



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0037437
(43) 공개일자 2024년03월22일

(51) 국제특허분류(Int. Cl.)
G16H 10/60 (2018.01) G06N 20/00 (2019.01)
G06N 3/08 (2023.01) G16H 50/70 (2018.01)
(52) CPC특허분류
G16H 10/60 (2021.08)
G06N 20/00 (2021.08)
(21) 출원번호 10-2022-0115852
(22) 출원일자 2022년09월14일
심사청구일자 2022년09월14일

(71) 출원인
재단법인 아산사회복지재단
서울특별시 송파구 올림픽로43길 88 (풍납동)
울산대학교 산학협력단
울산광역시 남구 대학로 93(무거동)
(72) 발명자
김영학
서울특별시 강남구 선릉로 120, 3동 1102호 (대치동, 개포1,2차우성아파트)
전태준
서울특별시 송파구 풍성로6길 15, 101동 708호(풍납동, 신동아아파트)
(74) 대리인
특허법인 무한

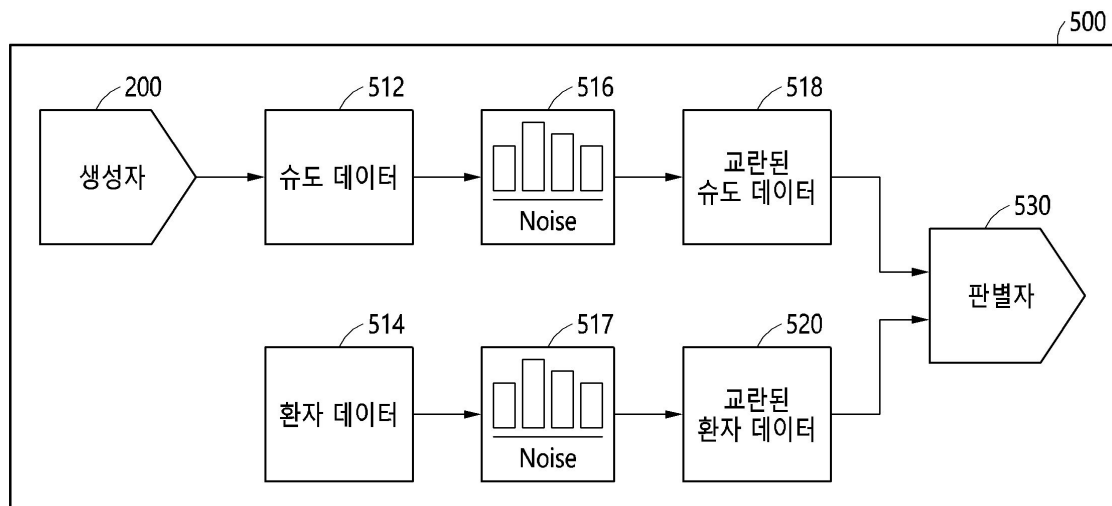
전체 청구항 수 : 총 17 항

(54) 발명의 명칭 지역 차분 프라이버시 기반 적대적 생성 네트워크를 활용한 합성 환자 데이터 생성 방법 및 장치

(57) 요약

일 실시예에 따른 보안 의료 데이터 생성 전자 장치는, 컴퓨터로 실행 가능한 명령어들(computer-executable instructions)이 저장된 메모리; 및 상기 메모리에 액세스(access)하여 상기 명령어들을 실행하는 프로세서를 포함하고, 상기 명령어들은, 전자의무기록(EMR, Electronic Medical Records)으로부터 추출된 환자 데이터로부터 교란된 환자 데이터를 생성하고, 잠재 공간(latent space)의 잠재 벡터(latent vector)로부터 생성자에 기초하여 생성된 슈도 데이터(pseudo data)로부터 교란된 슈도 데이터를 생성하고, 상기 교란된 환자 데이터 및 상기 교란된 슈도 데이터에 기초하여, 생성자(generator) 및 판별자(discriminator)를 포함하는 적대적 생성 네트워크(GAN; Generative Adversarial Network) 기반 모델을 학습시키고, 상기 생성자를 이용하여 보안 의료 데이터(secured medical data)를 생성할 수 있다.

대표도



(52) CPC특허분류

G06N 3/08 (2023.01)

G16H 50/70 (2018.01)

(72) 발명자

권한솔

서울특별시 노원구 초안산로5길 22, 105동 202호(월계동, 대우아파트)

안임진

경기도 의정부시 진등로19번길 20, 901동 1207호(녹양동, 녹양동 신도9차 전원마을아파트)

강희준

경기도 성남시 중원구 둔촌대로 300, 1004동 1004호(하대원동, 아튼빌아파트)

김윤하

서울특별시 성동구 서울숲길 25, 102동 304호(성수동1가, 현대아파트)

서혜람

서울특별시 강동구 양재대로124길 67-11, B102호(길동, 금석빌라)

조하나

서울특별시 송파구 오금로11길 41, 704호(방이동)

김민경

경기도 의정부시 오목로 36, 103동 1502호(용현동, 현대아파트)

한지예

서울특별시 양천구 목동동로 10, 1110동 601호(신정동, 목동신시가지아파트11단지)

기가은

서울특별시 서초구 서초대로16길 22, 302호(방배동, 삼풍맨션)

박서현

서울특별시 서초구 서초대로33길 84-24, B동 402호(방배동, 방배4차상지리츠빌)

최희정

서울특별시 강남구 봉은사로49길 21-10, 202호(논현동, 블루하우스(Blue house))

이 발명을 지원한 국가연구개발사업

과제고유번호	1465035447
과제번호	HR21C0198010022
부처명	보건복지부
과제관리(전문)기관명	한국보건산업진흥원
연구사업명	연구중심병원 육성 R&D
연구과제명	의료현장 통합 지식베이스 및 인공지능형 가상 임상시험 시스템 구축
기 여 율	1/2
과제수행기관명	서울아산병원
연구기간	2022.01.01 ~ 2022.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711152941
과제번호	2021-0-00982-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	ICT융합산업혁신기술개발사업
연구과제명	최적 치료 협진 서비스를 위한 다기관 분산지능 강화 페더레이션 기술 개발
기 여 율	1/2
과제수행기관명	한국전자통신연구원
연구기간	2022.01.01 ~ 2022.12.31

명세서

청구범위

청구항 1

프로세서에 의하여 수행되는 보안 의료 데이터 생성 방법에 있어서,

전자의무기록(EMR, Electronic Medical Records)으로부터 추출된 환자 데이터로부터 교란된 환자 데이터를 생성하는 단계;

잠재 공간(latent space)의 잠재 벡터(latent vector)로부터 생성자에 기초하여 생성된 슈도 데이터(pseudo data)로부터 교란된 슈도 데이터를 생성하는 단계;

상기 교란된 환자 데이터 및 상기 교란된 슈도 데이터에 기초하여, 생성자(generator) 및 판별자(discriminator)를 포함하는 적대적 생성 네트워크(GAN; Generative Adversarial Network) 기반 모델을 학습(training)시키는 단계; 및

상기 생성자를 이용하여 보안 의료 데이터(secured medical data)를 생성하는 단계를 포함하는 방법.

청구항 2

제1항에 있어서,

상기 추출된 환자 데이터가 이산(discrete) 데이터인 경우 무작위 응답(randomized response) 메커니즘에 기초하여, 변경 확률로 상기 환자 데이터와 반대되는 이산 데이터를 포함하는 상기 교란된 환자 데이터를 생성하는 단계; 및

상기 이산 데이터에 기초하여, 상기 변경 확률로 상기 슈도 데이터와 반대되는 이산 데이터를 포함하는 상기 교란된 슈도 데이터를 생성하는 단계

를 더 포함하는 방법.

청구항 3

제1항에 있어서,

상기 추출된 환자 데이터가 연속(continuous) 데이터인 경우 라플라스 메커니즘(Laplace mechanism)에 기초하여, 노이즈 데이터를 생성하는 단계;

상기 노이즈 데이터를 상기 환자 데이터에 부가하여 상기 교란된 환자 데이터를 생성하는 단계; 및

상기 노이즈 데이터를 상기 슈도 데이터에 부가하여 상기 교란된 슈도 데이터를 생성하는 단계

를 더 포함하는 방법.

청구항 4

제1항에 있어서,

사용자에 의해 선택된 보안성 조절 레벨에 기초하여, 상기 보안성 조절 레벨에 대응하는 분포에서 노이즈를 생성하는 단계;

상기 추출된 환자 데이터에 상기 노이즈를 부가함으로써 상기 교란된 환자 데이터를 생성하는 단계;

상기 슈도 데이터에 상기 노이즈를 부가함으로써 상기 교란된 슈도 데이터를 생성하는 단계;
를 더 포함하는 방법.

청구항 5

제4항에 있어서,
상기 노이즈를 생성하는 단계는,
무작위 응답 메커니즘, 라플라스 메커니즘, 가우시안 메커니즘(Gaussian mechanism), 또는 익스포넨셜 메커니즘(exponential mechanism) 중 적어도 하나의 메커니즘에 기초하여 상기 노이즈를 생성하는 단계
를 포함하는 방법.

청구항 6

제5항에 있어서,
상기 복수의 메커니즘 중 적어도 두 개에 기초하여, 제1 노이즈 및 제2 노이즈를 생성하는 단계;
상기 추출된 환자 데이터에 상기 제1 노이즈를 부가함으로써 상기 교란된 환자 데이터를 생성하는 단계; 및
상기 슈도 데이터에 상기 제2 노이즈를 부가함으로써 상기 교란된 슈도 데이터를 생성하는 단계
를 더 포함하는 방법.

청구항 7

제1항에 있어서,
상기 적대적 생성 네트워크 기반 모델을 학습시키는 단계는,
상기 교란된 슈도 데이터 및 상기 교란된 환자 데이터에 상기 판별자를 개별적으로 적용한 결과들에 기초하여 판별자 점수(discriminator score)를 획득하는 단계;
상기 판별자 점수에 기초하여, 상기 생성자가 상기 환자 데이터와 유사한 상기 슈도 데이터를 생성하도록 상기 적대적 생성 네트워크 기반 모델을 학습시키는 단계;
상기 판별자 점수에 기초하여, 상기 판별자가 상기 교란된 환자 데이터에 대해서는 참(true)을 나타내는 결과를 생성하도록 상기 적대적 생성 네트워크 기반 모델을 학습시키는 단계; 및
상기 판별자 점수에 기초하여, 상기 판별자가 상기 교란된 슈도 데이터에 대해서는 거짓(false)을 나타내는 결과를 생성하도록 상기 적대적 생성 네트워크 기반 모델을 학습시키는 단계
를 포함하는 방법.

청구항 8

제7항에 있어서,
상기 판별자 점수를 획득하는 단계는,
와서스타인 거리(Wasserstein distance)에 기초하는 상기 판별자의 목적함수를 통해 상기 판별자 점수를 획득하는 단계
를 포함하는 방법.

청구항 9

하드웨어와 결합되어 제1항 내지 제8항 중 어느 하나의 항의 방법을 실행시키기 위하여 컴퓨터 판독 가능한 기록매체에 저장된 컴퓨터 프로그램.

청구항 10

보안 의료 데이터 생성 전자 장치에 있어서,

컴퓨터로 실행 가능한 명령어들(computer-executable instructions)이 저장된 메모리; 및

상기 메모리에 액세스(access)하여 상기 명령어들을 실행하는 프로세서

를 포함하고,

상기 명령어들은,

전자의무기록(EMR, Electronic Medical Records)으로부터 추출된 환자 데이터로부터 교란된 환자 데이터를 생성하고,

잠재 공간(latent space)의 잠재 벡터(latent vector)로부터 생성자에 기초하여 생성된 슈도 데이터(pseudo data)로부터 교란된 슈도 데이터를 생성하고,

상기 교란된 환자 데이터 및 상기 교란된 슈도 데이터에 기초하여, 생성자(generator) 및 판별자(discriminator)를 포함하는 적대적 생성 네트워크(GAN; Generative Adversarial Network) 기반 모델을 학습시키고,

상기 생성자를 이용하여 보안 의료 데이터(secured medical data)를 생성하는

보안 의료 데이터 생성 전자 장치.

청구항 11

제10항에 있어서,

상기 프로세서는,

상기 추출된 환자 데이터가 이산(discrete) 데이터인 경우 무작위 응답(randomized response) 메커니즘에 기초하여, 변경 확률로 상기 환자 데이터와 반대되는 이산 데이터를 포함하는 상기 교란된 환자 데이터를 생성하고,

상기 이산 데이터에 기초하여, 상기 변경 확률로 상기 슈도 데이터와 반대되는 이산 데이터를 포함하는 상기 교란된 슈도 데이터를 생성하는

보안 의료 데이터 생성 전자 장치.

청구항 12

제10항에 있어서,

상기 프로세서는,

상기 추출된 환자 데이터가 연속(continuous) 데이터인 경우 라플라스 메커니즘(Laplace mechanism)에 기초하여, 노이즈 데이터를 생성하고,

상기 노이즈 데이터를 상기 환자 데이터에 부가하여 상기 교란된 환자 데이터를 생성하고,

상기 노이즈 데이터를 상기 슈도 데이터에 부가하여 상기 교란된 슈도 데이터를 생성하는

보안 의료 데이터 생성 전자 장치.

청구항 13

제10항에 있어서,

상기 프로세서는,

사용자에 의해 선택된 보안성 조절 레벨에 기초하여, 상기 보안성 조절 레벨에 대응하는 분포에서 노이즈를 생성하고,

상기 추출된 환자 데이터에 상기 노이즈를 부가함으로써 상기 교란된 환자 데이터를 생성하고,

상기 슈도 데이터에 상기 노이즈를 부가함으로써 상기 교란된 슈도 데이터를 생성하는

보안 의료 데이터 생성 전자 장치.

청구항 14

제13항에 있어서,

상기 프로세서는,

무작위 응답 메커니즘, 라플라스 메커니즘, 가우시안 메커니즘(Gaussian mechanism), 또는 익스포넬셜 메커니즘(exponential mechanism) 중 적어도 하나의 메커니즘에 기초하여 상기 노이즈를 생성하는

보안 의료 데이터 생성 전자 장치.

청구항 15

제14항에 있어서,

상기 프로세서는,

상기 복수의 메커니즘 중 적어도 두 개에 기초하여, 제1 노이즈 및 제2 노이즈를 생성하고,

상기 추출된 환자 데이터에 상기 제1 노이즈를 부가함으로써 상기 교란된 환자 데이터를 생성하고,

상기 슈도 데이터에 상기 제2 노이즈를 부가함으로써 상기 교란된 슈도 데이터를 생성하는

보안 의료 데이터 생성 전자 장치.

청구항 16

제10항에 있어서,

상기 프로세서는,

상기 교란된 슈도 데이터 및 상기 교란된 환자 데이터에 상기 판별자를 개별적으로 적용한 결과들에 기초하여 판별자 점수(discriminator score)를 획득하고,

상기 판별자 점수에 기초하여, 상기 생성자가 상기 환자 데이터와 유사한 상기 슈도 데이터를 생성하도록 상기 적대적 생성 네트워크 기반 모델을 학습시키고,

상기 판별자 점수에 기초하여, 상기 판별자가 상기 교란된 환자 데이터에 대해서는 참(true)을 나타내는 결과를 생성하도록 상기 적대적 생성 네트워크 기반 모델을 학습시키고,

상기 판별자 점수에 기초하여, 상기 판별자가 상기 교란된 슈도 데이터에 대해서는 거짓(false)을 나타내는 결과를 생성하도록 상기 적대적 생성 네트워크 기반 모델을 학습시키는

보안 의료 데이터 생성 전자 장치.

청구항 17

제16항에 있어서,

상기 프로세서는,

와서스타인 거리(Wasserstein distance)에 기초하는 상기 판별자의 목적함수를 통해 상기 판별자 점수를 획득하는

보안 의료 데이터 생성 전자 장치.

발명의 설명

기술 분야

[0001] 이하, 지역 차분 프라이버시 기반 적대적 생성 네트워크를 활용한 합성 환자 데이터 생성 방법 및 장치에 관한 기술이 제공된다.

배경 기술

[0002] 의료 분야는 인공 지능이 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 인공 지능이 특정 문제를 해결함에 있어서 전문의 수준의 판단이 가능함을 보였고 점점 영역을 확장하고 있다. 하지만 인공 지능이 효과적으로 의료영역에서 적용되기 위해서 해결해야 할 문제가 몇 가지 남아있다. 현재 가장 중요한 문제는 환자 데이터의 개인정보 보안에 관한 문제로 이는 의료 데이터를 다루는데 있어서 가장 복잡한 문제이다.

[0003] 위에서 설명한 배경기술은 발명자가 본원의 개시 내용을 도출하는 과정에서 보유하거나 습득한 것으로서, 반드시 본 출원 전에 일반 공중에 공개된 공지기술이라고 할 수는 없다.

발명의 내용

해결하려는 과제

과제의 해결 수단

[0004] 일 실시예에 따른 보안 의료 데이터 생성 전자 장치는, 컴퓨터로 실행 가능한 명령어들(computer-executable instructions)이 저장된 메모리; 및 상기 메모리에 액세스(access)하여 상기 명령어들을 실행하는 프로세서를 포함하고, 상기 명령어들은, 전자의무기록(EMR, Electronic Medical Records)으로부터 추출된 환자 데이터로부터 교란된 환자 데이터를 생성하고, 잠재 공간(latent space)의 잠재 벡터(latent vector)로부터 생성자에 기초하여 생성된 슈도 데이터(pseudo data)로부터 교란된 슈도 데이터를 생성하고, 상기 교란된 환자 데이터 및 상기 교란된 슈도 데이터에 기초하여, 생성자(generator) 및 판별자(discriminator)를 포함하는 적대적 생성 네트워크(GAN; Generative Adversarial Network) 기반 모델을 학습시키고, 상기 생성자를 이용하여 보안 의료 데이터(secured medical data)를 생성할 수 있다.

[0005] 상기 프로세서는, 상기 추출된 환자 데이터가 이산(discrete) 데이터인 경우 무작위 응답(randomized response) 메커니즘에 기초하여, 변경 확률로 상기 환자 데이터와 반대되는 이산 데이터를 포함하는 상기 교란된 환자 데이터를 생성하고, 상기 이산 데이터에 기초하여, 상기 변경 확률로 상기 슈도 데이터와 반대되는 이산 데이터를 포함하는 상기 교란된 슈도 데이터를 생성할 수 있다.

[0006] 상기 프로세서는, 상기 추출된 환자 데이터가 연속(continuous) 데이터인 경우 라플라스 메커니즘(Laplace mechanism)에 기초하여, 노이즈 데이터를 생성하고,

[0007] 상기 노이즈 데이터를 상기 환자 데이터에 부가하여 상기 교란된 환자 데이터를 생성하고, 상기 노이즈 데이터

를 상기 슈도 데이터에 부가하여 상기 교란된 슈도 데이터를 생성할 수 있다.

- [0008] 상기 프로세서는, 사용자에게 의해 선택된 보안성 조절 레벨에 기초하여, 상기 보안성 조절 레벨에 대응하는 분포에서 노이즈를 생성하고, 상기 추출된 환자 데이터에 상기 노이즈를 부가함으로써 상기 교란된 환자 데이터를 생성하고, 상기 슈도 데이터에 상기 노이즈를 부가함으로써 상기 교란된 슈도 데이터를 생성할 수 있다.
- [0009] 상기 프로세서는, 무작위 응답 메커니즘, 라플라스 메커니즘, 가우시안 메커니즘(Gaussian mechanism), 또는 익스포넬셜 메커니즘(exponential mechanism) 중 적어도 하나의 메커니즘에 기초하여 상기 노이즈를 생성할 수 있다.
- [0010] 상기 프로세서는, 상기 복수의 메커니즘 중 적어도 두 개에 기초하여, 제1 노이즈 및 제2 노이즈를 생성하고, 상기 추출된 환자 데이터에 상기 제1 노이즈를 부가함으로써 상기 교란된 환자 데이터를 생성하고, 상기 슈도 데이터에 상기 제2 노이즈를 부가함으로써 상기 교란된 슈도 데이터를 생성할 수 있다.
- [0011] 상기 프로세서는, 상기 교란된 슈도 데이터 및 상기 교란된 환자 데이터에 상기 판별자를 개별적으로 적용한 결과들에 기초하여 판별자 점수(discriminator score)를 획득하고, 상기 판별자 점수에 기초하여, 상기 생성자가 상기 환자 데이터와 유사한 상기 슈도 데이터를 생성하도록 상기 적대적 생성 네트워크 기반 모델을 학습시키고, 상기 판별자 점수에 기초하여, 상기 판별자가 상기 교란된 환자 데이터에 대해서는 참(true)을 나타내는 결과를 생성하도록 상기 적대적 생성 네트워크 기반 모델을 학습시키고, 상기 판별자 점수에 기초하여, 상기 판별자가 상기 교란된 슈도 데이터에 대해서는 거짓(false)을 나타내는 결과를 생성하도록 상기 적대적 생성 네트워크 기반 모델을 학습시킬 수 있다.
- [0012] 상기 프로세서는, 와서스타인 거리(Wasserstein distance)에 기초하는 상기 판별자의 목적함수를 통해 상기 판별자 점수를 획득할 수 있다.

도면의 간단한 설명

- [0013] 도 1은 멤버십 추론 공격(membership inference attack)을 도시한 도면이다.
- 도 2는 생성자에 대한 풀 블랙 박스 공격(full black-box attack)을 도시한 도면이다.
- 도 3은 생성자에 대한 부분 블랙 박스 공격(partial black-box attack)을 도시한 도면이다.
- 도 4는 판별자에 대한 판별자 화이트 박스 공격(discriminator-white-box attack)을 도시한 도면이다.
- 도 5는 일 실시예에 따른 보안 의료 데이터를 생성하기 위한 적대적 생성 네트워크(GAN; Generative Adversarial Network) 기반 모델을 도시한 도면이다.
- 도 6은 일 실시예에 따른 적대적 생성 네트워크 기반 모델을 학습시키는 방법을 도시한 도면이다.
- 도 7은 일 실시예에 따른 적대적 생성 네트워크를 학습시키는 방법을 도시한 흐름도이다.
- 도 8은 일 실시예에 따른 환자 데이터 항목 별 노이즈를 적용시키는 방법을 도시한 도면이다.
- 도 9는 일 실시예에 따른 무작위 응답(randomized response) 메커니즘을 도시한 도면이다.
- 도 10은 일 실시예에 따른 라플라스 메커니즘(Laplace mechanism)을 도시한 도면이다.
- 도 11은 일 실시예에 따른 보안성 조절 레벨에 기초하여 적대적 생성 네트워크를 학습시키기 위한 교란된 데이터들을 생성하는 방법을 도시한 도면이다.
- 도 12는 일 실시예에 따른 제1 메커니즘 및 제2 메커니즘을 적용하여 적대적 생성 네트워크를 학습시키기 위한 교란된 데이터들을 생성하는 방법을 도시한 도면이다.
- 도 13은 일 실시예에 따른 데이터의 효용성(utility)을 검증하는 방법을 도시한 도면이다.
- 도 14는 일 실시예에 따른 보안 의료 데이터 생성 장치를 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0014] 실시예들에 대한 특정한 구조적 또는 기능적 설명들은 단지 예시를 위한 목적으로 개시된 것으로서, 다양한 형태로 변경되어 구현될 수 있다. 따라서, 실제 구현되는 형태는 개시된 특정 실시예로만 한정되는 것이 아니며, 본 명세서의 범위는 실시예들로 설명한 기술적 사상에 포함되는 변경, 균등물, 또는 대체물을 포함한다.

- [0015] 제1 또는 제2 등의 용어를 다양한 구성요소들을 설명하는데 사용될 수 있지만, 이런 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 해석되어야 한다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소는 제1 구성요소로도 명명될 수 있다.
- [0016] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다.
- [0017] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 설명된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함으로써 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.
- [0018] 본 문서에서, "A 또는 B", "A 및 B 중 적어도 하나", "A 또는 B 중 적어도 하나", "A, B 또는 C", "A, B 및 C 중 적어도 하나", 및 "A, B, 또는 C 중 적어도 하나"와 같은 문구들 각각은 그 문구들 중 해당하는 문구에 함께 나열된 항목들 중 어느 하나, 또는 그들의 모든 가능한 조합을 포함할 수 있다.
- [0019] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 해당 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.
- [0020] 이하, 실시예들을 첨부된 도면들을 참조하여 상세하게 설명한다. 첨부 도면을 참조하여 설명함에 있어, 도면 부호에 관계없이 동일한 구성 요소는 동일한 참조 부호를 부여하고, 이에 대한 중복되는 설명은 생략하기로 한다.
- [0021] 도 1은 멤버십 추론 공격(membership inference attack)을 도시한 도면이다.
- [0022] 공격자(130)는 학습된 기계 학습 모델(machine learning model)(100)에 기초하여, 학습에 사용된 훈련 데이터(110)를 획득할 수 있다. 도 1에서는 공격자에 의한 예시적인 공격 시나리오가 설명된다. 예를 들어, 공격자(130)는 기계 학습 모델(100)의 학습 과정에 참가하지 않고 컴퓨터(computer) 내의 시스템(system)이나 프로그래밍(programming)에 관해 전문적인 지식을 가진 사람을 나타낼 수 있다. 후술하겠으나, 일 실시예에 따른 기계 학습 모델은 전술한 공격자에 의한 공격을 방어할 수 있다.
- [0023] 기계 학습 모델(100)은 기계 학습을 통해 생성될 수 있다. 학습 알고리즘은, 예를 들어, 지도형 학습(supervised learning), 비지도형 학습(unsupervised learning), 준지도형 학습(semi-supervised learning) 또는 강화 학습(reinforcement learning)을 포함할 수 있으나, 전술한 예에 한정되지 않는다. 기계 학습 모델(100)은, 복수의 인공 신경망 레이어들을 포함할 수 있다. 인공 신경망은 심층 신경망(DNN: deep neural network), CNN(convolutional neural network), RNN(recurrent neural network), RBM(restricted boltzmann machine), DBN(deep belief network), BRDNN(bidirectional recurrent deep neural network), 심층 Q-네트워크(deep Q-networks) 또는 상기 중 둘 이상의 조합 중 하나일 수 있으나, 전술한 예에 한정되지 않는다. 참고로 지도형 학습의 경우, 상술한 기계 학습 모델(100)은 훈련 데이터(110) 및 해당 학습 입력에 매핑(mapping)된 학습 출력(120)의 쌍을 포함하는 학습 데이터에 기초하여 학습될 수 있다. 예를 들어, 기계 학습 모델(100)은 훈련 데이터(110)으로부터 학습 출력(120)을 출력하도록 학습될 수 있다. 학습 중의 기계 학습 모델(100)은 훈련 데이터(110)에 응답하여 임시 출력을 생성할 수 있고, 임시 출력 및 학습 출력(120) 간의 손실이 최소화되도록 학습될 수 있다. 학습 과정 동안 기계 학습 모델(100)의 파라미터(예를 들어, 뉴럴 네트워크에서 노드들/레이어들 간의 연결 가중치)가 손실에 따라 업데이트될 수 있다. 이러한 학습은, 예를 들어, 기계학습 모델(100)이 수행되는 전자 장치 자체에서 수행될 수 있고, 별도의 서버를 통해 수행될 수도 있다. 학습이 완료된 기계 학습 모델(100)은 전자 장치의 메모리에 저장될 수 있다.
- [0024] 공격자(130)는 임의 데이터를 기계 학습 모델(100)에 적용하여 획득한 임의 출력에 기초하여, 훈련 데이터(110)를 획득할 수 있다. 전술한 공격자(130)가 훈련 데이터(110)를 획득하는 공격 시나리오는 하기 도 2 내지 도 4에서 후술한다.
- [0025] 도 2는 생성자에 대한 풀 블랙 박스 공격(full black-box attack)을 도시한 도면이다.
- [0026] 공격자(130)는 생성자(200)에서 무작위 데이터들을 생성할 수 있다. 예를 들어, 공격자(130)는 생성된 무작위

데이터들에 기초하여, 타겟 데이터(240)와 유사한 복원 데이터(230)를 획득할 수 있다. 복원 데이터(230)는 공격자(130)에 의해 생성된 무작위 데이터들 중 타겟 데이터(240)와 가장 유사한 데이터를 나타낼 수 있다. 타겟 데이터(240)는 생성자(200) 학습에 사용된 학습 데이터를 나타낼 수 있다. 도 2에서 생성된 영역(210)은 공격자(130)가 생성자(200)를 통해 생성된 무작위 데이터들의 영역을 나타낼 수 있다. 무작위 데이터들은 하기 도 5에서 후술하는 슈도 데이터를 나타낼 수 있다. 실제 영역(220)은 생성자(200)의 학습에 사용된 학습 데이터의 영역을 나타낼 수 있다. 구체적으로, 공격자(130)는 복원 데이터(230) 및 타겟 데이터(240)의 거리 데이터가 타겟 데이터(240)의 임계 거리(250) 안에 속하는 경우, 복원 데이터(230)가 생성자(200)의 학습에 사용된 학습 데이터로 평가할 수 있다.

[0027] 거리 데이터는 다음 수식에 의해 계산될 수 있다:

[0028] [수식 1]

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

[0030] 여기서, X 는 타겟 데이터(240)를 나타낼 수 있고, $\hat{X}$ 은 생성된 영역(210)에 속하는 데이터를 나타낼 수 있다. 또한, L 은 거리 함수(distance function)을 나타낼 수 있다. 예를 들어, 거리 함수는 유클리디안 거리(Euclidean distance)를 사용할 수 있다.

[0031] 도 3은 생성자에 대한 부분 블랙 박스 공격(partial black-box attack)을 도시한 도면이다.

[0032] 공격자(130)는 잠재 벡터(300)에 기초하여 생성자(200)에서 무작위 데이터들을 생성할 수 있다. 잠재 벡터(300)는 생성자(200)에서 무작위 데이터들을 생성하기 위한 생성자(200)의 입력 데이터를 나타낼 수 있다. 예를 들어, 공격자(130)는 생성된 무작위 데이터에 기초하여, 타겟 데이터(240)와 유사한 복원 데이터(230)를 획득할 수 있다. 참고로, 공격자(130)는, 풀 블랙 박스 공격과 달리 부분 블랙 박스 공격에서, 잠재 벡터(300)를 변형함으로써 복원 데이터(230)를 획득할 수 있다. 또한, 공격자(130)는 잠재 벡터(300) 및 복원 데이터(230)에 기초하여 최적의 잠재 벡터를 획득할 수 있다.

[0033] 최적의 잠재 벡터는 다음 수식에 의해 계산될 수 있다:

[0034] [수식 2]

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

[0036] 여기서, $G(z)$ 는 잠재 벡터(300)에 기초하여 생성자(200)에서 생성된 무작위 데이터를 나타낼 수 있다.

[0037] 복원 데이터(230)는 공격자(130)에 의해 생성된 무작위 데이터들 중 타겟 데이터(240)와 가장 유사한 데이터를 나타낼 수 있다. 공격자(130)는 복원 데이터(230) 및 타겟 데이터(240)의 거리 데이터가 타겟 데이터(240)의 임계 거리(250) 안에 속하는 경우, 복원 데이터(230)가 생성자(200)의 학습에 사용된 학습 데이터로 평가할 수 있다.

[0038] 도 4는 판별자에 대한 판별자 화이트 박스 공격(discriminator-white-box attack)을 도시한 도면이다.

[0039] 공격자(130)는 판별자(400)의 판별자 점수(420)를 통해 학습에 사용된 데이터를 획득할 수 있다. 예를 들어, 공격자(130)는 복수의 데이터들(410)의 판별자 점수(420) 중 임계 값을 초과하는 판별자 점수를 포함하는 유추 데이터들(430)을 획득할 수 있다. 복수의 데이터들(410)은 판별자(400)의 학습에 사용된 데이터 또는 판별자(400)의 학습에 사용되지 않았던 데이터 중 적어도 하나를 포함할 수 있다. 판별자(400)는 판별자(400)의 학습에 사용된 데이터에 대해 보다 높은 판별자 점수(420)를 출력하도록 학습될 수 있다. 이와 반대로, 판별자(400)는 판별자(400)의 학습에 사용되지 않았던 데이터에 대해 보다 낮은 판별자 점수(420)를 출력하도록 학습될 수 있다. 여기서, 공격자(130)는 상술한 판별자(400)의 특성을 통해 유추 데이터들(430)을 획득할 수 있다.

[0040] 도 5는 일 실시예에 따른 보안 의료 데이터를 생성하기 위한 적대적 생성 네트워크(GAN; Generative Adversarial Network) 기반 모델을 도시한 도면이다.

[0041] 일 실시예에 따른 전자 장치는 교란된 데이터들을 적대적 생성 네트워크 기반 모델(500)에 적용할 수 있다. 예를 들어, 적대적 생성 네트워크 기반 모델(500)은 생성자(generator)(200) 및 판별자(discriminator)(530)를 포함할 수 있다. 생성자(200)는 도 6에서 후술하는 잠재 공간(latent space)에 포함된 하나의 랜덤 벡터를 입력 받아 슈도 데이터(512)를 출력하는 기계 학습 모델을 나타낼 수 있다. 생성자(200)는 판별자(530)로 하여금 교란된 슈도 데이터(518) 및 교란된 환자 데이터(520)를 구별(discriminate)하지 못하게 슈도 데이터(512)를 생성할 수 있다. 환자 데이터(514)는 복수의 환자에 대한 개인 정보(private information)를 포함하는 데이터를 나타낼 수 있다. 예를 들어, 환자 데이터는 환자의 성별, 나이, 키, 수축기 혈압(Systolic blood pressure), 이완기 혈압(Diastolic blood pressure), 체질량 지수(BMI, body mass index), 또는 시술 경험 중 적어도 하나를 포함할 수 있다. 슈도 데이터(512)는 생성자(200)에 의해 생성된 가짜 데이터를 나타낼 수 있다. 슈도 데이터(512)는 환자 데이터(514)의 종류(type)에 따라 이산 데이터 또는 연속 데이터 중 적어도 하나를 포함할 수 있다. 예를 들어, 이산 데이터는 환자 데이터(514) 중 이산 형태의 값을 가지는 데이터를 나타낼 수 있다. 연속 데이터는 환자 데이터(514) 중 연속적인 실수 값을 가지는 데이터를 나타낼 수 있다. 환자 데이터(514)가 이산 데이터를 포함하는 경우, 슈도 데이터(512)도 이산 데이터를 포함할 수 있고, 환자 데이터(514)가 연속 데이터를 포함하는 경우 슈도 데이터(512)도 연속 데이터를 포함할 수 있다.

[0042] 교란된 데이터들은 원본 데이터에 노이즈가 적용됨으로써 원본과 달라진 데이터를 나타낼 수 있다. 본 명세서에서는 교란된 데이터의 예시로서, 교란된 슈도 데이터(518) 및 교란된 환자 데이터(520)를 설명한다. 교란된 슈도 데이터(518)는 슈도 데이터(512)에 노이즈(516)를 적용하여 생성될 수 있다. 교란된 환자 데이터(520)는 환자 데이터(514)에 노이즈(517)를 적용하여 생성될 수 있다. 생성자(200)는 판별자(530)가 교란된 슈도 데이터(518) 및 교란된 환자 데이터(520)를 구분하지 못하게 학습될 수 있다. 적대적 생성 네트워크 기반 모델(500)을 학습시키는 방법은 하기 도 6에서 후술한다.

[0043] 노이즈(516, 517)는 차분 프라이버시(differential privacy)에 기초하여 수학적으로 디자인된 임의의 실수 값을 나타내는 노이즈 데이터 또는 변경 확률이 적용된 데이터 중 적어도 하나를 나타낼 수 있다. 차분 프라이버시는 프라이버시(privacy)를 정량적으로 모델화하여 프라이버시 보호 정도를 측정하기 위한 방법론을 나타낼 수 있다. 차분 프라이버시를 통해 노이즈(516, 517)를 획득하는 방법은 하기 도 8 내지 도 10에서 후술한다.

[0044] 도 6은 일 실시예에 따른 적대적 생성 네트워크 기반 모델을 학습시키는 방법을 도시한 도면이다.

[0045] 일 실시예에 따른 전자 장치는 보안성 조절 레벨(600)에 기초하여 적대적 생성 네트워크 기반 모델(예: 도 5의 적대적 생성 네트워크 기반 모델(500))을 학습시킬 수 있다. 예를 들어, 보안성 조절 레벨(600)은 노이즈(예: 도 5의 노이즈(516, 517))의 강도(intensity)를 나타내는 파라미터(parameter)로서, 사용자에게 의해 입력될 수 있다.

[0046] 예를 들어, 전자 장치는, 사용자로부터 증가된 보안성 조절 레벨(600)을 입력 받은 경우, 증가된 노이즈 강도로 교란된 데이터들을 획득할 수 있다. 전자 장치는 사용자에게 의해 입력된 증가된 보안성 조절 레벨(600)에 대응하는 노이즈 강도의 노이즈를 데이터에 적용시킬 수 있다. 증가된 강도의 노이즈를 통해, 데이터(예: 슈도 데이터(512) 또는 환자 데이터(514))의 보안성이 개선될 수 있다. 다른 예를 들어, 전자 장치는 데이터의 효용성을 높이기 위해서, 감소된 강도의 노이즈를 나타내는 보안성 조절 레벨(600)에 기초하여, 교란된 데이터들을 생성할 수 있다.

[0047] 구체적으로, 전자 장치는 다음의 수식을 만족시키는 보안성 조절 레벨(600)을 슈도 데이터(512) 및 환자 데이터(514)에 적용시킬 수 있다.

[0048] [수식 3]

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

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

[0050] 여기서 D 은 보안성 조절 레벨(600)을 나타낼 수 있고, D' 와 D' 은 하나의 데이터 포인트에서만 차이가 있는 데이터들의 집합을 나타낼 수 있다. 예를 들어, 데이터들의

집합이 환자 데이터(514)들의 집합인 경우, D 와 D' 은 복수의 환자 데이터 중 하나의 환자 데이터에서만 차이가 있는 데이터들의 집합을 나타낼 수 있다. 또한, A 는 차분 프라이버시 메커니즘(mechanism)을 나타낼 수 있고, S 는 상술한 데이터 집합에서 차분 프라이버시 메커니즘을 적용한 결과들 중 부분 집합(subset)을 나타낼 수 있다. 결과적으로, 전자 장치는 수식 1을 만족하는 보안성 조절 레벨(600)을 슈도 데이터(512) 및 환자 데이터(514)에 적용하여 교란된 슈도 데이터(518) 및 교란된 환자 데이터(520)를 생성할 수 있다. 교란된 데이터들에 기초하여 적대적 생성 네트워크 모델을 학습시키는 방법은 하기 후술한다.

[0051] 일 실시예에 따른 전자 장치는 판별자(530)의 목적함수(610)를 통해 생성자(200) 및 판별자(530)를 학습시킬 수 있다.

[0052] 판별자의 목적함수(610)는 다음 수식에 의해 표현될 수 있다:

[0053] [수식 4]

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

[0054]

[0055] 여기서, $D(x)$ 는 교란된 환자 데이터(520)를 판별자(530)에 적용시킴으로써 획득한 판별자 점수(discriminator score)를 나타낼 수 있고, $D(G(z))$ 는 교란된 슈도 데이터(518)를 판별자(530)에 적용시킴으로써 획득한 판별자 점수를 나타낼 수 있다. 또한, $E_{x \sim p_x} [\log(D(x))]$ 는 교란된 환자 데이터(520) 분포를 따르는 샘플(sample) 환자 데이터에 기초한 $\log(D(x))$ 의 기대값을 나타낼 수 있다. $E_{z \sim p_z} [\log(1 - D(G(z)))]$ 는 교란된 슈도 데이터(518) 분포를 따르는 샘플 슈도 데이터에 기초한 $\log(1 - D(G(z)))$ 의 기대값을 나타낼 수 있다. 판별자 점수는 판별자(530)에 적용된 데이터가 진짜(real) 데이터(예: 교란된 환자 데이터(520))인지 또는 가짜(fake) 데이터(예: 교란된 슈도 데이터(518))인지 여부를 포함하는 점수를 나타낼 수 있다. 예를 들어, 판별자 점수는 판별자(530)에 입력된 데이터가 진짜인 데이터인 경우, 1에 가까운 스코어를 포함할 수 있다. 구체적으로, 전자 장치는 학습이 진행될수록 구별 능력이 좋아지는 판별자(530) 및 판별자(530)의 구별 능력을 속일 수 있는 정교한 생성자(200)를 획득할 수 있다.

[0056] 일 실시예에 따른 전자 장치는 판별자 점수를 획득하기 위해, 와서스타인 거리(Wasserstein distance)에 기초하는 판별자의 목적함수를 통해 상기 판별자 점수를 획득할 수 있다. 와서스타인 거리는 두 개의 확률분포의 연관성을 측정하여 거리의 기대값이 가장 작을 때의 거리를 나타낼 수 있다.

[0057] 와서스타인 거리는 다음의 수식에 의해 계산될 수 있다:

[0058] [수식 5]

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

[0059]

[0060] 여기서, x 는 입력 데이터(예: 환자 데이터 또는 슈도 데이터 중 적어도 하나)를 나타낼 수 있고, $f(x)$ 는 판별자의 출력(예: 판별자 점수)를 나타낼 수 있다. 또한, P_r 은 환자 데이터의 분포를 나타낼 수 있고, P_θ 는 슈도 데이터의 분포를 나타낼 수 있고, $\sup_{\|f\|_{L \leq 1}} |f|$ 는 $|f|$ 의 기울기의 절대값이 1을 넘지 않는 상한(supremum)값을 나타낼 수 있다.

[0061] 도 7은 일 실시예에 따른 적대적 생성 네트워크를 학습시키는 방법을 도시한 흐름도이다.

[0062] 단계(710)에서, 전자 장치는 전자의무기록(EMR, Electronic Medical Records)에서 환자 데이터를 추출할 수 있다. 예를 들어, 전자기록의무는 의료 시스템을 이용하는 환자에 관한 데이터를 효율적이고 통일적으로 관리될

수 있도록 기록된 전산정보를 나타낼 수 있다. 구체적으로, 전자기록의무는 도 8 내지 도 10에서 후술하는 이산 데이터 또는 연속 데이터 중 적어도 하나를 포함할 수 있다.

[0063] 단계(720)에서, 전자 장치는 잠재 공간(latent space)의 잠재 벡터(latent vector)로부터 생성자(예: 도 2의 생성자(200))에 기초하여, 슈도 데이터(예: 도 5의 슈도 데이터(512))를 생성할 수 있다. 슈도 데이터는, 환자 데이터와 달리, 생성자에 의해 단순히 출력된 데이터를 포함할 수 있다. 구체적으로, 슈도 데이터는 환자의 개인 정보를 포함하지 않기 때문에 기관에 완전히 종속되지 않을 수 있고 복수의 법적 규정(legal regulation)에서 상대적으로 자유롭게 활용할 수 있는 데이터를 포함할 수 있다.

[0064] 단계(730)에서, 전자 장치는 환자 데이터 및 슈도 데이터에 지역 차분 프라이버시(local differential privacy)를 적용할 수 있다. 예를 들어, 전자 장치는 환자 데이터가 이진(binary) 데이터인 경우에 기초하여, 지역 차분 프라이버시를 통해 생성된 변경 확률을 환자 데이터 및 슈도 데이터에 적용할 수 있다. 전자 장치는, 환자 데이터가 연속(continuous) 데이터인 경우에 기초하여, 지역 차분 프라이버시를 통해 생성된 노이즈 데이터를 환자 데이터 및 슈도 데이터에 추가할 수 있다. 전자 장치가 환자 데이터 및 슈도 데이터에 지역 차분 프라이버시를 적용하는 방법은 하기 도 8 내지 도 10에서 후술한다.

[0065] 단계(740)에서, 전자 장치는 교란된 환자 데이터(예: 도 5의 교란된 환자 데이터(520) 및 교란된 슈도 데이터(예: 도 5의 교란된 슈도 데이터(518))에 기초하여 적대적 생성 네트워크 기반 모델(예: 도 5의 적대적 생성 네트워크 기반 모델(500))을 학습시킬 수 있다. 구체적으로, 전자 장치는 교란된 환자 데이터 및 교란된 슈도 데이터에 기초하여 적대적 생성 네트워크 기반 모델을 학습시킴으로써, 환자 데이터에 포함된 개인 정보가 외부 공격자로부터 누출되는 것을 방지할 수 있다.

[0066] 도 8은 일 실시예에 따른 환자 데이터 항목 별 노이즈를 적용시키는 방법을 도시한 도면이다.

[0067] 일 실시예에 따른 전자 장치는 보안성 조절 레벨(600)에 기초하여 환자 데이터(514)에 노이즈를 적용시킬 수 있다. 예를 들어, 환자 데이터(514)는 전자 의무 기록(800)으로부터 추출될 수 있다. 환자 데이터(514)는 이산(discrete) 데이터(812) 또는 연속(continuous) 데이터(814) 중 적어도 하나를 포함할 수 있다.

[0068] 일 실시예에 따른 전자 장치는 이산 데이터(812)에 제1 노이즈(830)를 적용하여 교란된 이산 데이터(850)를 획득할 수 있다. 예를 들어, 이산 데이터(812)는 이산 값으로 표현되는 정보를 포함할 수 있다. 구체적으로, 이산 데이터(812)는 이진(binary) 값으로 표현되는 데이터들을 포함할 수 있다. 본 명세서에서는 설명의 편의를 위해 이산 데이터(812)의 예시로 환자의 성별에 대응하는 성별 이진 값(813)을 주로 설명하겠으나, 이로 한정하는 것은 아니다. 성별 이진 값(813)은 남성 환자의 경우 1 및 여성 환자의 경우 0인 이진 값으로 표현되는 데이터를 포함할 수 있다. 제1 노이즈(830)는 무작위 응답(randomized response) 메커니즘(832) 및 변경 확률(834)을 포함할 수 있다. 예를 들어, 무작위 응답 메커니즘(832)은 후술하는 지역 차분 프라이버시의 메커니즘 중 이산 데이터에 관한 대표적인 메커니즘으로 성별 이진 값(813)에 기초하여, 변경 확률(834)에 대응하는 이진 값들 중 다른 이진 값으로 전환된 이진 값을 생성하는 메커니즘을 나타낼 수 있다. 여기서, 전자 장치는 보안성 조절 레벨(600)을 무작위 응답 메커니즘(832)에 적용하여 변경 확률(834)을 획득할 수 있다. 전자 장치는, 획득된 변경 확률(834)을 성별 이진 값(813)에 적용하여 교란된 이산 데이터(850)를 획득할 수 있다. 교란된 이산 데이터(850)는 변경 확률에 의해 원본 데이터(예: 성별 이진 값(813))중 다른 이진 값으로 전환된 이진 값(852)을 포함할 수 있다. 무작위 응답 메커니즘(832)에 관한 설명은 하기 도 9에서 후술한다.

[0069] 일 실시예에 따른 전자 장치는 연속 데이터(814)에 제2 노이즈(840)를 적용하여 교란된 연속 데이터(860)를 획득할 수 있다. 예를 들어, 연속 데이터(814)는 연속 값으로 표현되는 정보를 포함할 수 있다. 본 명세서에서는 설명의 편의를 위해 연속 데이터(814)의 예시로 환자의 신장(height) 정보(815)로 주로 설명하겠으나, 이로 한정하는 것은 아니다. 제2 노이즈(840)는 라플라스 메커니즘(Laplace mechanism)(842) 및 노이즈 데이터(844)를 포함할 수 있다. 예를 들어, 라플라스 메커니즘(842)은 후술하는 지역 차분 프라이버시의 메커니즘 중 연속 데이터에 관한 대표적인 메커니즘으로, 연속 데이터에 노이즈 데이터(844)를 부가하는 메커니즘을 나타낼 수 있다. 여기서, 전자 장치는 보안성 조절 레벨(600)을 라플라스 메커니즘(842)에 적용하여 노이즈 데이터(844)를 획득할 수 있다. 전자 장치는, 획득된 노이즈 데이터(844)를 환자의 신장 정보(815)에 부가하여 교란된 연속 데이터(860)를 획득할 수 있다. 교란된 연속 데이터(860)는 노이즈 데이터(844)가 부가되어 교란된 환자의 신장 정보를 포함할 수 있다. 라플라스 메커니즘(842)에 관한 설명은 하기 도 10에서 후술한다.

[0070] 도 9는 일 실시예에 따른 무작위 응답(randomized response) 메커니즘을 도시한 도면이다.

[0071] 일 실시예에 따른 전자 장치는 이산 데이터(812)에 무작위 응답 메커니즘(832)을 적용하여 교란된 이산 데이터

(850)를 획득할 수 있다. 이산 데이터(812)는 성별 이진 값(813)을 포함할 수 있다. 예를 들어, 전자 장치는 사용자(user)의 입력 또는 미리 결정된(predetermined) 값 중 적어도 하나에 기초하는 보안성 조절 레벨(600)을 무작위 응답 메커니즘(832)에 적용하여 변경 확률(900) 및 유지(maintenance) 확률(902)을 획득할 수 있다. 전자 장치는 변경 확률(900) 및 유지 확률(902)에 기초하여 교란된 이산 데이터(850)를 획득할 수 있다.

[0072] 변경 확률(900)은 다음의 수식에 의해 계산될 수 있다:

[0073] [수식 6]

[0074]

$$p = \frac{e^{\varepsilon}}{e^{\varepsilon} + 1}$$

[0075]

여기서 ε 은 보안성 조절 레벨(600)을 나타낼 수 있고, e 는 자연 로그의 밑(base of the natural logarithm)을 나타낼 수 있고, p 는 변경 확률(900)을 나타낼 수 있다.

[0076]

교란된 이산 데이터(850)는 이산 데이터(812)의 예시인 성별 이진 값(813)에 무작위 응답 메커니즘(832)을 적용한 결과를 포함할 수 있다. 예를 들어, 교란된 이산 데이터(850)는 변경 이산 데이터(904) 및 유지 이산 데이터(906)를 포함할 수 있다. 변경 이산 데이터(904)는 성별 이진 값(813)에서 변경 확률(900)에 의해 변경된 이산 데이터를 나타낼 수 있다. 유지 이산 데이터(906)는 성별 이진 값(813)에서 유지 확률(902)에 의해 유지되는 이산 데이터를 나타낼 수 있다. 결과적으로, 전자 장치는 교란된 이산 데이터(850)를 교란된 슈도 데이터(예: 도 5의 교란된 슈도 데이터(518)) 또는 교란된 환자 데이터(예: 도 5의 교란된 환자 데이터(520)) 중 적어도 하나로서 적대적 생성 네트워크 기반 모델(예: 도 5의 모델(500))을 학습시킬 수 있다.

[0077]

도 10은 일 실시예에 따른 라플라스 메커니즘(Laplace mechanism)을 도시한 도면이다.

[0078]

일 실시예에 따른 전자 장치는 연속 데이터(814)에 라플라스 메커니즘(842)을 적용하여 교란된 연속 데이터(860)를 획득할 수 있다. 연속 데이터(814)는 환자의 신장 정보(815)를 포함할 수 있다. 예를 들어, 전자 장치는 사용자(user)의 입력 또는 미리 결정된(predetermined) 값 중 적어도 하나에 기초하는 보안성 조절 레벨(600)을 라플라스 메커니즘(842)에 적용하여 라플라스 분포를 획득할 수 있다. 본 명세서에서는 설명의 편의를 위해서 복수의 라플라스 분포들을 3개의 라플라스 분포로서 예시를 설명하겠으나 이로 한정하는 것은 아니다. 예를 들어, 전자 장치는 복수의 보안성 조절 레벨(600)에 기초하여 제1 분포(1000) 내지 제3 분포(1020)를 획득할 수 있다. 제1 분포(1000)는 제2 분포(1010) 및 제3 분포(1020)보다 보안성 조절 레벨이 작은 값을 포함할 수 있다.

[0079]

일 실시예에 따른 전자 장치는 연속 데이터(814)에 노이즈 데이터(844)를 부가(add)하여 교란된 연속 데이터(860)를 획득할 수 있다. 교란된 연속 데이터(860)는 연속 데이터(814)의 예시인 환자의 신장 정보(815)에 라플라스 메커니즘(842)을 적용한 결과를 포함할 수 있다. 예를 들어, 교란된 연속 데이터(860)는 환자의 신장 정보(815)에 노이즈 데이터(844)가 더해진 데이터를 포함할 수 있다. 결과적으로, 전자 장치는 교란된 연속 데이터(860)를 교란된 슈도 데이터(예: 도 5의 교란된 슈도 데이터(518)) 또는 교란된 환자 데이터(예: 도 5의 교란된 환자 데이터(520)) 중 적어도 하나로서 적대적 생성 네트워크 기반 모델(예: 도 5의 모델(500))을 학습시킬 수 있다.

[0080]

도 11은 일 실시예에 따른 보안성 조절 레벨에 기초하여 적대적 생성 네트워크를 학습시키기 위한 교란된 데이터들을 생성하는 방법을 도시한 도면이다.

[0081]

일 실시예에 따른 전자 장치는 보안성 조절 레벨(600)에 기초하여 적대적 생성 네트워크 기반 모델을 학습시킬 수 있다. 예를 들어, 전자 장치는 복수의 메커니즘들 중 하나에 보안성 조절 레벨(600)을 적용하여 적대적 생성 네트워크 기반 모델을 학습시킬 수 있다. 복수의 메커니즘들은 무작위 응답 메커니즘(832), 라플라스 메커니즘(842), 가우시안(Gaussian) 메커니즘(1102), 또는 익스포�셜(exponential) 메커니즘(1104) 중 적어도 하나를 포함할 수 있다.

[0082]

일 실시예에 따른 전자 장치는 슈도 데이터(512)에 노이즈 파라미터(1100)를 적용하여 교란된 슈도 데이터(518)를 생성할 수 있다. 전자 장치는 환자 데이터(514)에 노이즈 파라미터(1100)를 적용하여 교란된 환자 데이터(520)를 생성할 수 있다. 노이즈 파라미터(1100)는 환자 데이터(514)의 종류에 따라 선택된 메커니즘에 기초하

여 데이터에 부가되는 노이즈 강도에 관련된 파라미터를 나타낼 수 있다. 구체적으로, 노이즈 파라미터(1100)는 변경 확률(834) 또는 노이즈 데이터(844) 중 적어도 하나를 포함할 수 있다. 예를 들어, 환자 데이터(514)가 이산 데이터(예: 도 8의 이산 데이터(812))인 것에 기초하여, 전자 장치는 무작위 응답 메커니즘(832)으로 변경 확률(834)을 생성할 수 있다.

[0083] 일 실시예에 따른 전자 장치는 라플라스 메커니즘(842)을 통해 복수의 노이즈 데이터(844)들을 생성할 수 있다. 예를 들어, 전자 장치는 보안성 조절 레벨(600)에 기초하여 라플라스 메커니즘(842)에 의해 제1 분포(예: 도 10의 제1 분포(1000))를 획득할 수 있다. 전자 장치는 제1 분포에서 복수의 노이즈 데이터(844)들을 생성할 수 있다. 전자 장치는 복수의 노이즈 데이터(844)에서 2개의 노이즈 데이터를 환자 데이터(514) 및 슈도 데이터(512) 각각에 적용할 수 있다.

[0084] 도 12는 일 실시예에 따른 제1 메커니즘 및 제2 메커니즘을 적용하여 적대적 생성 네트워크를 학습시키기 위한 교란된 데이터들을 생성하는 방법을 도시한 도면이다.

[0085] 일 실시예에 따른 전자 장치는 제1 메커니즘(1202) 및 제2 메커니즘(1204)을 적용하여 교란된 데이터를 생성할 수 있다. 예를 들어, 전자 장치는 복수의 메커니즘(1200) 중에서 적어도 각각 다른 2개의 메커니즘을 선택할 수 있다. 구체적으로, 환자 데이터(514)가 이산 데이터(예: 도 8의 이산 데이터(812))인 경우, 제1 메커니즘(1202) 및 제2 메커니즘(1204)은 모두 무작위 응답 메커니즘(832)을 나타낼 수 있다. 환자 데이터(514)가 연속 데이터(예: 도 8의 연속 데이터(814))인 경우, 제1 메커니즘(1202) 및 제2 메커니즘(1204)은 라플라스 메커니즘(842), 가우시안 메커니즘(1102), 또는 익스포넨셜 메커니즘(1104) 중 적어도 각각 다른 2개의 메커니즘을 나타낼 수 있다.

[0086] 도 13은 일 실시예에 따른 데이터의 효용성(utility)을 검증하는 방법을 도시한 도면이다.

[0087] 일 실시예에 따른 데이터의 효용성은 산점도(scatter plot)(1330) 또는 상관 관계(correlation) 중 적어도 하나에 의해 평가될 수 있다. 데이터의 효용성은 보안 의료 데이터(1300) 및 환자 데이터(1310)간의 유사한 정도를 나타낼 수 있다. 산점도(1330)는 x축의 환자 데이터(1310)와 y축의 보안 의료 데이터(1300)가 만나는 지점을 나타낼 수 있다. 예를 들어, 산점도(1330)는 데이터 항목의 특성에 맞는 스코어(score)를 계산해서 좌표 평면 상에 점으로 나타내는 지점을 포함할 수 있다. 스코어는 데이터 항목의 특성 또는 파악하려는 정보에 따라 DWS(dimension wise statistics), DWA(dimension wise average), 또는 DWP(dimension wise prediction) 중 적어도 하나에 기초하여 획득될 수 있다. 보안 의료 데이터(1300)는 학습된 적대적 생성 네트워크 기반 모델(예: 도 5의 적대적 생성 네트워크 기반 모델(500))의 생성자(예: 도 2의 생성자(200))에 의해 생성된 데이터를 나타낼 수 있다.

[0088] 일 실시예에 따른 데이터의 효용성은 거리 데이터(1340)를 포함할 수 있다. 거리 데이터(1340)는 평가된 데이터들의 효용성을 나타내는 지표를 나타낼 수 있다. 도 13에 도시된 바와 같이, 보안 의료 데이터(1300)의 효용성은 산점도(1330) 및 정비례 효용성 지표(1320)에 기초하여 획득될 수 있다. 정비례 효용성 지표(1320)는 환자 데이터(1310) 및 보안 의료 데이터(1300)가 동일함을 나타낼 수 있다. 결과적으로, 전자 장치는 환자 데이터(1310)의 보안성을 위해 교란된 데이터들(예: 도 5의 교란된 슈도 데이터(518) 또는 교란된 환자 데이터(520))을 적대적 생성 네트워크 기반 모델의 학습을 위해 사용할 수 있고, 실제 환자 데이터(1310)에 유사하게 학습된 생성자에 의해 생성된 보안 의료 데이터(1300)를 통해 데이터의 효용성을 획득할 수 있다.

[0089] 일 실시예에 따른 데이터의 효용성은 보안 의료 데이터(1300) 및 환자 데이터(1310)의 상관 관계에 의해 평가될 수 있다.

[0090] 상관 관계는 다음의 수식에 의해 계산될 수 있다:

[0091] [수식 7]

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

[0092]

[0093] 여기서 CM_{real} 은 환자 데이터(1310)의 상관 관계 행렬(correlation matrix)을 나타낼 수 있고, CM_{fake} 은 보안 의료 데이터(1300)의 상관 관계 행렬을 나타낼 수 있다.

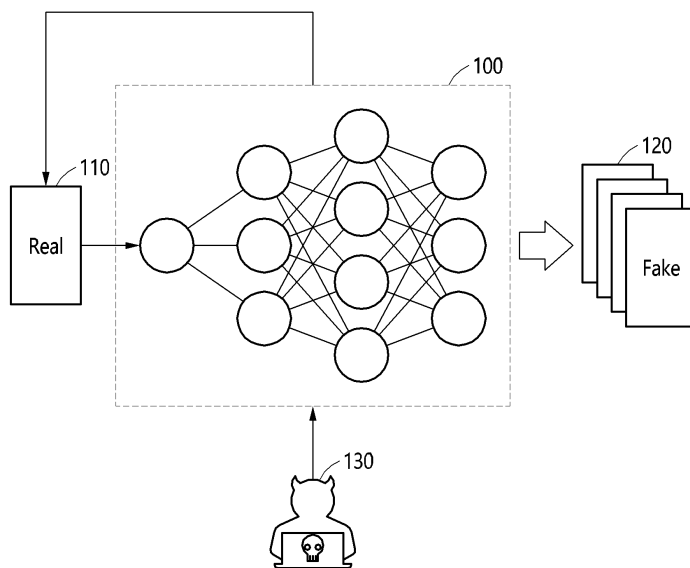
[0094] 도 14는 일 실시예에 따른 보안 의료 데이터 생성 장치를 도시한 도면이다.

- [0095] 일 실시예에 따른 전자 장치(1400)는 프로세서(1410), 메모리(1420), 입출력 인터페이스(1440), 및 통신 모듈(1450)을 포함할 수 있다.
- [0096] 프로세서(1410)는, 교란된 환자 데이터(예: 도 5의 교란된 환자 데이터(520)) 및 교란된 슈도 데이터(예: 도 5의 교란된 슈도 데이터(518))에 기초하여, 생성자(예: 도 5의 생성자(200)) 및 판별자(예: 도 5의 판별자(530))를 포함하는 적대적 생성 네트워크 기반 모델(예: 도 5의 적대적 생성 네트워크 기반 모델(500))을 학습시킬 수 있다. 프로세서(1410)는 소프트웨어를 실행할 수 있고, 프로세서(1410)에 연결된 적어도 하나의 다른 구성요소(예: 하드웨어 또는 소프트웨어 구성요소)를 제어할 수 있다. 프로세서(1410)는 이외에도 다양한 데이터 처리 또는 연산을 수행할 수 있다.
- [0097] 메모리(1420)는 컴퓨터로 실행 가능한 명령어들(1430)을 포함할 수 있다. 메모리(1420)는 적대적 생성 네트워크 기반 모델을 학습시키기 위해 요구되는 다양한 데이터 및/또는 정보를 임시적으로 및/또는 영구적으로 저장할 수 있다. 예를 들어, 메모리(1420)는 교란된 환자 데이터, 교란된 슈도 데이터, 또는 적대적 생성 네트워크 기반 모델 중 적어도 하나를 저장할 수 있다.
- [0098] 입출력 인터페이스(1440)는 입력 장치 및 출력 장치를 포함할 수 있다. 입력 장치는 촉각, 비디오, 오디오 또는 터치 입력을 통해 사용자로부터 입력을 수신할 수 있다. 예를 들어, 입력 장치는 키보드, 마우스, 터치 스크린, 마이크로폰, 또는 사용자로부터 입력을 검출하고, 검출된 입력을 전자 장치(1400)에 전달할 수 있는 임의의 다른 장치를 포함할 수 있다. 출력 장치는 시각적, 청각적 또는 촉각적인 채널을 통해 사용자에게 전자 장치(1400)의 출력을 제공할 수 있다. 출력 장치는 예를 들어, 디스플레이, 터치 스크린, 스피커, 진동 발생 장치 또는 사용자에게 출력을 제공할 수 있는 임의의 다른 장치를 포함할 수 있다. 통신 모듈(1450)은 유선 또는 무선 네트워크를 통해 외부 장치와 통신할 수 있다.
- [0099] 이상에서 설명된 실시예들은 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치, 방법 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 컨트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.
- [0100] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(signal wave)에 영구적으로, 또는 일시적으로 구체화(embodiment)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.
- [0101] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있으며 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

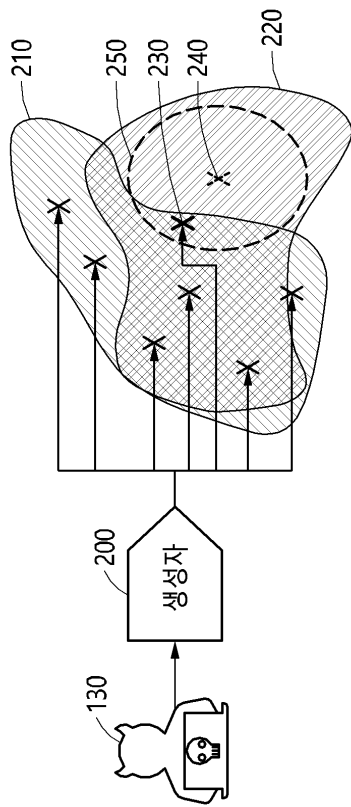
- [0102] 위에서 설명한 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 또는 복수의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.
- [0103] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 이를 기초로 다양한 기술적 수정 및 변형을 적용할 수 있다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.
- [0104] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

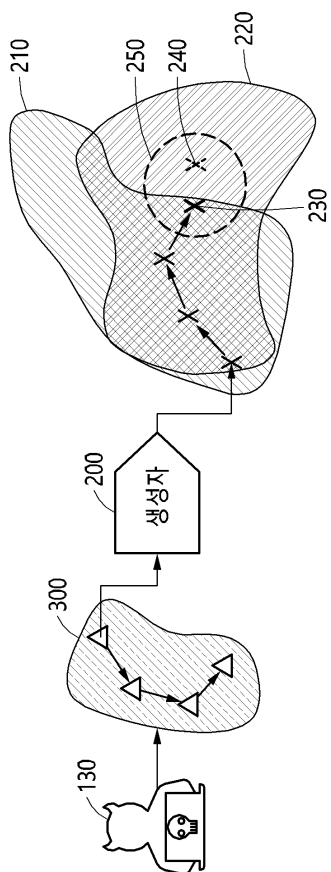
도면1



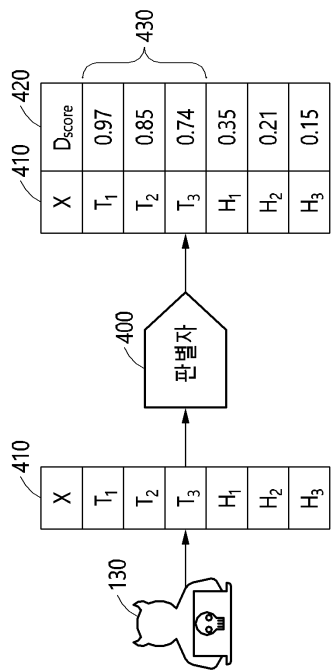
도면2



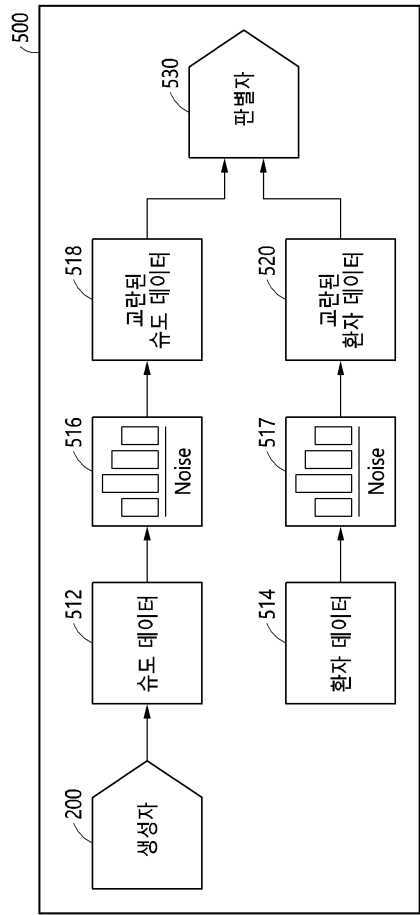
도면3



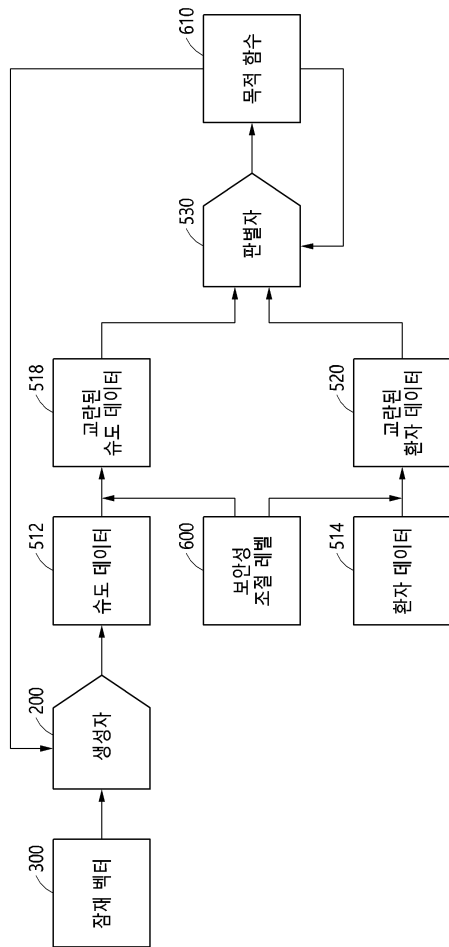
도면4



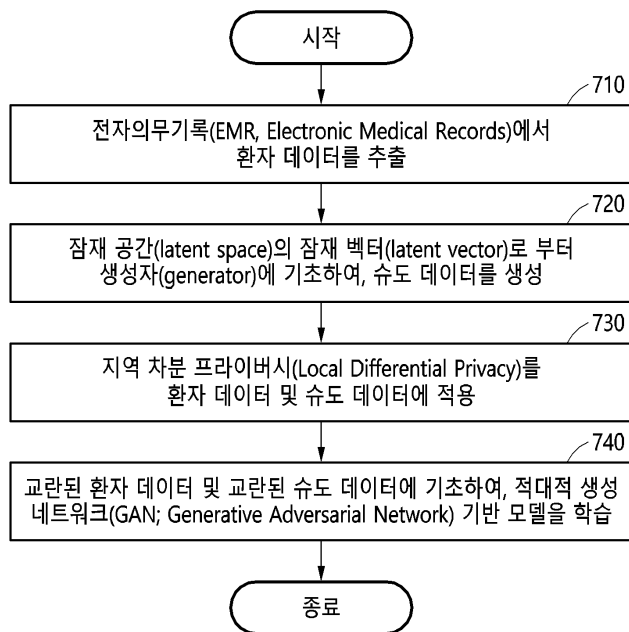
도면5



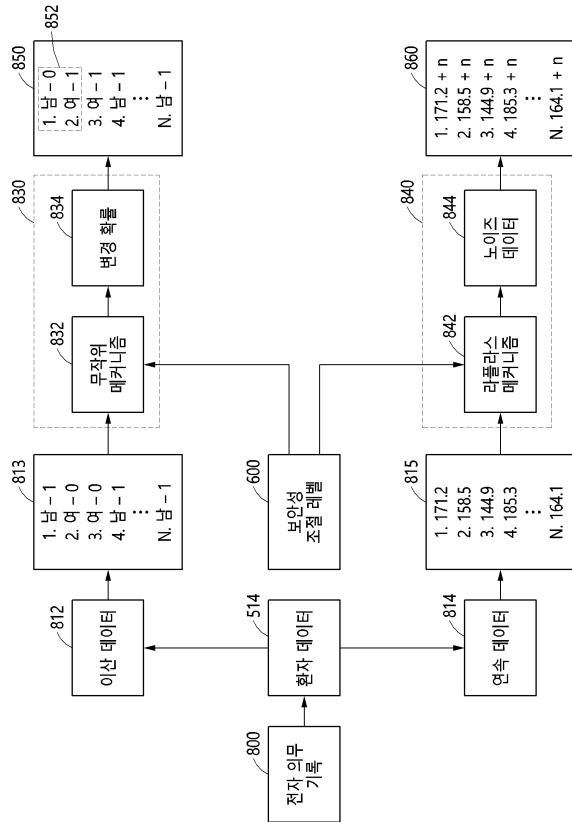
도면6



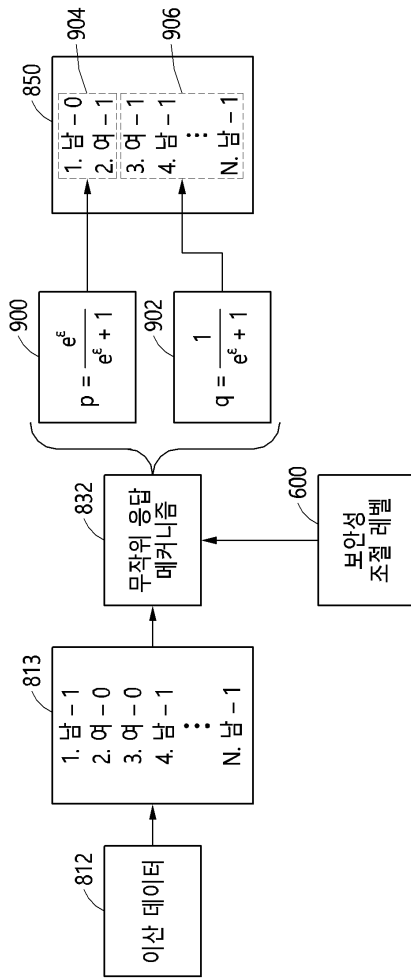
도면7



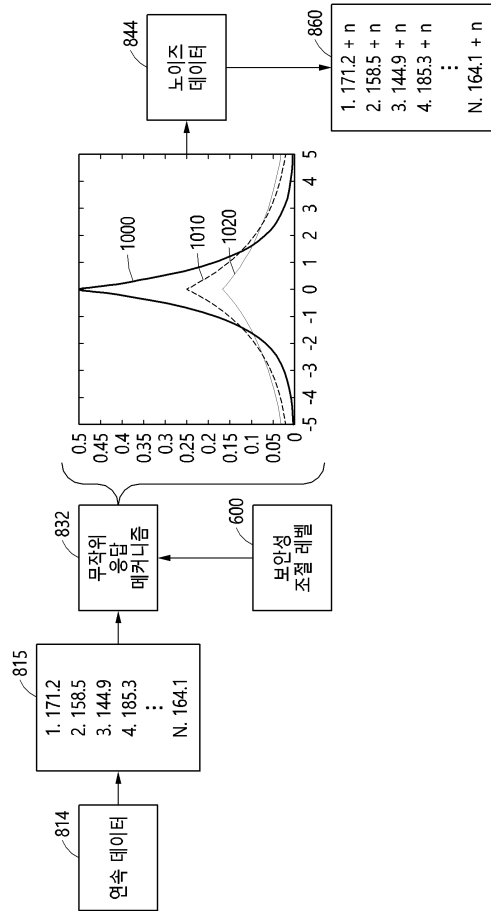
도면8



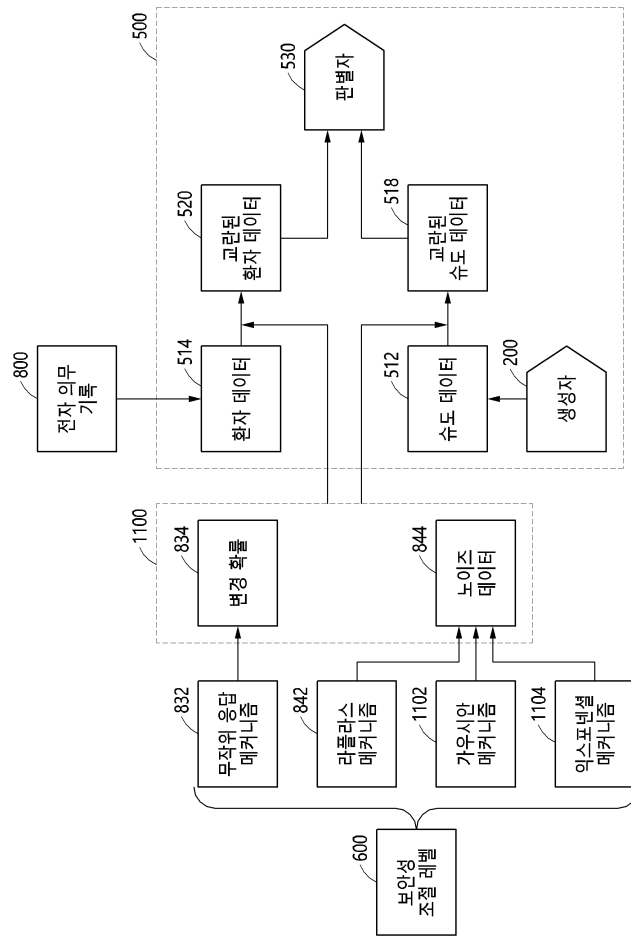
도면9



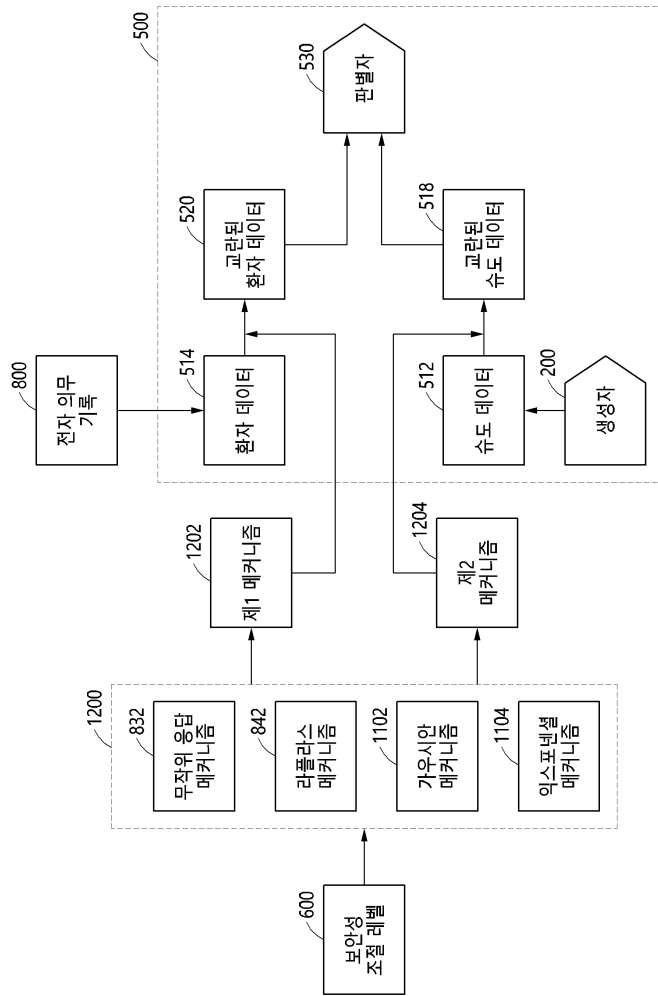
도면10



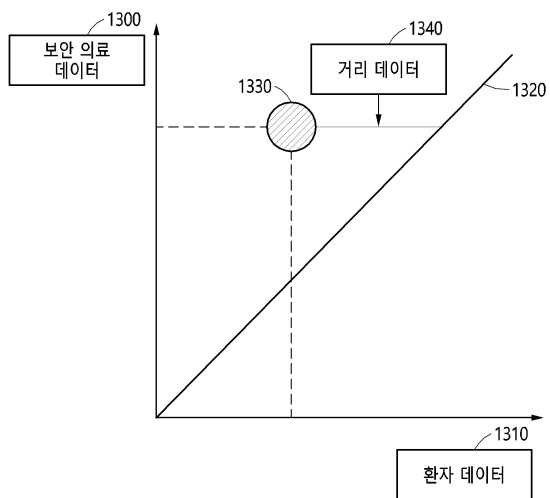
도면11



도면12



도면13



도면14

