정보의 양이 가장 제한된 상태에서 적용가능한 Attack이며 [그림 3 - Full Black-box Attack의 도식]은 이 과정을 보여준다. Full-black-box-attack에서 Attacker는 오직 생성된 데이터에만 접근 가능하며 데이터를 무

작위로 생성한 후 Target data와 가장 가까운 데이터를 복원 데이터로 사용한다. 이때 복원 데이터 R은 다음과 같이 나타낸다.

$$R = \underset{\hat{x}}{argmin}\ L(X, \hat{X})$$

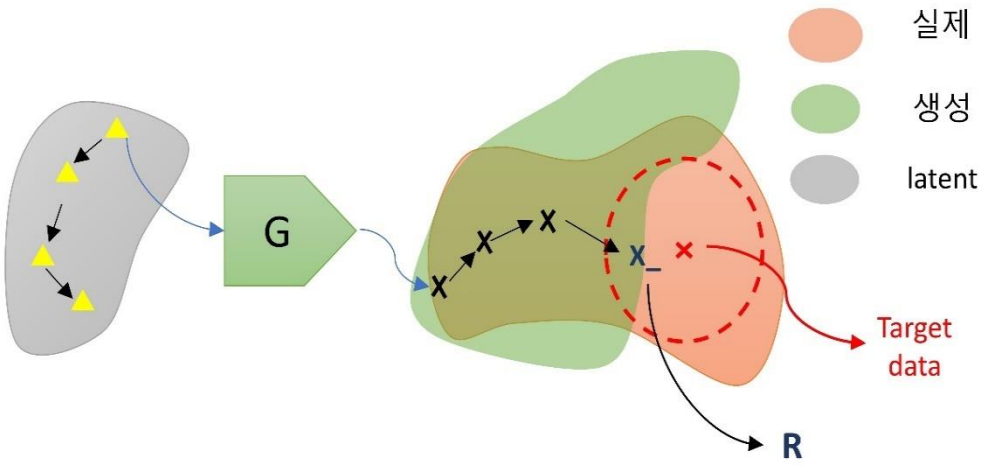
[수식 13 - Full Black-box Attack의 데이터 복원]

이때 L은 Distance function으로 이 연구에서는 유클리디안 거리(Euclidean distance)를 사용한다. Attack은 이 R이 Target X 와의 거리 L(X,R)이 Threshold보다 작으면 X가 모델에 의해 유출됐다고 추론하며, 이 추론의 정확도가 Attack의 성능을 결정한다. Attack의 과정은 [Full Black-box Attack 수식]과 같다. [그림3 - Full Black-box Attack의 도식]은 Full Black-box Attack을 도식화하여 나타낸다. 그림에서 붉은색 영역은 실제 데이터의 분포를 나타나며 초록색 영역은 합성 데이터의 분포를 나타낸다. 붉은색 X는 Target data를 나타내고 검은색 X들은 생성된 샘플들은 의미한다. X_는 X와 가장 가까운 생성 샘플인 R을 나타낸다. 빨간색 점선원은 X의 Threshold 범위를 의미하며 X_가 점선 원 안으로 들어오면 attacker는 X가 훈련 데이터셋에 포함되어 있다고 추론한다. 이 방식은 [수식 14 - Full Black-box Attack의 출력]과 같이 나타낼 수 있다.

$$A(X) = \begin{cases} 1 & (L(X,R) < \text{threshold}) \\ 0 & (L(X,R) < \text{threshold}) \end{cases}$$

[수식 14 - Full Black-box Attack의 출력]

3.2 Partial black box attack



[그림4 – Partial Black-box Attack의 도식. 잠재 벡터를 움직여 합성 데이터를 Target data에 접근시킨다.]

Partial Black-box Attack은 Full Black-box Attack에서 Latent space z에 추가적으로 접근이 가능한 경우에 적용가능한 Attack이다. X와 가장 가까운 R을 얻기 위해 최적의 z를 찾는다. 이 최적의 latent vector z*으로 부터 나온 R은 다음과 같이 나타낼 수 있다.

$$R = G(z^*)$$

[수식 15 - Partial Black-box Attack의 데이터 복원]

이때 z*는 다음과 같다.

$$z^* = \underset{z}{argmin}\, L(X, G(z))$$

[수식 16 - 최적의 잠재 벡터]

R과 X의 거리가 Threshold보다 작아지는 z*을 구할 수 있으면 Attacker는 X가 훈련 데이터 셋에 포함된다고 추론한다. [그림4 - Partial Black-box Attack의 도식]은 Partial Black-box Attack의 과정을 보여준다. 여기서 회색 영역은 잠재 벡터인 z의 영역을 나타내며 Attacker는 이 영역에서 z를 조금씩 움직여 G(z)가 Target data point(빨간색 X)에 최대한 가까워지도록 최적화 과정을 수행한다. 최적화는 Nelder-mead 알고리즘[22]을 통해 수행되

며 이때 N개의 항목을 가진 X에 대한 최적화 과정을 다음과 같이 나타낼 수 있다.

$$\underset{z}{minimize}\ \frac{1}{N}\sum_{i=1}^{N}\sqrt{\left(X_i - G(z)_i\right)^2}$$

[수식 17 - 최적의 잠재 벡터]

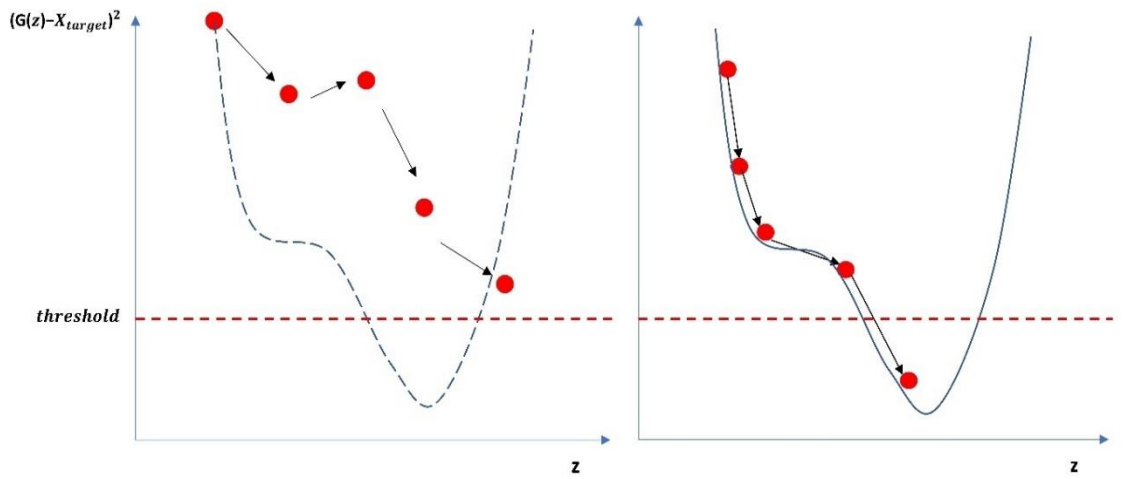
Nelder-mead의 최적화로 얻어진 z*에 대해 X와 G(z*)의 거리가 Threshold이하라면 Attack은 X가 학습에 사용됐다고 추론한다.

$$A(X) = \begin{cases} 1\ (L(X, G(z*)) \leq \text{threshold}) \\ \\ 0\ (L(X, G(z*)) > \text{threshold}) \end{cases}$$

[수식 18 - Partial Black-box Attack의 출력]

3.3 White-box Gradient Attack

White-box Gradient Attack은 Partial Black-box Attack에서 모델의 Gradient를 추가적으로



(a) **(b)**

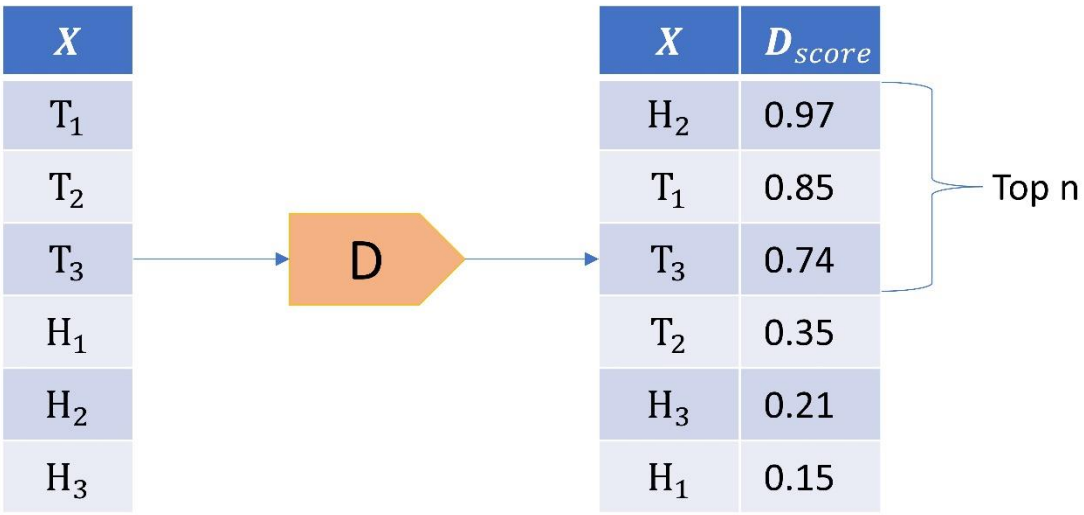
[그림5 – White-box Gradient Attack과 Partial Black-box Attack의 차이. (a) Partial Black-box Attack의 최적화 과정. 그래프는 목적함수의 최적 단면이며 Partial Black-box Attack에서는 이를 알 수 없어 점선으로 표시되었다. (b) White-box Gradient Attack의 최적화 과정. Gradient 정보를 활용할 수 있어 목적함수 그래프가 실선으로 표시되었다.]

활용 가능한 상황에서의 Attack 시나리오다. 전체적인 과정은 Partial Black-box Attack과 같으며 목적함수에 대한 Gradient를 사용하여 최적화를 진행한다. Partial Black-box Attack과 마찬가지로 z를 움직여 목적함수인 L(X, G(z))를 최소화하도록 최적화를 진행한다. 이 결과로 얻어진 z*과 이로부터 생성된 G(z*)와 X 사이의 유클리디안 거리를 기반으로 X가 훈련 데이터인지 추론한다. 전체적인 과정은 [그림4 - Partial Black-box Attack의 도식]과 동일하다. 그러나 White-box Gradient Attack 시나리오에서는 목적함수에 대한 미분이 가능하기 때문에 더 적은 시행착오로 더 정확한 최적화를 가능하게 한다. 두 Attack의 차이는 [그림5- White-box Gradient Attack과 Partial Black-box Attack의 차이]에서 확인할 수 있다. Partial Black-box Attack은 [그림5 - (a)]에서 보듯이 매 스텝마다 시행착오를 반복하며 내려가는 반면 White-box Gradient Attack은 Gradient의 정보를 활용해 최적의 루트로 최적화를 수행한다[그림5 - (b)]. 이 연구에서는 Gradient를 활용한 공격을 위해 L-BGFS 알고리즘[28]을 사용했으며 Attack의 출력은 다음과 같이 나타낼 수 있다.

$$A(X) = \begin{cases} 1 \ (L(X, G(z*)) \leq \text{threshold}) \\ \\ 0 \ (L(X, G(z*)) > \text{threshold}) \end{cases}$$

[수식 19 - White-box Gradient Attack의 출력]

3.4 Discriminator-white-box-attack



[그림6- White-box Discriminator Attack의 예시. 데이터의 판별자 스코어를 기준으로 데이터가 학습에 사용된 여부를 판단한다.]

White-box Discriminator Attack은 Attacker가 판별자에 접근가능한 상황에서 활용가능한 Attack 시나리오다. Black-box Attack이나 White-box Gradient Attack과 다르게 실제 데이터와 생성데이터 사이의 거리를 기반으로 추론을 하는 이 아닌 판별자의 스코어를 활용한다. GAN은 학습 과정에서 실제 데이터와 합성 데이터를 구분하도록 학습되는데 이를 위해 실제 데이터가 입력으로 들어오는 경우는 큰 값을, 합성 데이터가 입력으로 들어오는 경우는 낮은 값을 출력하도록 유도된다. 이상적인 판별자는 훈련 데이터와 Holdout 데이터에 대해 비슷한 수준으로 높은 스코어를 출력하고 합성 데이터에 대해서는 낮은 스코어를 출력하게 된다. 만약 판별자가 훈련 데이터셋에 과적합한 경우 판별자는 훈련 데이터셋에 대해 상대적으로 더 높은 스코어를 출력한다. 이 경우 Attacker는 판별자 스코어를 기반으로 데이터가 훈련에 사용된 여부를 판단할 수 있고 Membership inference attack에 대상이 될 수 있다. Attack은 데이터들의 판별자 스코어를 구하는 것으로 시작한다. 훈련 데이터와 Holdout 데이터를 각각 k개씩 입력으로 넣어 판별자 스코어를 구한 후 이 스코어를 기준으로 데이터들을 내림차순으로 정렬한다. 이후 상위 k개의 데이터를 학습 데이터라고 추론한다. 이 과정은 [그림6- White-box Discriminator Attack의 예시]에 나타나 있으며 이 추론이 정확할수록 모델은 Attack에 취약하다고 할 수 있다.

$$A(X) = \begin{cases} 1\left(X \in Top_k\right) \\ 0\left(X \notin Top_k\right) \end{cases}$$

[수식 20 - White-box Discriminator Attack의 출력]

결 과

1. 효용성 테스트 결과

1.1 Dimension Wise Statistics

[그림7 - Dimension Wise Statistics의 결과]은 모델들의 Dimension Wise Statistics 테스트의 결과를 나타낸다. [그림7-(a)]는 LDP-GAN의 Dimension Wise Statistics의 결과를 나타내고 [그림7-(b)]는 DP-GAN의 Dimension Wise Statistics의 결과를 나타낸다. 그래프의 행을 따라서 ε값의 변화에 따른 결과를 나타낸다. DWS는 보통 두 데이터셋에서 특정 질병을 가진 환자 수 같은 binary 항목의 비율을 비교한다. 환자 비율을 실제 데이터와 얼마나 유사하게 생성하는지를 기준으로 합성 데이터의 유효성을 평가한다. ε은 0.1에서 20까지 변하며 ε이 클수록 약한 잡음을 의미한다. 그래프에서 Distance가 줄어드는 것은 효용성이 증가함을 의미하는데 LDP-GAN은 모든 ε이 증가함에 따라 선형적으로 Distance가 줄어드는 반면 DP-GAN은 ε=2.5까지는 큰 Distance를 보이다 ε=5에서 급격하게 Distance가 줄어드는 모습을 보였다. 또한 DP-GAN은 ε에 따른 선형적 변화가 아닌 진동하는 모습이 나타났다. ε 조절해 효용성을 통제하는 관점에서 보면 LDP-GAN이 더욱 안정적인 모습을 보였다. 전체적으로 LDP-GAN에서 좋은 DWS 스코어를 보였고 최고의 스코어도 LDP-GAN에서 나타났다.

1.2 Dimension Wise Average

Dimension Wise Average 테스트는 Integer 항목에 대해 시행되었으며 [그림8 - Dimension Wise Average의 결과]는 모델들의 테스트 결과를 보여준다. [그림8-(a)]는 LDP-GAN의 결과이며 [그림8-(b)]는 DP-GAN의 결과이다. 세로축은 ε을 의미하며 0.1에서 20까지 변화한다. DWS의 결과와 마찬가지로 LDP-GAN은 ε에 따라 전반적으로 선형적인 변화를 보이는 반면 DP-GAN은 급격한 변화와 진동하는 모습을 보인다. 또한 전체적인 DWA 스코어나 최고 DWA 스코어에서 LDP-GAN이 좋은 결과를 보였고, 이 결과는 DWA 관점에서 LDP-GAN이 더 유효한 데이터를 생성함을 의미한다.

1.3 Dimension Wise Prediction

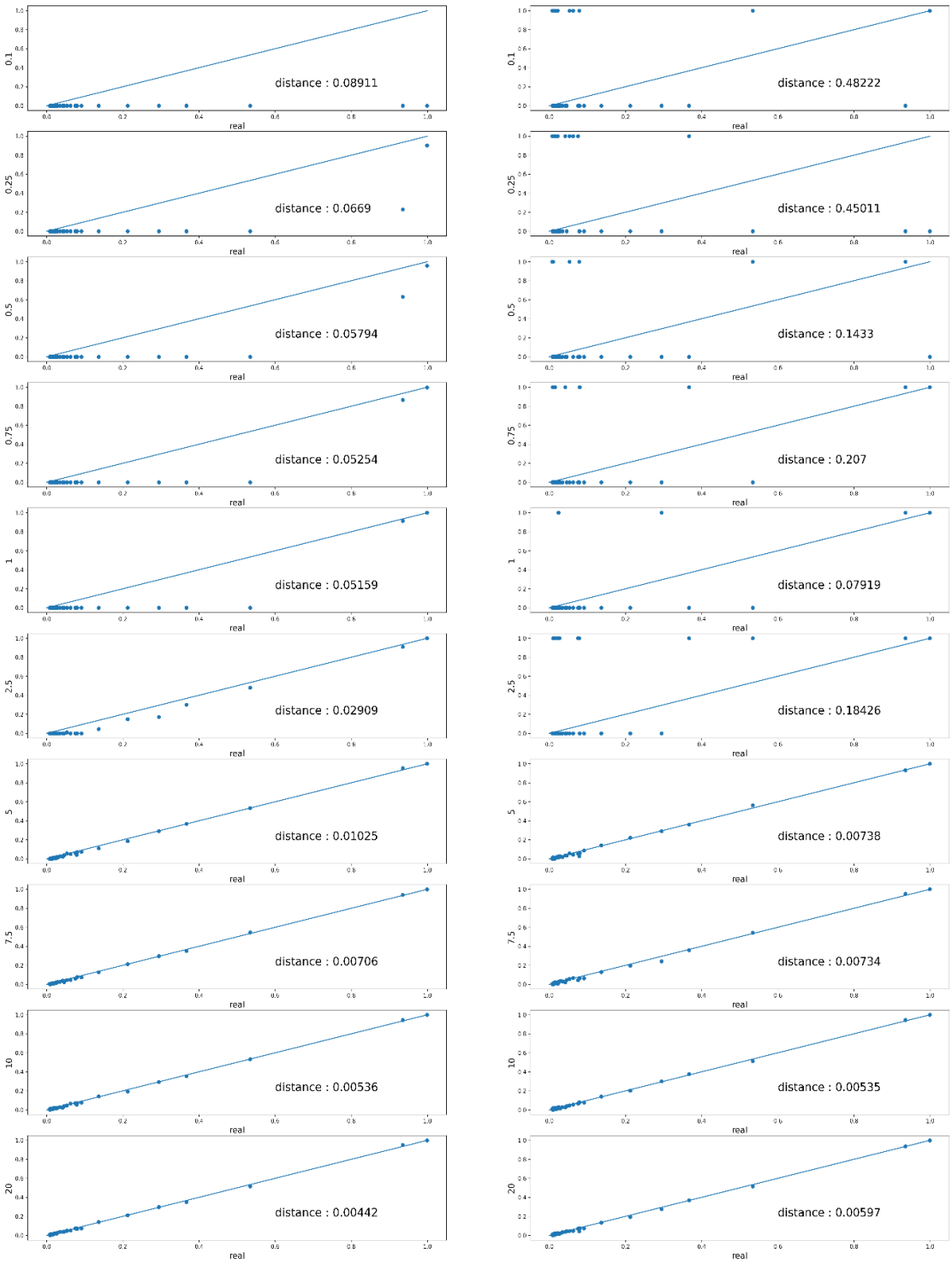
Dimension Wise Prediction 테스트는 EMR데이터의 구조적 특성이 생성데이터에서 잘 반영됐는지 평가한다. 예를 들어 어떤 질병 A를 가지고 있는 환자는 B Lab테스트에서 높은 수치를 띄는 경향이 있다면 합성 데이터에서도 이러한 경향이 반영이 되어야 실제 데이터와의 DWP에서 낮은 Distance를 가질 수 있다. DWP는 기계학습의 결과로써 생성데이터를 기계학습에 사용하고자 하는 경우에는 데이터 퀄리티를 판단하는 가장 직접적인 지표가 될 수 있다.

[그림9 - Dimension Wise Prediction의 결과]는 모델들의 Dimension Wise Prediction으로 각 항목을 기계학습방법으로 예측한 결과이다. 이 연구에서는 Decision Tree로 각 항목을 예측했는데 Count 항목이나 Continuous 항목은 MSE(Mean Squared Error), Binary 항목은 ROC-AUC(Receiver Operating Characteristic - Area Under the Curve)를 스코어로 산점도에 나타냈다. [그림9 - Dimension Wise Prediction의 결과]의 푸른 점은 ROC-AUC를 나타내고

붉은 점은 MSE를 나타낸다. DWP는 전체적으로 DWS나 DWA보다 점들이 더 산개해 있고 대각선과 멀리 떨어진 모습을 보였다. MSE(붉은 점)는 LDP-GAN에선 선형적으로 서서히 직선에 수렴하는 반면 DP-GAN에선 ε=5이전 까지는 수렴하지 못하다가 ε=5에서 급격한 수렴을 보였다. ROC-AUC(푸른 점)은 ε=1일때까지 LDP-GAN과 DP-GAN의 그래프에서 전부 0.5에 해당하는 모습을 보였는데 이는 합성 데이터가 실제 데이터의 특성을 전혀 반영하지 못하며 데이터로서 가치가 낮음을 나타낸다. LDP-GAN은 ε=2.5일 때부터 푸른 점들의 Distance가 단계적으로 감소하는 모습을 보였지만 DP-GAN은 ε=5에서 한번 급격한 변화를 보인 이후에 진동하는 모습을 보였다. 이 두가지 결과들은 LDP-GAN은 ε에 대해 안정적으로 반응함을 의미한다. 또한 전체적인 DWP 스코어와 최고 스코어 모두 LDP-GAN에서 우위를 보였는데, 모든 결과들을 종합하면 LDP-GAN이 더 안정적으로 실제 데이터와 구조가 유사한 데이터를 생성할 수 있음을 의미한다.

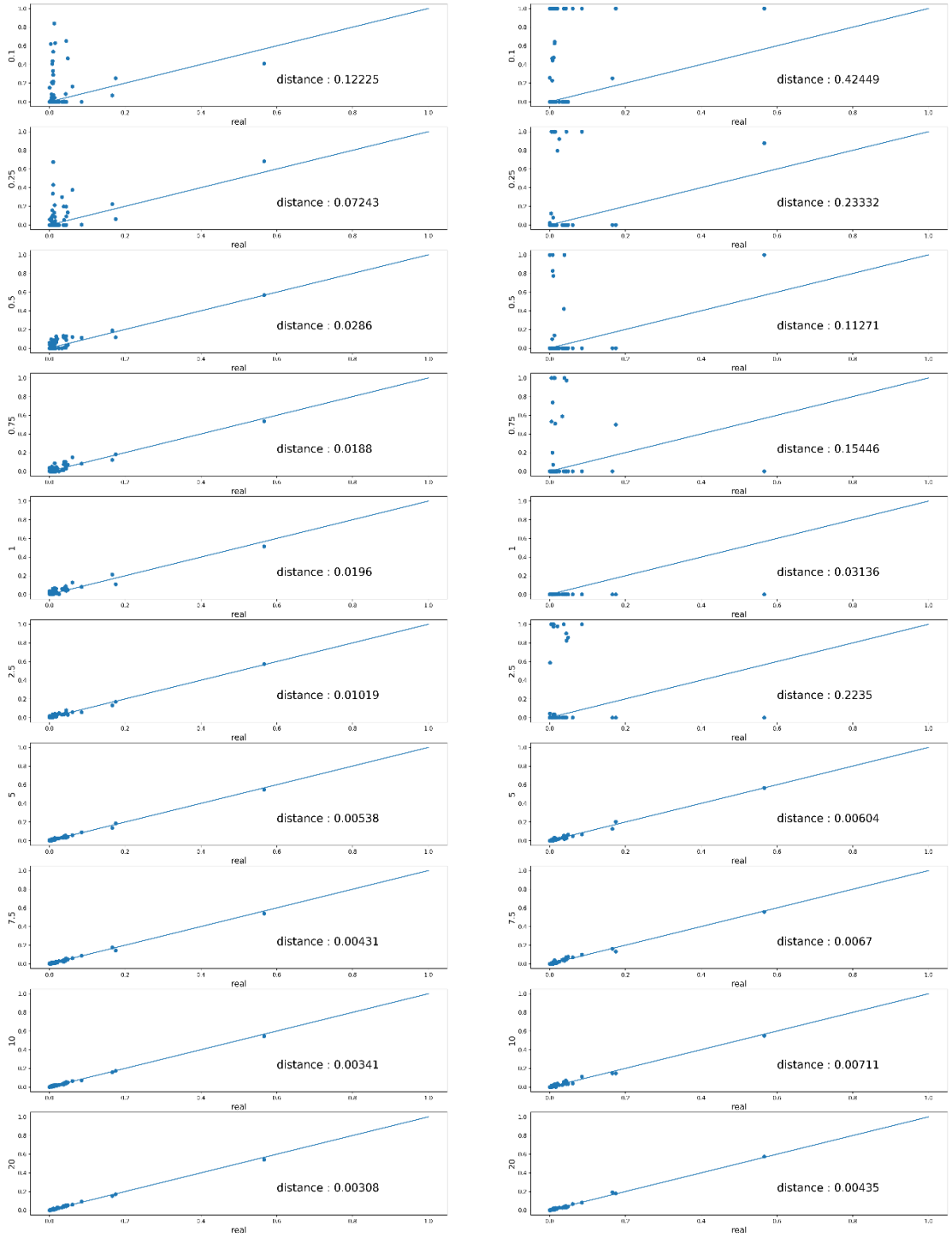
1.4 상관관계

이 Section에서는 DWS, DWA, DWP와 다르게 산점도가 아닌 Correlation matrix를 기준으로 효용성을 평가했다. [표5- LDP-GAN과 DP-GAN에서 생성된 데이터의 상관관계 결과값]는 [수식 12 - Correlation matrix의 차이를 구하는 공식]을 통해 얻은 결과들을 보여준다. 이 수치들은 실제 데이터와 합성 데이터의 차이를 보여주는 값으로 작을수록 유효성이 높다. LDP-GAN은 ε=0.25일 때 한번 상승한 이후 계속해서 수치가 감소했으며 이는 ε을 증가시킬 때 기대하는 결과와 일치한다. DP-GAN은 진동과 급격한 변화를 반복하면서 ε에 대해 불안정한 반응을 보였다. 실험 결과는 LDP-GAN이 더 안정적으로 더 유효한 데이터를 생성함을 증명했다.



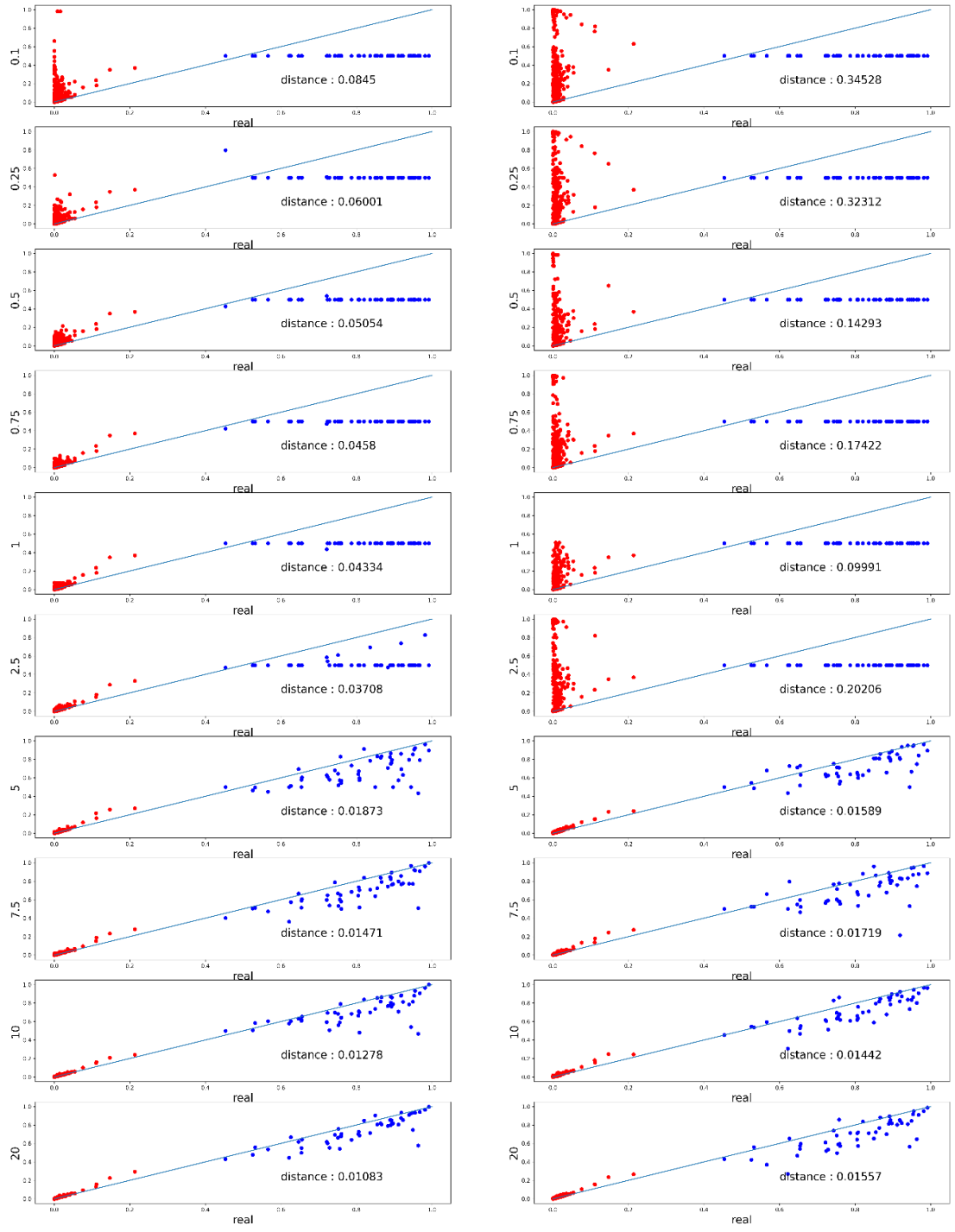
(a) **(b)**

[그림7 - Dimension Wise Statistics의 결과. 세로축은 ε을 나타낸다. (a) LDP-GAN의 DWS 결과. (b) DP-GAN의 DWS 결과.]



(a) **(b)**

[그림8 – Dimension Wise Average의 결과. 세로축은 ε을 나타낸다. (a) LDP-GAN의 Dimension Wise Average 결과. (b) DP-GAN의 Dimension Wise Average 결과.]



(a) **(b)**

[그림9 – Dimension Wise Prediction의 결과. 붉은 점은 MSE, 푸른점은 ROC-AUC의 결과를 나타내고 세로축은 ε을 의미한다. (a) LDP-GAN의 Dimension Wise Prediction의 결과. (b) DP-GAN의 Dimension Wise Prediction의 결과.]

ε	*LDP-GAN*	*DP-GAN*
0.1	0.2291	0.3299
0.25	0.5984	0.4228
0.5	0.3125	0.5013
0.75	0.2164	0.4164
1	0.1708	0.4846
2.5	0.1023	0.2393
5	0.0687	0.0811
7.5	0.0610	0.0778
10	0.0534	0.0787
20	0.0507	0.0651

[표5 - LDP-GAN과 DP-GAN에서 생성된 데이터의 상관관계 결과값. 값은 실제 데이터와 생성데이터의 Correlation matrix간 차이의 절대갑의 평균값을 나타낸다.]

2. 프라이버시 테스트 결과

2.1 Full Black-box Attack

Full Black-box Attack(FBA)은 학습 데이터셋 중 랜덤으로 추출된 1000개의 데이터셋에 대해 수행됐다. 이 Attack은 각각의 추출된 데이터에 대해 500개의 합성 데이터를 무작위로 생성한 후 이중 Target 데이터에 충분히 근접한 데이터가 있는지 여부를 판단했다. 이때 유클리디안 거리가 0.05미만인 경우 모델을 통해 Target 데이터를 재구성할 수 있다고 평가했다.

[그림10 - Full Black-box Attack의 결과]은 1000개의 샘플 Target 데이터 중 Attack에 성공한 데이터의 개수를 보여준다. 왼쪽은 이 연구에서 제안하는 모델인 LDP-GAN의 결과이고 오른쪽은 비교 모델인 DP-GAN의 결과이다. 세로축을 따라 ε의 변화가 나타난다. 바의 길이는 1000개의 샘플 중 Attack에 성공한 Target 데이터의 개수를 나타내는데 바의 길이가 길수록 모델이 Attack에 취약함을 의미한다. LDP-GAN과 DP-GAN 모두 ε=2.5이하에서 0에 가까운 수치를 기록하며 거의 완벽한 프라이버시 수준을 보였다. 특히 LDP-GAN은 이 구간에서 효용성이 증가했던 것을 감안하면 이 구간에 한정해서 프라이버시 손실 없이 효용성을 증가시키는 이상적인 Trade-off관계를 가진다고 볼 수 있다. ε=2.5이후에는 전체적인 수치는 LDP-GAN이 더 높았으나 최고점은 거의 비슷했고 ε에 따른 선형성도 LDP-GAN에서 더 좋은 모습을 보였다.

2.2 Partial Black-box Attack

Partial Black-box Attack(PBA)역시 학습 데이터셋에서 랜덤으로 추출된 1000개의 데이터 샘플들에 대해 수행됐다. Attacker는 Nelder-mead 알고리즘을 사용해 Target X에 대하여 $L(z) = (X - G(z))^2$ 을 줄이는 최적화를 수행했고, L(z)가 0.025미만으로 수렴한다면 X는 학습 데이터셋에 포함된다고 추론했다.

[그림11 - Partial Black-box Attack의 결과]가 Attack의 결과를 나타내며 바의 길이는 1000개의 샘플 데이터 중 L(z)가 0.025밑으로 수렴한 데이터를 나타낸다. Partial Black-box Attack은 LDP-GAN에서 ε=5까지 ε에 비례해서 증가하다가 이후에 다시 감소하는 모습을 보였다. 이 결과는 기대하는 결과와 다른 모습으로 LDP-GAN은 Partial Black-box Attack에서는 불안정함을 의미한다. 또한 전체적인 수치나 최고치도 DP-GAN대비 높아 Partial Black-box Attack에는 상대적으로 취약함을 보였다.

2.3 White-box Gradient Attack

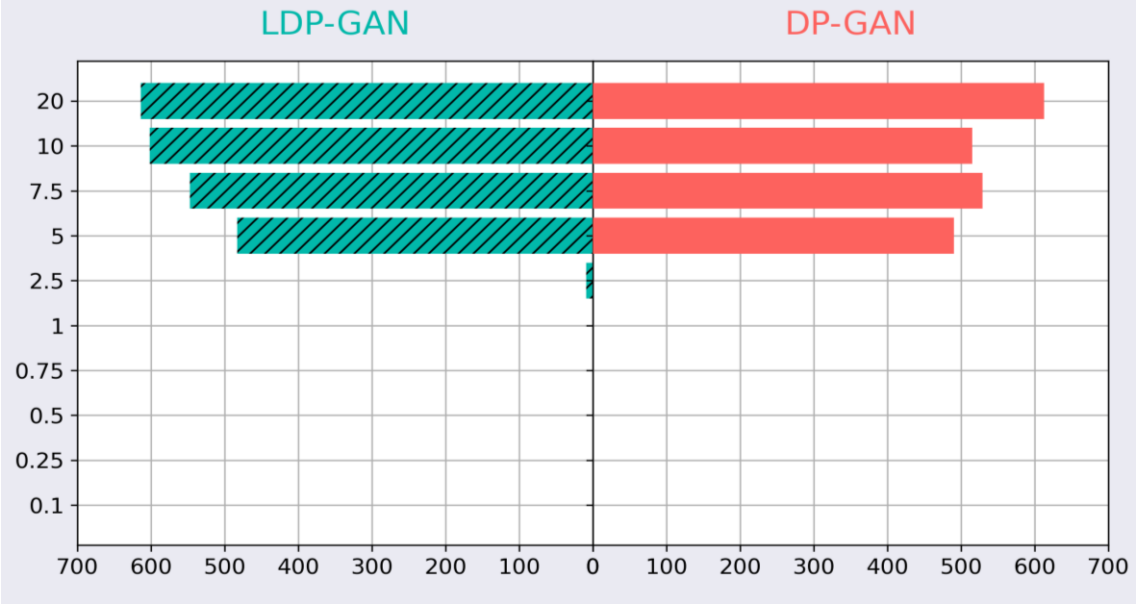
White-box Gradient Attack(WGA)은 Partial Black-box Attack과 비슷하게 진행되었으며, Gradient를 추가적으로 사용해 1000개의 데이터에 대해서 L-BFGS알고리즘으로 최적화를 수행했다. Threshold는 0.005가 사용되었다. Target 데이터로부터 거리가 Threshold미만인 데이터를 생성 가능하면 Attacker는 X가 훈련 데이터에 포함됐다고 주장한다.

[그림12 -White-box Gradient Attack의 결과]는 White-box Gradient Attack의 결과를 보여준다. LDP-GAN은 ε=0.75부터 유효한 Attack이 보였고, DP-GAN은 ε=5부터 유효한 Attack이 보였다. 또한 전체적인 수치와 최고치도 LDP-GAN이 높았다. DP-GAN은 급격한 수치의 변화도 보이지 않으면서 ε에 대해 선형적으로 반응했다. 이런 결과들은 전체적으로 DP-GAN이 Gradient를 활용한 공격에 더 안정적임을 의미한다. 이 결과는 Gradient에 잡음을 가하는 DP-GAN의 학습방식 때문에 Gradient를 이용하여 가하는 Attack에 더 안정적인 것으로 보인다.

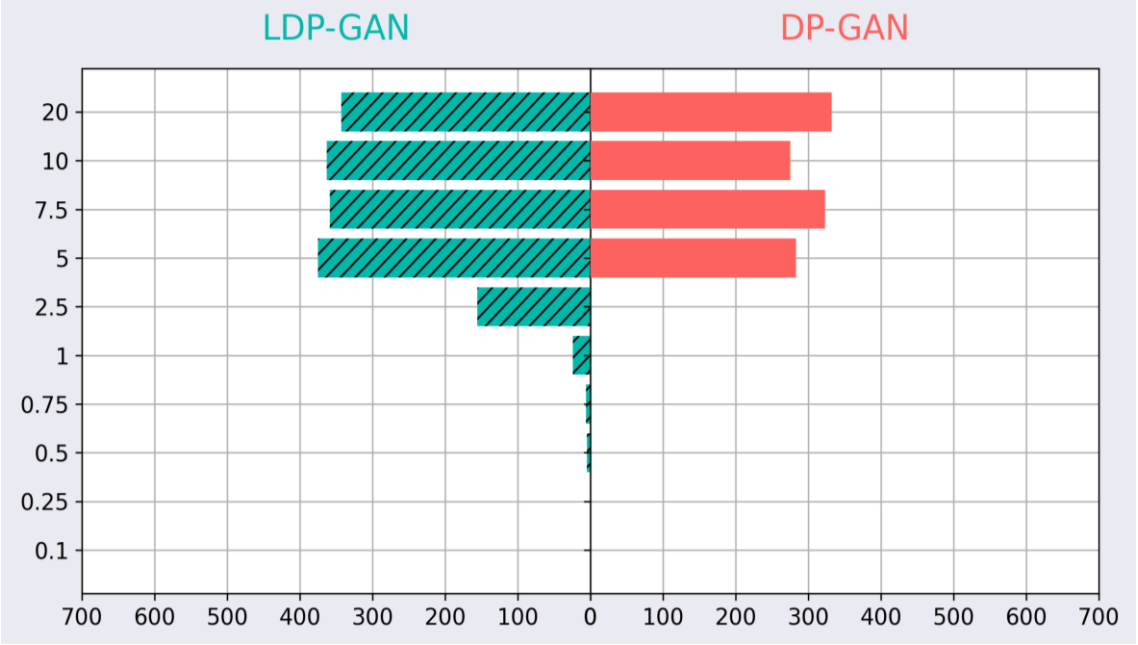
2.4 White-box Discriminator Attack

White-box Discriminator Attack(WDA)은 Attacker가 판별자(Discriminator)에 접근가능 할 때 이를 통해 가하는 Attack 시나리오이다. 학습에 사용된 샘플과 Holdout 샘플의 Discriminator 출력 값을 기준으로 내림차순으로 정렬 후 상위 k개의 데이터를 훈련데이터라 주장한다. 상위 k개의 데이터에 훈련데이터의 비중이 높을수록 공격이 더 유효 해진다. 이 연구에서는 훈련 데이터셋 Holdout 데이터 각각 625개를 사용했고 이중 상위 200개를 기준으로 Attack을 수행했다. 최악의 경우는 상위 200개의 데이터가 전부 학습 데이터인 경우인데, 이런 경우는 Attacker가 판별자의 출력 값을 기준으로 거의 완벽하게 훈련데이터를 구별해 낼 수 있다. 가장 이상적인 경우는 상위 200개의 데이터에 훈련 데이터와 Holdout데이터가 100개씩 속해 있는 경우이다. 이런 경우 Attacker는 판별자로부터 아무런 정보를 얻을 수 없어 공격이 불가능해진다. 판별자가 학습 데이터셋에 과적합할수록 이 상위 k개의 데이터 중 학습 데이터셋의 비중이 늘어나는데, 이 비중을 통해 판별자의 과적합정도를 알 수 있다. 학습 데이터셋에 과적합 됐다는 자체로 여러 공격에 취약해지기 때문에 과적합 정도는 모델의 보안성을 대략적으로 평가하는 지표가 될 수 있다.

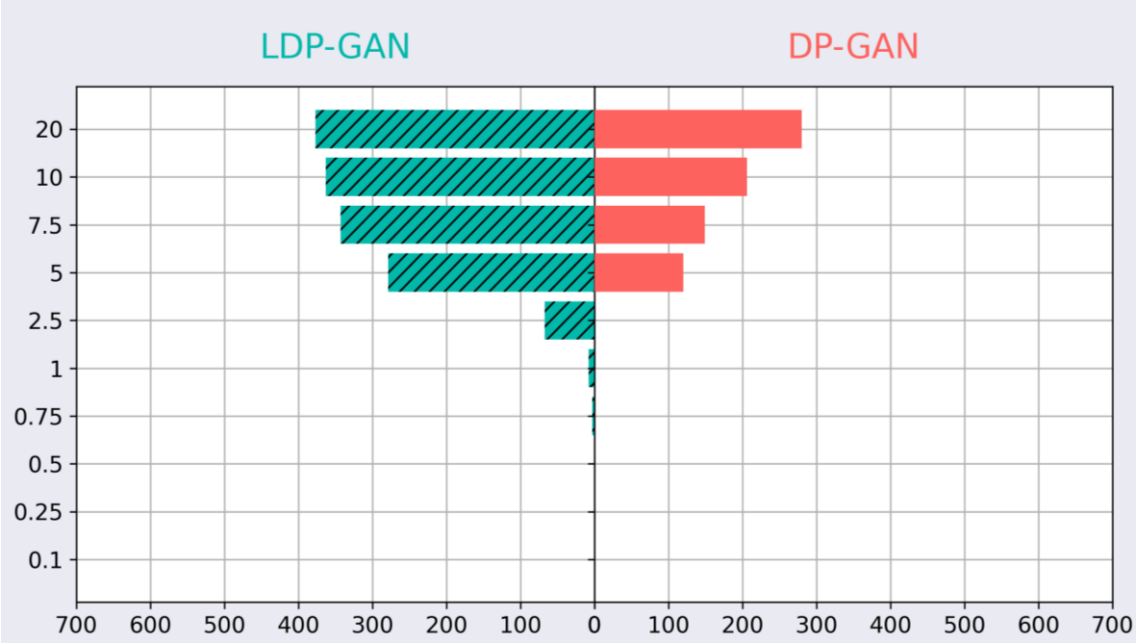
[그림13-White-box Discriminator Attack의 결과]은 White-box Discriminator Attack에서 상위 k개의 데이터 중 학습 데이터셋의 개수를 나타낸다. 그래프가 점선에 가까울수록 이상적인 프라이버시를 의미하고 점선을 많이 넘어갈수록 Attack에 취약함을 의미한다. 전체적으로 LDP-GAN에서 높은 수치를 보였으며 최고치 역시 LDP-GAN에서 보였다. 다만 DP-GAN은 ε과 어떠한 비례관계도 보이지 못했고, LDP-GAN은 적절한 비례관계를 보였다. ε과 비례관계를 보이지 않으면 프라이버시 수준을 원하는 대로 조절하기 힘들어진다. 이러한 모델은경우에 따라 높은 프라이버시 수준을 만들더라도 좋은 프라이버시 모델이라고 보기 어렵다.



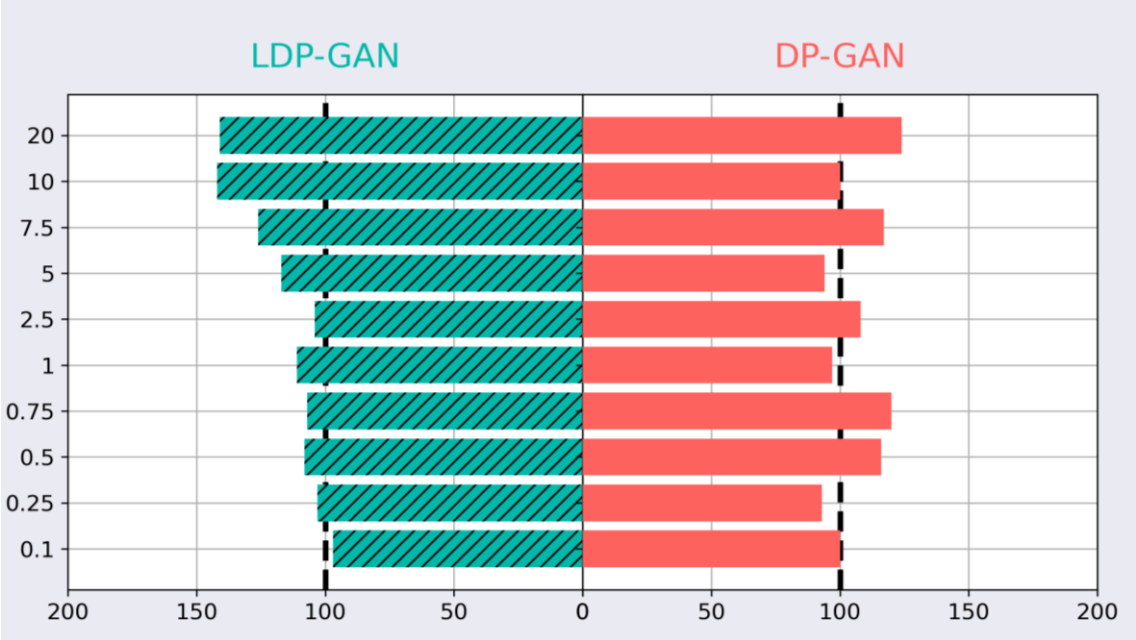
[그림10 – Full Black-box Attack의 결과. 세로축은 ε을 나타낸다.]



[그림11– Partial Black-box Attack의 결과. 세로축은 ε을 나타낸다.]

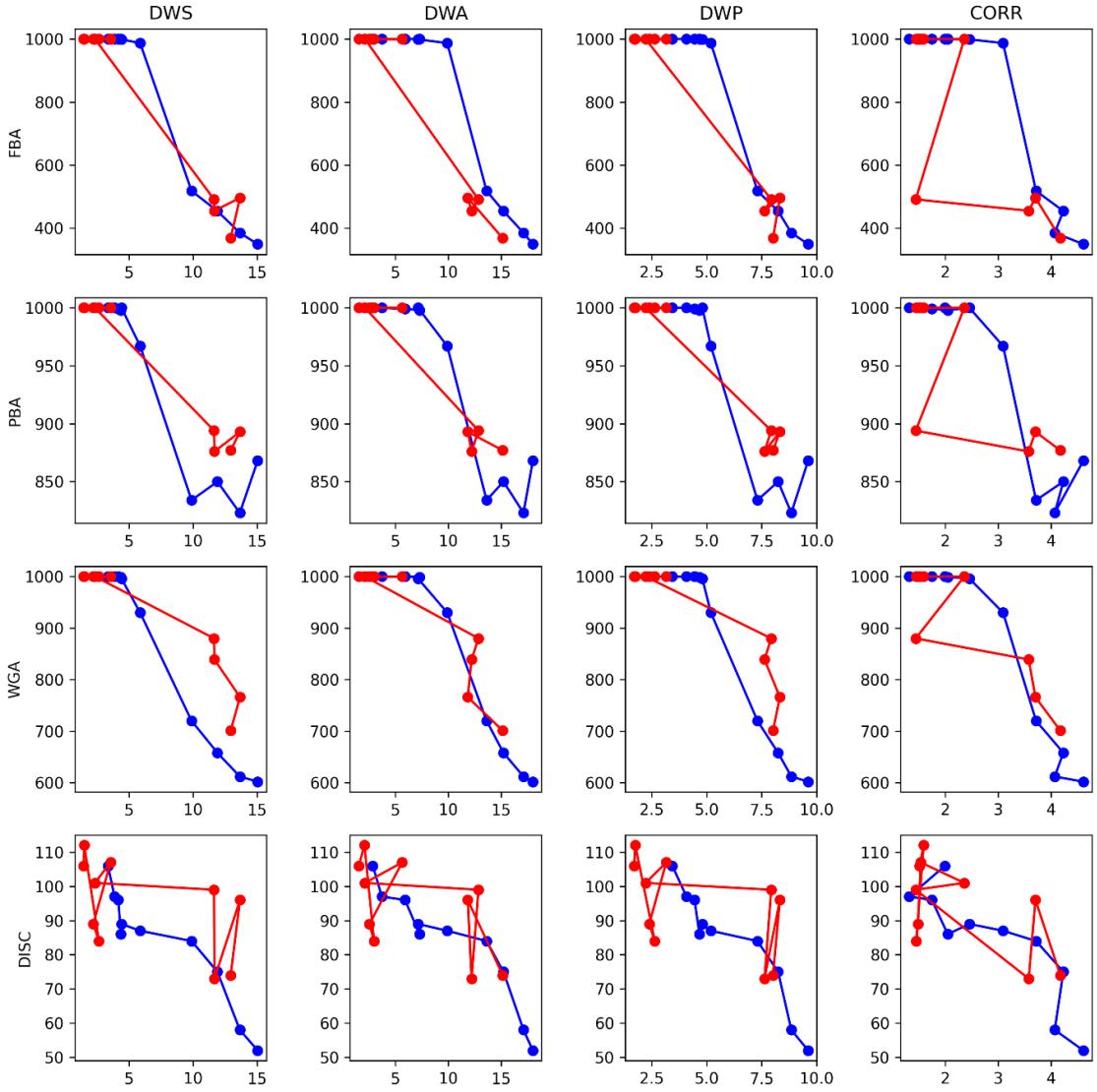


[그림12– White-box Gradient Attack의 결과. 세로축은 ε을 나타낸다]



[그림13 White-box Discriminator Attack의 결과. 세로축은 ε을 나타낸다.]

3. Trade-off 테스트 결과



[그림14 – Trade-off 테스트의 결과. 세로축은 프라이버시, 가로축은 효용성을 나타낸다. 푸른색 그래프는 LDP-GAN의 결과, 빨간색 그래프는 DP-GAN의 결과이다.]

Trade-off 관계는 효용성 지표를 가로축에, 프라이버시 지표를 세로축에 표시하여 좌표평면에 그래프로 시각화 할 수 있다. 가장 이상적인 경우는 효용성이 증가하여도 프라이버시 지표가 감소하지 않는 경우이나 현실에서는 존재하기 어렵고 일반적으로 효용성이 증가할 때 프라이버시는 감소한다. 효율적인 Trade-off 관계에서는 효용성이 증가할 때 프라이버시 지표가 적게 감소하며 이 경우 그래프는 감소폭이 완만하고 상대적으로 위쪽, 그리고 오른쪽에 그려지게 된다. 또한 Trade-off 안정성도 좋은 Trade-off를 평가하는 기준이 될 수 있

다. 안정성 있는 Trade-off는 그래프가 매끄럽게 하강하며 그래프가 진동하는 경우는 안정성이 떨어진다고 볼 수 있다.

이 연구에서 사용한 4가지 효용성 지표(DWS, DWA, DWP, Correlation)와 4가지 프라이버시 지표 (FBA, PBA, WGA, WDA)를 1대1로 조합하여 총 16가지 테스트가 수행됐다[그림14 - Trade-off 테스트의 결과]. 이때 푸른색 그래프는 LDP-GAN의 결과를 나타내며 붉은색 그래프는 DP-GAN의 결과를 나타낸다. ε은 [0.1, 0.25, 0.5, 0.75, 1, 2.5, 5, 7.5, 10, 20]의 값을 가지며 각 점은 하나의 ε에 따른 결과를 나타낸다. 점을 잇는 선들은 점들을 ε 순서대로 순차적으로 이어서 나타냈다. FBA를 프라이버시 지표로 사용했을 때만 LDP-GAN의 그래프가 더 위쪽에 그려지는 경향을 보였고 나머지의 경우에선 ε이 작을 때는 LDP-GAN이, ε이 클 때는 DP-GAN이 약간 우위를 보였다. 가장 큰 차이는 안정성에서 보였는데 LDP-GAN의 그래프는 ε이 증가함에 따라 매끄럽게 감소하는 반면 DP-GAN은 큰 진동을 보였다. 특히 DP-GAN의 그래프는 전체적으로 진동을 하다가 한번의 큰 이동 후 다시 진동을 하는 모습을 보였다. 이 상황에선 ε의 변화에 따른 Trade-off 관계에서 적절한 파라미터를 선택하기 어렵다. 결과적으로 이 실험에서 사용한 ε 범위에선 LDP-GAN이 DP-GAN대비 ε값에 대해 더 안정적인 Trade-off관계를 보였고 경우에 따라 더 효율적인 Trade-off를 보임을 알 수 있었다.

결론 및 논의

우리는 이 연구에서 프라이버시를 유출시키지 않는 GAN 생성모델에 대해서 연구했다. 이 목적을 위해 Ambient-GAN 구조를 활용하고 지역 차분 프라이버시(Local Differential Privacy)를 접목시켰다. 이 모델은 교란된 데이터로만 학습한다는 특징이 있다. 이 모델은 전역 차분 프라이버시(Global Differential Privacy)가 적용된 DP-GAN과 비교됐으며, 효용성과 프라이버시측면에 나누어 두 모델을 평가했다. 비교 모델인 DP-GAN에 대비해서 우리가 제시한 모델인 LDP-GAN은 이 연구에서 설정한 대부분의 경우에서 잡음의 강도에 따른 좋은 선형적 변화를 보였고 이론과 일치하는 모습을 보였다. LDP-GAN은 특히 잡음이 강한 상황에서 DP-GAN대비 높은 효용성을 보였는데 이러한 차이들은 학습 과정에서 잡음을 가하는 방식의 차이에 따른 것으로 보인다. 교란된 데이터로 학습하며 실제 데이터의 분포를 추론하는 LDP-GAN의 학습방식이 강하게 교란된 상황에서도 좋은 효용성을 만들어내고 학습을 안정시킨 결과로 보여진다. 이러한 학습방식은 아직 연구할 여지가 많이 남아있으며 이는 향후 연구의 중요한 방향이 될 것이다. 또한, 이 연구에서 지역 차분 프라이버시를 채택함으로써 얻은 큰 장점 중 하나는 항목별로 특성을 고려한 프라이버시를 적용할 수 있다는 점이다. 전역 차분 프라이버시를 적용하는 DP-GAN과 달리 데이터의 특성을 고려한 잡음을 디자인할 수 있고 더욱 효율적인 학습이 가능하다. 또한 데이터 특성에 맞게 다양하고 정교하게 디자인된 잡음은 높은 효용성을 유지하면서 원하는 수준의 프라이버시를 얻을 수 있게 한다. 본 연구에서는 데이터의 범위나 데이터 타입(Type)등만 고려했지만 더 복잡한 특성을 고려하여 잡음을 디자인하는 것 또한 중요한 연구주제 중 하나가 될 것이다.

실험결과들은 이 연구에서 제안한 모델이 기존의 모델보다 안정적임을 보여주기는 했으나 Trade-off에서 크게 효율적임 보여주진 않았다. 이러한 효율성을 개선하는 것은 프라이버시 모델의 가장 중요한 과제 중 하나로, 이 연구의 최종 목표 역시 효율적인 Trade-off를 가지는 생성모델을 개발하는 것이다. 이 목표는 앞서 언급한 학습방식의 개선과 잡음의 디자인을 통해 달성할 수 있을 것으로 기대한다. 효율성을 개선함으로써 높은 프라이버시 수준을 유지하면서 임상적으로 실제와 거의 유사한 데이터를 생성하는 생성모델을 개발하여 의료 데이터가 가지는 여러가지 제약을 해결할 수 있을 것으로 예상한다.

참고 문헌

[1] W Nicholson Price and I Glenn Cohen. Privacy in the age of medical big data. Nature medicine, 25(1):37-43, 2019.

[2] Latanya Sweeney. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05):557-570, 2002.

[3] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond kanonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1):3-es, 2007.

[4] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In 2007 IEEE 23rd international conference on data engineering, pages 106-115. IEEE, 2007.

[5] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, pages 1322-1333, 2015.

[6] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In 2017 IEEE symposium on security and privacy (SP), pages 3-18. IEEE, 2017.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems, 27, 2014.

[8] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. Advances in neural information processing systems, 30, 2017.

[9] Cynthia Dwork. Differential privacy: A survey of results. In International conference on theory and applications of models of computation, pages 1-19. Springer, 2008.

[10] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, pages 308-318, 2016.

[11] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. arXiv preprint arXiv:1802.06739, 2018.

[12] Ashish Bora, Eric Price, and Alexandros G. Dimakis. AmbientGAN: Generative models from lossy measurements. In International Conference on Learning Representations, 2018.

[13] Ziqi Zhang, Chao Yan, Diego A Mesa, Jimeng Sun, and Bradley A Malin. Ensuring electronic medical record simulation through better training, modeling, and evaluation. Journal of the American Medical Informatics Association, 27(1):99-108, 2020.

[14] Mrinal Kanti Baowaly, Chia-Ching Lin, Chao-Lin Liu, and Kuan-Ta Chen. Synthesizing electronic health records using improved generative adversarial networks. Journal of the American Medical Informatics Association, 26(3):228-241, 2019.

[15] Sumit Mukherjee, Yixi Xu, Anusua Trivedi, Nabajyoti Patowary, and Juan Lavista Ferres. privgan: Protecting gans from membership inference attacks at low cost to utility. Proc. Priv. Enhancing Technol., 2021(3):142-163, 2021.

[16] Yi Liu, Jialiang Peng, JQ James, and Yi Wu. Ppgan: Privacy-preserving generative adversarial network. In 2019 IEEE 25Th international conference on parallel and distributed systems (ICPADS), pages 985- 989. IEEE, 2019.

[17] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In ICML deep learning workshop, volume 2, page 0. Lille, 2015

[18] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), volume 2, pages 1735-1742. IEEE, 2006

[19] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4):600-612, 2004

[20] Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In 2013 IEEE Global Conference on Signal and Information Processing, pages 245- 248. IEEE, 2013.

[21] Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. Dp-gan: diversity-promoting generative adversarial network for generating informative and diversified text. arXiv preprint arXiv:1802.01345, 2018.

[22] Brett K Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, Ran Lee, Sanjeev P Bhavnani, James Brian Byrd, and Casey S Greene. Privacy preserving generative deep neural networks support clinical data sharing. Circulation: Cardiovascular Quality and Outcomes, 12(7):e005122, 2019.

[23] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In International conference on machine learning, pages 2642-2651. PMLR, 2017

[24] Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 60(309):63-69, 1965.

[25] Imjin Ahn, Wonjun Na, Osung Kwon, Dong Hyun Yang, Gyung-Min Park, Hansle Gwon, Hee Jun Kang, Yeon Uk Jeong, Jungsun Yoo, Yunha Kim, et al. Cardionet: a manually curated database for artificial intelligence-based research on cardiovascular diseases. BMC Medical Informatics and Decision Making, 21(1):1-15, 2021.

[26] J. Ross Quinlan. Induction of decision trees. Machine learning, 1(1):81- 106, 1986. [22] John A Nelder and Roger Mead. A simplex method for function minimization.

The computer journal, 7(4):308-313, 1965.

[28] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. Mathematical programming, 45(1):503- 528, 1989.

[29] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, pages 837- 845, 1988.

[30] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. Gan-leaks: A taxonomy of membership inference attacks against generative models. In Proceedings of the 2020 ACM SIGSAC conference on computer and communications security, pages 343-362, 2020

영문 요약

The electronic medical records (EMR) are a type of medical data containing the patient's health condition, treatment results, and prescription information. It contains a lot of information on patients, so it can be used in various ways, and has the potential to improve the quality of medical care. In particular, machine learning which has recently made great progress, has been introduced into the medical field, eventually leading to the increased usage of EMR. However, EMR contain a number of sensitive personal information of patients, making it difficult to collect, utilize, and share. These characteristics make it difficult to study and utilize the EMR. Thus, the generative model can be a great solution to the previous difficulties.

A generative model refers to a model that generates synthetic data similar to actual data. By utilizing synthetic data generated in this generative model, restrictions on personal information can be avoided. Although there are many types of generative models, recently, generative models using deep learning are the most noteworthy. In fact, deep learning generative models have made great strides in the field of images, and are able to generate high-resolution images that are difficult to determine authenticity with the human eye. Moreover, the deep learning generative model was also applied to medical data and was able to generate clinically meaningful data. Though deep learning generative models show good performance, they do not completely solve personal information problem. In the past, several studies have dealt with attacks on deep learning models, and it has been found that training data can be inferred based on the output values of the models. The result indicates that even when using a deep learning generative model, there still remains a risk to privacy, and protection of the model is therefore necessary if the model is used for the purpose of protecting personal information.

Further, the objective of this study is to develop a deep learning generative model that is safe from membership inference attacks. To achieve this objective, we used WGAN-GP, a type of Generative adversarial network(GAN), as a basic model, and adopted differential privacy to protect privacy. The differential privacy protects privacy through mathematically designed noise, and uses ε, a parameter related to noise intensity, to adjust the trade-off relationship between utility and privacy protection levels. In this study, we developed a method for learning a model using only perturbation data by introducing regional differential privacy among differential privacy. Also, because training is performed only on perturbed data, the original data can be strongly protected from attacks on the model.

Next, the performance of the model trained in this way was evaluated in terms of utility and privacy. Both evaluation indicators showed significant changes according to ε, and it was shown that it is possible to obtain an optimal model by appropriately adjusting the trade-off relationship between the two indicators. The results of this

experiment signifies that the training data can be protected from attacks on the model if appropriate noise is applied. Through the finding of this experiment, it is expected that the limitations caused by the personal information issues on EMR can be resolved to some extent.

Key words: EMR, deep-learning, Differential Privacy, GAN, Trade-off