Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.35, NO.1, Feb. 2025 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2025.35.1.87

멤버십 추론 공격에 대한 견고성 향상을 위한 심전도 신호 비식별화 방안*

이 한 주,[1†] 김 광 남,[1] 최 석 환[2‡]
[1,2]연세대학교 (대학원생, 교수)

Electrocardiogram Signal De-Identification Methods for Enhancing Robustness against Membership Inference Attacks*

Han-Ju Lee,[1†] Gwang-Nam Kim,[1] Seok-Hwan Choi[2‡]
[1,2]Yonsei University (Graduate student, Professor)

요 약

최근 정보보안의 중요성이 부각됨에 따라, 생체인식 기술이 사용자 신원 확인의 핵심 수단으로 급부상하고 있다. 특히, 심전도 신호의 경우 위변조가 어렵고 높은 정확도를 제공하여 다양한 용도로 활용되고 있으며, 최근에는 인공지능 모델을 활용한 심전도 신호 분석 기술이 활발히 등장하고 있다. 하지만, 이러한 상황에서 심전도 데이터는 멤버십 추론 공격 같은 특정 보안 위협에 취약할 수 있다. 본 논문에서는 멤버십 추론 공격에 대한 견고성을 향상하기 위한 심전도 신호 비식별화 기법을 제안한다. 구체적으로, 제안하는 비식별화 기법은 랜덤 라운딩, 가우시안 잡음, 임펄스 잡음, 정현파 잡음을 기반으로 심전도 신호 데이터에 대한 비식별화를 수행한다. MIT-BIH Arrhythmia Database를 대상으로 한 실험을 통해 제안하는 비식별화 방법이 심전도 신호 분류 모델의 정확도 감소를 최소화하며 멤버십 추론 공격에 대한 견고성을 제공함을 확인하였다.

ABSTRACT

As the importance of information security has become prominent, biometric technologies have rapidly emerged as a key for user identity verification. Especially, electrocardiogram (ECG) signals are used for various purposes in many applications because they are difficult to falsify and provide high accuracy. Moreover, ECG signal analysis technology which analyzes ECG signals using artificial intelligence models has recently been emerging. However, in such scenarios, ECG signal data can be vulnerable to specific security threats like membership inference attack. In this paper, we proposed de-identification methods for ECG signal data to enhance robustness against membership inference attacks. Specifically, the proposed de-identification method uses random rounding, Gaussian noise, impulse noise, and sinusoidal noise to de-identify ECG signal data. From the experimental results using the MIT-BIH Arrythmia Database, it is observed that the proposed de-identification method provides robustness against membership inference attacks while minimizing the decrease in accuracy of ECG signal classification models.

Keywords: De-Identification, Electrocardiogram, Membership Inference Attack, Artificial Intelligence

Received(10. 15. 2024), Modified(12. 23. 2024), Accepted(12. 24. 2024)
* 본 연구는 정부(과학기술정보통신부)의 재원으로 한국 연구재단(RS-2023-00243075)의 지원을 받아 수행된 연구임.
† 주저자, hanleju@yonsei.ac.kr
‡ 교신저자, sh.choi@yonsei.ac.kr(Corresponding author)

I. 서 론

최근 정보 보안이 중요한 이슈로 부상하면서 사용자 신원 확인을 위한 편리하고 안전한 방법으로 생체인식 기술이 크게 주목받고 있다. 생체인식은 개인의 생리학적 및 행동적 특징을 활용하여 사용자의 신원을 확인하는 기술이다. 생체인식에 사용되는 생리학적 특징으로는 홍채, 지문, 안면, 심전도 등이 존재하며 행동적 특징으로는 음성, 걸음걸이 등이 존재한다. 생체인식 기술은 기존의 비밀번호나 별도의 인증키를 사용하는 것과 달리, 별도로 소지하거나 기억할 필요가 없어 사용자에게 편리함을 제공한다[1].

하지만 홍채, 지문 등 생체인식에 사용되는 몇 가지 특징들은 위조 또는 도용에 취약하기에 이를 보완하기 위해 추가적인 보안 시스템이 필요하다[1]. 비밀번호, 핀 번호 등 다른 형태의 인증을 병행하는 멀티팩터인증(MFA) 등이 추가적인 보안 시스템에 해당이 된다[2].

반면 심전도의 경우, 타 생체인식 기술과 마찬가지로 고유한 개인의 신원을 확인하는 데 사용할 수 있으며 타 생체인식 기술과 비교했을 때 위변조가 어렵고 높은 정확도를 제공한다[3]. 이러한 특징을 기반으로 심전도 신호는 생체인식 기술뿐 아니라 다양한 용도로 활용되고 있다. 특히, 인공지능 모델을 통해 환자의 질병 판단, 스트레스 수준 분석, 운동 효율성 평가 등 일상생활과 밀접하게 사용되고 있다[4][5][6]. 하지만, 심전도 신호 역시 보안상 완벽하지 않다. 현재 사용되는 웨어러블 기기나 기타 의료기기를 통해 심전도를 측정하는 과정에서 암호화, 위변조 방지, 인증 등 보안 기법을 적용하기에는 기술적 장벽이 존재한다. 또한, 심전도 신호는 개인의 중요한 정보를 담고 있어, 이러한 데이터의 유출은 심각한 보안 문제를 일으킬 수 있다[7]. 이를 해결하기 위해, 현재 규모가 큰 기업이나 국가에서는 개인 인증을 위한 생체정보를 서버에 저장할 경우, 인증에 필요한 정보만 추출하여 보안 하드웨어에 저장함으로써 보안 위협을 방지하는 경우가 존재한다. 그러나 중앙 집중식 관리 시스템을 사용하거나 규모가 작은 기업에서 이러한 방식을 적용하기에 금전적, 기술적 한계가 존재한다[8].

이러한 상황에서, 심전도 데이터는 멤버십 추론 공격 같은 특정 보안 위협에 취약할 수 있다. 멤버십 추론 공격은 인공지능 모델 내 특정 개인의 데이터 존재 여부를 파악하는 공격 방식으로, 심전도 데이터의 경우 이를 통해 개인의 신원이나 건강 상태와 같은 중요한 정보가 노출될 위험이 있다[9]. 따라서, 심전도 신호가 사용되는 모든 기술은 보안 측면에서 지속적인 개선과 보완이 필요하며, 개인정보 보호를 위한 추가적인 조치가 필수적이다.

본 논문에서는 심전도 신호에 대한 비식별화 절차를 수립하고 이를 통해 보안 위협 중 하나인 멤버십 추론 공격에 대한 견고성을 향상시키는 방안을 제시한다. 구체적으로, 심전도 신호의 비식별화를 위해 데이터 범주화 기법에 해당하는 랜덤 라운딩 기법과 데이터 마스킹 기법에 해당하는 임의 잡음 추가 기법을 사용한다. 이어서 심전도 신호 비식별화 기법이 멤버십 추론 공격에 어떠한 영향을 미치는지 실험적으로 분석한다. 이를 위해 심전도 신호 데이터 비식별화 기법 적용 여부에 따른 심전도 신호 분류 모델의 정확도와 멤버십 추론 공격 성능을 비교하여 비식별화에 따른 멤버십 추론 공격의 견고성 향상을 검증한다.

본 논문의 구성을 요약하면 다음과 같다. 2장에서는 심전도 신호와 비식별화, 멤버십 추론 공격에 대해 소개한다. 3장에서는 본 논문에서 제안하는 비식별화 기법에 대해 상세히 기술한다. 4장에서는 제안하는 기법에 대한 성능 검증 결과를 기술하며 5장에서는 전반적인 내용을 요약 기술한다.

II. 연구 배경

2.1 심전도 신호의 생체인식

심전도는 심장박동 과정에서 발생하는 전기적 신호를 피부에 부착된 전극과 신체 외부의 장비를 통해 측정하여 시각적인 파동을 보여주는 생체신호이다[10].

심전도 신호는 높낮이와 간격에 따라 P파, QRS 복합체, T파로 구성된다. P파는 심방 탈분극이라고 하며, 심방의 수축을 시작하는 신호로 심장의 상부 심방이 수축하여 혈액을 심실로 보내는 과정을 나타낸다. QRS 복합체는 심실 탈분극이라고 하며, 심실에 전기가 흘러가 심실이 기계적 수축하는 심박동 상태를 나타낸다. 마지막으로 T파는 심실 재분극이라고 하며, 심장의 활동이 다시 평상시로 돌아가는 것으로 심실이 다음 박동을 위해 준비하는 과정을 나타낸다[11].

심전도 신호는 심장 질환의 유무, 심장의 위치,

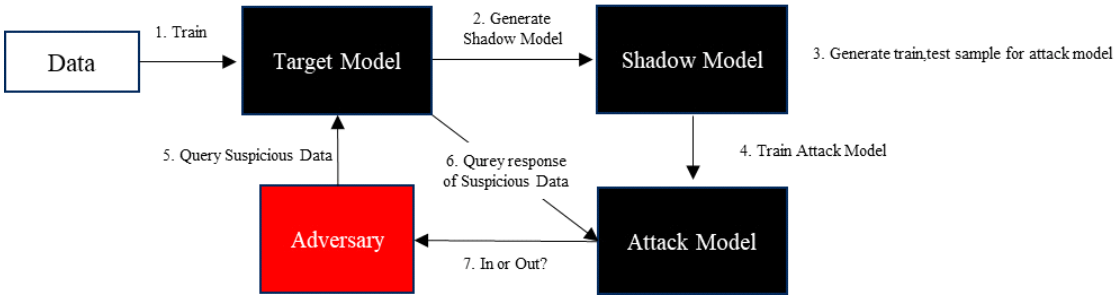

Fig. 1. Overview of Membership Inference Attack Process

크기 등 생리학적 요인에 따라 개인별 고유한 특징을 보유한다. 따라서 심전도 신호는 개인식별 특성 인증에 있어 좋은 수단이 될 수 있다. 또한 시간에 따라 끊임없이 변화하기에 위변조가 어렵다는 장점이 존재한다[3]. 그러나 이러한 고유성 때문에 심전도 신호가 유출될 경우 개인의 신원이 추적될 수 있는 위험성도 존재한다. 특히, 심전도 신호를 활용한 개인 인증이나 의료 진단 과정에서 데이터가 누출되면 개인의 신원 정보뿐만 아니라 의료 기록까지도 악용될 우려가 있어 심전도 신호는 매우 민감한 정보로 간주된다[12].

2.2 의료데이터 비식별화

비식별화란 데이터 내에 개인을 식별할 수 있는 정보가 있는 경우, 이의 일부 또는 전부를 삭제하거나 속성 정보로 대체 처리함으로써 다른 정보와 결합하여도 특정 개인을 식별하기 어렵게 조치하는 것이다[13].

최근 의료데이터 활용에 대한 전망과 기대가 늘어남에 따라 의료데이터 내 민감 정보에 대한 정보 보호 기술과 재식별 위험을 낮추기 위한 비식별화 기술에 대한 요구 또한 늘어나고 있다.

대표적인 비식별화 기법으로는 삭제, 가명처리, 총계처리, 데이터 범주화, 데이터 마스킹이 있다. 삭제는 가장 기본적이고 강력한 비식별화 기법으로 개인 또는 관련한 사물에 고유하게 부여된 값 또는 이름을 삭제함으로써 데이터의 비식별화를 수행한다. 가명처리는 개인식별이 가능한 데이터를 직접적으로 식별할 수 없는 다른 값으로 대체하는 기법으로 일정한 규칙에 따라 데이터를 변형하는 휴리스틱 가명화, 식별 가능한 정보에 암호화 알고리즘을 적용하는 암호화, 기존의 정보를 사전에 정해진 변수와 교환을 하는 교환 방법이 존재한다. 총계처리는 통계값을 적용하여 특정 개인을 식별할 수 없도록 하는 기법이다. 데이터 범주화는 데이터의 값을 범주의 값으로 변환하여 본래의 값을 감추는 기법이다. 마지막으로 데이터 마스킹은 쉽게 개인을 식별할 수 있는 정보를 대체 값으로 변환하여 주요 개인 식별자가 보이지 않도록 처리하는 기법을 의미한다. 상기 비식별화 기법들은 사용 목적에 맞게 단독 또는 복합적으로 사용된다[11].

2.3 멤버십 추론 공격

멤버십 추론 공격은 인공지능 모델에 대한 보안 위협 중 하나로 특정 데이터가 모델의 학습 과정에 사용되었는지를 확인하는 것을 목표로 한다[9].

Fig. 1.은 멤버십 추론 공격의 동작 절차를 나타낸다. 멤버십 추론 공격에서 공격자는 특정 데이터로 학습된 대상 모델(1)에 대한 질의를 수행하여 각 질의에 대한 모델의 예측 결과를 수집한다. 이러한 질의와 모델의 예측 결과는 대상 모델이 특정 입력에 대해 어떻게 반응하는지를 이해하는 데 사용된다. 공격자는 질의의 결과를 바탕으로 대상 모델의 행동을 모방하는 그림자 모델(Shadow Model)을 구축하고(2), 이 모델을 활용하여 학습 데이터셋과 비학습 데이터셋에 대한 예측 결과를 수집한다(3). 이후, 공격자는 수집된 예측 결과에 대한 레이블을 부여하고, 학습 데이터셋에 포함된 데이터와 포함되지 않은 데이터를 구분하는 공격 모델(Attack Model)을 학습한다(4). 학습된 공격 모델은 이진 분류기로서, 특정 데이터에 대한 대상 모델의 예측 결과를 입력받아 해당 데이터가 대상 모델의 학습 데이터셋에 존재하는지를 판별한다. 이어서, 공격자는 대상 모델에 특정 데이터를 질의하여(5) 그 결과를 공격 모델에 입력함으로써(6) 해당 데이터의 학습 사용 여부를 판별한다(7)[9].

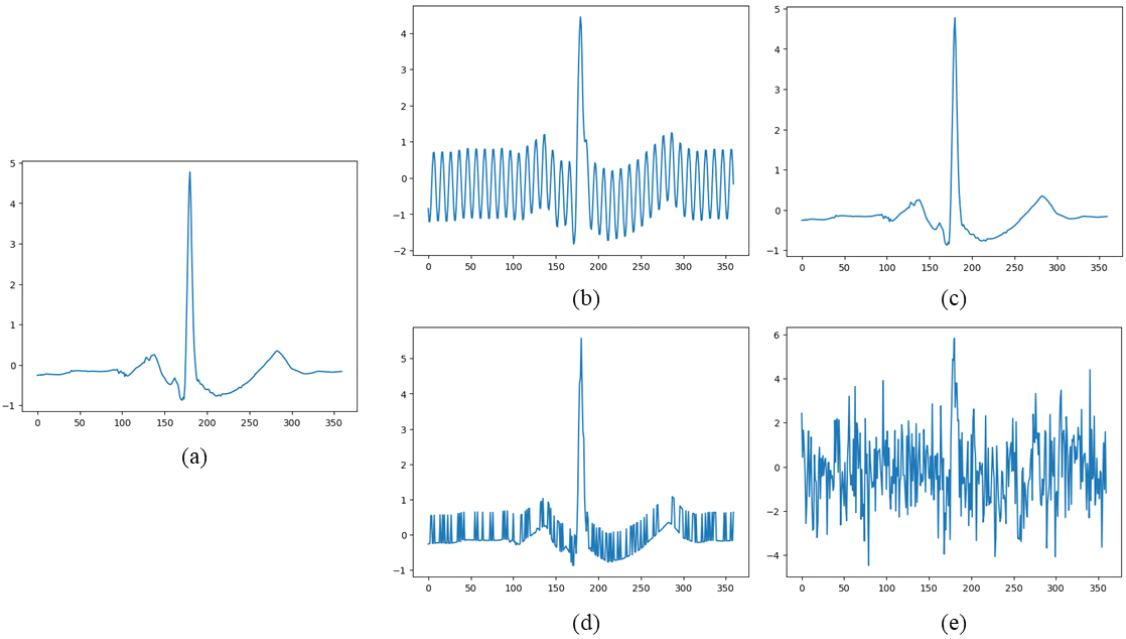

Fig. 2. Transformations of ECG Data with Different De-identification Methods: (a) Original ECG signal, (b) ECG with sinusoidal noise added, (c) ECG post Rounding, (d) ECG with Impulse noise added, (e) ECG with Gaussian noise added

멤버십 추론 공격은 개인 정보 보호와 관련하여 중요한 위협이 된다. 만약, 학습 데이터에 포함된 개인 데이터가 외부에 노출될 경우, 해당 개인의 프라이버시 침해 문제가 발생할 수 있으므로 이러한 공격에 대한 적절한 대응 방안이 요구된다.

III. 본 론

본 장에서는 멤버십 추론 공격에 대한 견고성 향상을 위해 제안하는 심전도 신호 비식별화 방안을 기술한다. 본 논문에서는 4가지 유형의 심전도 신호 비식별화 방안을 제안하며, 데이터 범주화를 기반으로 하는 랜덤 라운딩 기반 비식별화와 데이터 마스킹을 기반으로 하는 가우시안 잡음 기반 비식별화, 임펄스 잡음 기반 비식별화, 정현파 잡음 기반 비식별화로 구분할 수 있다.

3.1 랜덤 라운딩 기반 비식별화

랜덤 라운딩 기반 비식별화는 원본 심전도 신호의 전반적인 추세는 유지하면서 개별 측정값의 식별 가능성을 감소시키기 위해 원본 심전도 신호 데이터셋의 각 데이터 포인트 즉, 심전도 신호 중 하나인 QRS 복합체의 가장 높은 점을 특정 기준에 의해 반올림하여 심전도 신호를 비식별화한다. 모든 데이터 포인트를 가장 가까운 정수로 반올림하거나, 소수점 아래 두 자리까지만 표현하여 나머지를 버리는 방법 등이 이에 해당한다. 본 논문에서는 자릿수(일의 자리, 십의 자리, 백의 자리)를 하이퍼파라미터로 사용하였다.

3.2 가우시안 잡음 기반 비식별화

가우시안 잡음 기반 비식별화는 원본 심전도 신호의 각 데이터 포인트에 정규 분포를 따르는 무작위 잡음을 추가함으로써 원본 심전도 신호 데이터셋의 패턴을 유지하지만 개별 값들을 왜곡하여 심전도 신호를 비식별화한다. 식 (1)은 가우시안 잡음을 나타낸다.

$$N(x\,|\,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (1)$$

여기서 N은 확률 밀도 함수를 나타내며 x는 입

력 변수, μ는 평균, σ^2는 분산을 의미한다. 가우시안 분포의 정규 분포는 평균값이 0에 근접하도록 설정되고 표준편차(σ)를 조절하여 잡음의 강도를 결정한다[14][15]. 본 논문에서는 표준편차를 하이퍼파라미터로 사용하였다.

3.3 임펄스 잡음 기반 비식별화

임펄스 잡음 기반 비식별화는 무작위로 선택된 원본 심전도 신호 데이터셋의 데이터 포인트에 특정 값을 추가하여 데이터 일부의 극단적인 변경을 통해 원본 신호의 일부를 손실시키며, 식별이 가능한 특정 패턴이나 특징을 흐리도록 심전도 신호를 비식별화한다. 본 논문에서는 잡음의 크기를 하이퍼파라미터로 사용하였다.

3.4 정현파 잡음 기반 비식별화

정현파 잡음 기반 비식별화는 원본 심전도 데이터셋의 주기적인 특성을 마스킹하거나 변형시키기 위해 심전도 신호에 특정 주기와 진폭을 가진 정현파 형태의 잡음을 추가하여 심전도 신호를 비식별화한다. 식 (2)는 정현파 잡음을 나타낸다.

$$N(t) = Asin(2\pi ft + \phi) \quad (2)$$

여기서 N은 시간(t)에 대한 잡음 신호를 나타내며, A는 잡음의 진폭, f는 잡음의 주파수, ϕ는 신호가 시작하는 시점의 각도로 잡음 신호의 위상을 나타낸다[15]. 정현파 잡음은 주파수, 진폭, 위상의 값을 변경하여 다양한 형태의 잡음을 생성할 수 있다. 본 논문에서는 진폭을 하이퍼파라미터로 사용하였다.

IV. 실험 및 결과

본 장에서는 제안하는 심전도 신호 비식별화 기법이 멤버십 추론 공격에 대한 견고성에 미치는 영향을 실험을 통해 분석한다. 이를 위해, 먼저 원본 심전도 신호 데이터셋으로 학습한 분류 모델과 비식별화 처리된 심전도 신호 데이터셋으로 학습한 분류 모델의 성능을 비교하여 제안하는 비식별화 기법이 모델의 분류 정확도에 미치는 영향을 분석한다. 이어서, 원본 심전도 신호 데이터셋으로 학습한 분류 모델과 비식별화된 심전도 신호 데이터셋으로 학습한 분류 모델을 대상으로 멤버십 추론 공격을 수행하여 제안하는 비식별화 기법이 멤버십 추론 공격에 대한 견고성에 미치는 영향을 분석한다.

4.1 데이터셋

본 논문에서는 MIT-BIH Arrhythmia Database를 사용하여 제안하는 비식별화 기법의 평가를 진행하였다. MIT-BIH Arrhythmia Database는 심장 질환 진단과 연구에 사용되는 데이터셋으로, 47명의 개인으로부터 얻은 심전도 기록을 포함하고 있으며 기록의 각 채널은 초당 360개의 샘플로 디지털화되어 있다[17]. 본 논문에서는 데이터셋에 포함된 19개의 비트 레이블 중 발생 빈도가 높은 5가지 비트, 즉 정상 박동(N), 좌심실 분지 블록 박동(L), 우심실 분지 블록 박동(R), 심방 조기 박동(A), 심실 조기 박동(V)을 대상으로 실험을 진행하였다.

4.2 실험 설계

본 실험에서 사용된 심전도 신호 데이터는 시간에 따라 변화하는 시계열 데이터로 시계열 데이터의 복잡한 패턴과 시간적 의존성을 학습하는 데 강점을 가지는 LSTM 모델과 GRU 모델을 분류 모델로 선정하여 실험을 진행하였다[18]. 또한, 본 실험에서는 학습 데이터셋과 검증 데이터셋 모두 비식별화 기법을 적용하여 실험을 진행하였다.

멤버십 추론 공격 실험은 TensorFlow-Privacy에서 제공하는 멤버십 추론 공격 기법을 활용하여 진행되었으며, 모델 학습에는 배치 크기 512, 총 25에포크로 설정하였다. 각 공격 결과는 실험을 10회 반복하여 평균값을 사용하였다. 멤버십 추론 공격에 필요한 그림자 모델은 대상 모델과 동일하게 사용하여 대상 모델의 행동을 모방하였다. 이 과정에서 전체 학습 데이터셋을 세 가지로 분할하여 사용하였다. 첫 번째 데이터셋은 대상 모델을 학습하기 위해 사용했으며, 두 번째 데이터셋은 그림자 모델을 학습하는 데 사용하였다. 이후 공격 모델에 대한 데이터셋을 구축하기 위해 마지막 데이터셋은 공격 모델의 학습에 사용하였다. 본 논문에서는 멤버십 추론 공격의 공격 모델로 Multi-Layered-Perceptron(MLP)을 사용하였다. MLP는 다층 구조로 이루어져 있어 데이터의 복잡한 특성을 다차원적으로 학습할 수 있게 해주며 비선형 활성화 함수를 사용하기 때문에 심

Table 1. LSTM Classification Results

Method	Hyper Parameter	Train ACC	Test ACC
Origin	None	0.987	0.986
Rounding	1	0.955	0.957
	2	0.965	0.963
	3	0.943	0.941
Gaussian	0.3	0.973	0.971
	0.8	0.967	0.957
	1.5	0.870	0.869
Impulse	1	0.985	0.985
	2	0.986	0.985
	4	0.979	0.981
sinusoidal	0.3	0.986	0.986
	0.5	0.988	0.986
	1	0.984	0.982
DP	1	0.891	0.884
	2	0.935	0.926
	3	0.960	0.961

Table 2. GRU Classification Results

Method	Hyper Parameter	Train ACC	Test ACC
Origin	None	0.988	0.975
Rounding	1	0.989	0.974
	2	0.967	0.965
	3	0.962	0.964
Gaussian	0.3	0.982	0.975
	0.8	0.981	0.971
	1.5	0.864	0.867
Impulse	1	0.978	0.976
	2	0.986	0.974
	4	0.958	0.963
sinusoidal	0.3	0.984	0.977
	0.5	0.982	0.979
	1	0.971	0.969
DP	1	0.883	0.871
	2	0.910	0.904
	3	0.951	0.949

전도 신호와 같은 복잡한 시계열 데이터에서 발견되는 비선형 패턴을 효과적으로 포착한다는 장점이 있다[19].

또한, 제안한 비식별화 기법의 성능을 평가하기 위해 대표적인 기존 비식별화 기법인 차등 프라이버시(Differential Privacy, DP) 기법을 적용하여 비교 분석을 수행하였다. 차등 프라이버시는 데이터에 노이즈를 추가하여 민감한 정보의 노출을 방지하는 강력한 프라이버시 보호 방법으로 널리 사용되고 있다. 차등 프라이버시 기법은 설정된 epsilon 값에 따라 노이즈의 크기를 조절하며, epsilon 값이 작을수록 더 큰 노이즈가 추가되는 것을 의미한다. 이는 프라이버시와 데이터 정확도 간의 균형을 맞추는 중요한 요소로 작용한다[20]. 본 연구에서는 라플라스 분포 기반의 노이즈를 적용하여 데이터를 변형하였고, 제안된 기법과의 성능 비교 진행하였다.

4.3 실험 결과

4.3.1 심전도 신호 분류 모델 성능 비교

Table 1과 Table 2는 제안하는 심전도 신호 데이터 비식별화 기법과 차등 프라이버시를 적용한 결과에 따른 심전도 신호 분류 모델의 정확도를 나타낸다. 실험 결과에 따르면, 제안하는 비식별화 기법은 심전도 신호 분류 모델의 정확도에 큰 영향을 미치지 않았다. 예를 들어, LSTM 모델을 사용한 Table 1에서 원본 심전도 신호 데이터셋으로 학습한 분류 모델의 경우 테스트 정확도가 98.6%를 나타냈으며 임펄스 잡음 기반 비식별화 데이터셋으로 학습한 분류 모델의 경우 테스트 정확도가 98.5%를 나타내었다. 이는 제안하는 비식별화 기법이 심전도 신호 데이터의 유용성을 보존할 수 있음을 의미한다.

또한, 제안하는 각각의 비식별화 기법은 파라미터 값에 따라 다양한 성능을 보였다. 특히, 가우시안 잡음 기반 비식별화가 두 번의 하이퍼파라미터 증가에 따른 분류 모델 성능에 있어 97.1%에서 86.9%로 감소하여 파라미터값에 가장 민감함을 보였다. 이어서, 랜덤 라운딩 기반 비식별화 기법이 하이퍼파라미터의 증가로 분류 모델의 성능이 95.7%에서 94.1%로 감소하여 가우시안 잡음 기반 비식별화 다음으로 파라미터값에 민감함을 보였다. 마지막으로, 제안하는 심전도 신호 비식별화 기법 중 임펄스 잡음 기반 비식별화와 정현파 잡음 기반 비식별화는 하이퍼파라미터 증가에 따라 각각 98.5%에서 98.1%, 98.6%에서 98.2%로 분류 모델의 정확도가 감소하여 두 가지 비식별화의 하이퍼 파라미터값은 분류 모델의 정확도에 영향을 덜 미치는 것을 확인할 수 있

Table 3. LSTM Membership Inference Attack Results

Method	Hyper parameter	AUC
Origin	None	0.684
Rounding	2	0.672
Gaussian	0.8	0.538
Impulse	2	0.676
sinusoidal	0.5	0.629
DP	3	0.695

었다.

Table 2에서 GRU 모델을 사용한 결과, 각 비식별화 기법에 따른 성능이 Table 1에서 관찰된 모델들과 유사한 경향을 보이며 비식별화 적용 후에도 GRU 모델의 테스트 정확도는 기존 데이터 대비 유사하거나 약간의 차이를 보이는 수준으로 유지되었다. 이는 GRU 모델에서도 비식별화 기법이 심전도 신호 데이터의 유용성을 효과적으로 보존할 수 있음을 입증한다.

반면, Table 1과 2의 차등 프라이버시 기법을 적용한 실험 결과에서, 차등 프라이버시 기법의 경우 epsilon 값이 감소할수록 심전도 신호 분류 모델의 정확도가 점차 낮아지는 경향을 확인할 수 있다. 즉, 차등 프라이버시 기법은 심전도 신호 데이터의 유용성을 보존하지 못함을 의미한다.

4.3.2 멤버십 추론 공격 성능 비교

본 논문에서는 멤버십 추론 공격에 대한 제안하는 심전도 신호 비식별화 방법을 평가하기 위해 Receiver Operating Characteristics(ROC) curve의 Area Under the Curve(AUC)를 평가 메트릭으로 활용하였다. ROC curve는 멤버십 추론 공격에 대한 취약성을 나타내는 지표로 사용되며 ROC curve의 AUC 값이 0.5는 무작위 추측과 동일한 수준을 의미하며, 1.0은 공격자가 완벽하게 추론할 수 있음을 의미한다[20].

Table 3을 통해 LSTM 모델을 사용했을 때의 공격 실험 결과를 확인할 수 있다. 원본 심전도 신호 데이터셋으로 학습한 분류 모델에 대한 멤버십 추론 공격 결과, AUC 값은 0.684로 나타났다. 반면에, 제안하는 심전도 신호 비식별화 기법이 적용된 데이터셋으로 학습한 분류 모델의 경우 AUC 값이 전반적으로 하락하는 경향을 보였다. 이는 제안하는 비식별화 기법이 데이터에 대한 멤버십 추론 공격의 견고성을 향상시킨다는 것을 의미한다. 특히, 정현파 잡음 기반 비식별화 기법과 가우시안 잡음 기반 비식별화 기법이 검증 정확도 감소 대비 공격 취약성이 큰 폭으로 감소하는 것을 확인할 수 있었다. 구체적으로, 가우시안 잡음 기반 비식별화 시 원본 데이터 대비 검증 정확도 감소율은 약 2.94%이지만, 모델의 공격 취약성 감소율은 21.55%로 나타났다. Fig. 3.에서 원본 데이터의 멤버십 추론 공격 AUC 결과와 가우시안 잡음 기반 비식별화 시 AUC 결과를 확인할 수 있다.

Table 4. GRU Membership Inference Attack Results

Method	Hyper parameter	AUC
Origin	None	0.552
Rounding	2	0.547
Gaussian	0.8	0.540
Impulse	2	0.526
sinusoidal	0.5	0.543
DP	3	0.583

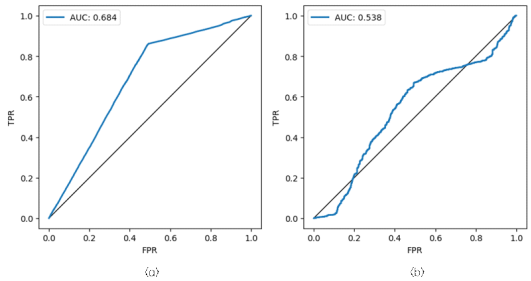

Fig. 3. (a) AUC results from a membership inference attack on the original (b) AUC results after de-identification, where the guassian noise's standard deviation value is set to 0.8

Table 4에서는 GRU 모델을 사용한 멤버십 추론 공격 실험 결과를 확인할 수 있다. GRU 모델을 사용한 실험 결과는 앞서 LSTM 모델에서 관찰된 경향과 유사하게 나타났다. 원본 심전도 신호 데이터셋으로 학습한 분류 모델에 대한 멤버십 추론 공격 결과, AUC 값은 0.552로 나타났다. 이후, 제안하는 심전도 신호 비식별화 기법이 적용된 데이터셋으로 학습한 분류 모델의 경우 AUC 값이 전반적으로

Table 5. Top-5 epsilon lower bounds of original signal

Slice	Epsilon 1	Epsilon 2	Epsilon 3	Epsilon 4	Epsilon 5
Class=0 (N)	1.6641	1.4386	1.3414	1.3412	1.3411
Class=1 (L)	1.1943	1.1941	1.1935	1.1934	1.1931
Class=2 (R)	1.7686	1.6837	1.6162	1.6031	1.5912
Class=3 (A)	0.6376	0.6140	0.6006	0.4354	0.4342
Class=4 (V)	1.0304	1.0295	1.0293	1.0291	1.0288

Table 6. Top-5 epsilon lower bounds of de-identification with gaussian noise set to a standard deviation of 0.8

Slice	Epsilon 1	Epsilon 2	Epsilon 3	Epsilon 4	Epsilon 5
Class=0 (N)	1.3653	1.3653	1.3653	1.3653	1.3653
Class=1 (L)	1.2133	1.2117	1.2108	1.2103	1.2098
Class=2 (R)	1.4017	1.3990	1.3978	1.301	1.3892
Class=3 (A)	1.4061	1.3989	1.3877	1.3794	1.3738
Class=4 (V)	1.0973	0.7887	0.7177	0.7000	0.6901

하락하는 경향을 보였다. 특히, 임펄스 잡음 기반 비식별화 결과 모델의 공격 취약성 감소율은 4.72%로 검증 정확도 감소율 대비 큰 폭으로 감소한 것을 확인할 수 있다.

또한, Table 3, Table 4의 결과를 통해 심전도 신호 데이터의 유용성을 보존하는 수준에서 비식별화 정도를 추가했을 때, 제안된 비식별화 기법은 멤버십 추론 공격에서 AUC 값이 하락하는 경향을 보인 것을 확인할 수 있다. 반면, 차등 프라이버시 기법을 적용하였을 때 멤버십 추론 공격에서 AUC 값이 증가하는 경향을 확인할 수 있다. 이는 제안된 비식별화 기법이 차등 프라이버시 기법 대비 데이터 유용성을 유지하면서도 멤버십 추론 공격에 대한 견고성을 효과적으로 확보할 수 있음을 나타낸다.

Table 5와 6은 LSTM 모델을 사용했을 때 제안하는 비식별화 기법 적용 여부에 따른 멤버십 추론 공격의 epsilon 값을 나타낸다. Epsilon은 각 클래스별 프라이버시 손실을 나타내는 지표로, 값이 낮을수록 적대적 공격에 대한 견고성이 높다는 것을 의미한다[21]. 원본 심전도 신호 데이터셋으로 학습한 분류 모델의 결과(Table 5)와 제안하는 가우시안 잡음 기반 비식별화 기법을 적용한 데이터셋으로 학습한 분류 모델의 결과(Table 6)를 비교하였을 때, MIT-BIH Arrhythmia Database 대부분의 클래스에서 Epsilon 값이 감소하는 것을 확인 할 수 있었다.

V. 결 론

본 논문에서는 멤버십 추론 공격에 대한 견고성 향상을 위한 심전도 신호 비식별화 방안을 제안하고, 이를 MIT-BIH Arrhythmia Database를 이용해 검증하였다. 실험 결과를 통해 제안하는 비식별화 방법은 심전도 신호 분류 모델의 정확도 감소를 최소화하며 멤버십 추론 공격에 대한 견고성을 제공함을 보였다. 그러나 본 논문에서 제시한 비식별화 기법에는 한계점이 존재한다. 전체 데이터셋에 대한 일괄적인 비식별화는 식별에 중요한 역할을 하는 데이터까지도 영향을 받게 만들어 식별이나 데이터 활용의 관점에서 볼 때 중요한 문제점이 될 수 있다. 따라서 향후 연구에서는 인공지능 모델에서 중요한 데이터를 구별하는 Explainable AI(XAI) 등을 사용할 예정이다. 이를 활용하여 핵심 데이터를 식별하고 주요하지 않은 데이터에 대해서만 비식별화를 적용하여 데이터 활용성과 비식별화 효과 간의 균형을 효과적으로 맞추는 방향으로 연구를 진행할 계획이다.

References

[1] M. Ingale, P. Kulkarni, and S. R. Kolhe, “Ecg biometric authentication: A comparative analysis,” IEEE Access, vol. 8, pp. 117853 - 117866, Aug. 2020.

[2] H. Abdi and D. Valentin, "Multiple factor analysis (MFA)," Encyclopedia of Measurement and Statistics, pp. 657 - 663, 2007.
[3] R. D. Labati, A. Genovese, E. Munoz, and F. Scotti, "Deep-ECG: Convolutional neural networks for ECG biometric recognition," Pattern Recognition Letters, vol. 126, pp. 78 - 85, May 2019.
[4] K. C. Siontis, P. Grammatikopoulos, P. G. Kallianos, and A. E. Attia, "Artificial intelligence-enhanced electrocardiography in cardiovascular disease management," Nature Reviews Cardiology, vol. 18, no. 7, pp. 465 - 478, Jul. 2021.
[5] Q. Lin, H. Zhao, and Z. Wang, "Advanced artificial intelligence in heart rate and blood pressure monitoring for stress management," Journal of Ambient Intelligence and Humanized Computing, vol. 12, pp. 3329 - 3340, Nov. 2021.
[6] B. R. Chaitman, J. R. Lesperance, and A. Bourassa, "Improved efficiency of treadmill exercise testing using a multiple lead ECG system and basic hemodynamic exercise response," Circulation, vol. 57, no. 1, pp. 71 - 79, Jan. 1978.
[7] W. Yang and S. Wang, "A privacy-preserving ECG-based authentication system for securing wireless body sensor networks," IEEE Internet of Things Journal, vol. 9, no. 8, pp. 6148 - 6158, Apr. 2021.
[8] M. I. Mihailescu and S. L. Nita, "A searchable encryption scheme with biometric authentication and authorization for cloud environments," Cryptography, vol. 6, no. 1, p. 8, Jan. 2022.
[9] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), pp. 3 - 18, May 2017.
[10] Z. F. M. Apandi, R. Ikeura, and S. Hayakawa, "Arrhythmia detection using MIT-BIH dataset: A review," Proceedings of the 2018 International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA), pp. 1 - 6, Aug. 2018.
[11] P. U. Ilavarasi and N. Jeevitha, "Automatic detection of arrhythmia using LabVIEW and MATLAB," Australian Journal of Basic and Applied Sciences, vol. 10, no. 5, pp. 20 - 23, 2016.
[12] A. Ghazarian, J. Zheng, and C. Rakovski, "Privacy-preserving ECG data analysis with differential privacy: A literature review and a case study," arXiv preprint arXiv:2406.13880, Jun. 2024.
[13] S. Garfinkel, De-identification of personal information, National Institute of Standards and Technology, 2015.
[14] M. A. Selami and A. F. Fadhil, "A study of the effects of Gaussian noise on image features," Kirkuk University Journal-Scientific Studies, vol. 11, no. 3, pp. 152 - 169, 2016.
[15] V. M. Hidalgo, J.-C. Letelier, and J. Díaz, "The amplitude modulation pattern of Gaussian noise is a fingerprint of Gaussianity," arXiv preprint arXiv:2203.16253, Mar. 2022.
[16] P. K. Kythe, Sinusoids: Theory and Technological Applications, CRC Press, 2014.
[17] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," IEEE Engineering in

Medicine and Biology Magazine, vol. 20, no. 3, pp. 45 - 50, May 2001.

[18] B. Lindemann, F. Santoro, and K. Schwenk, "A survey on long short-term memory networks for time series prediction," Procedia CIRP, vol. 99, pp. 650 - 655, 2021.

[19] T. Bikku, "Multi-layered deep learning perceptron approach for health risk prediction," Journal of Big Data, vol. 7, no. 1, p. 19, 2020.

[20] M. Abadi, A. Chu, I. Goodfellow, and H. B. McMahan, "Deep learning with differential privacy," Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308 - 318, Oct. 2016.

[21] H. Hu, A. Sen, and K. K. Kothari, "Membership inference attacks on machine learning: A survey," ACM Computing Surveys (CSUR), vol. 54, no. 11s, pp. 1 - 37, Dec. 2022.

〈저자소개〉

이 한 주 (Han-Ju Lee) 학생회원
2024년 2월: 연세대학교 컴퓨터정보통신공학부 공학사
2024년 3월~현재: 연세대학교 전산학과 석사과정
〈관심분야〉 정보보안, 적대적 공격, 인공지능 등

김 광 남 (Gwang-Nam Kim) 학생회원
2024년 2월: 연세대학교 컴퓨터정보통신공학부 공학사
2024년 3월~현재: 연세대학교 전산학과 석사과정
〈관심분야〉 인공지능, 딥러닝, AI 보안 등

최 석 환 (Seok-Hwan Choi) 정회원
2016년 8월: 부산대학교 정보컴퓨터공학부 공학사
2022년 8월: 부산대학교 정보융합공학사 공학박사
2022년~현재: 연세대학교 소프트웨어학부 교수
〈관심분야〉 AI 보안, 정보보안, 네트워크 보안 등