

심전도 사용자인증과 적대적 공격

스마트서비스응용

202350092 이새봄

프로젝트 목표

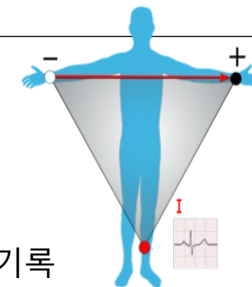
- 최근 바이오 인증 시스템의 위·변조 위험을 해결하고자 심전도 사용자 인증 기술 연구가 활발히 진행되고 있음
- 그러나, 딥러닝을 이용한 심전도 사용자 인증 모델은 사용자의 오분류를 유도하는 적대적 공격에 취약함
- 따라서, 적대적 예제를 사용하여 심전도 사용자 인증에서의 적대적 공격이 보안에 얼마나 취약한지 분석하고자 함

데이터셋

• 심전도 데이터셋 (비공개)

측정기간	2016.08.23 ~ 2016.12.27	측정담당자	최규호
측정 인원	100명 : 조선대학교 IT융합 대학 대학원생 및 학부생	피험자 상태 및 조건	의자에 앉은 편안한 상태
측정 시간	1회 측정 시간 : 10초 총 60회 측정	데이터 sampling rate	50만 Hz
심전도 유형	심전도 Lead-I	전극 유형	습식 전극

Table 1. ECG Dataset



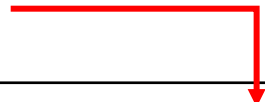
* Lead I : 오른팔이 기준 전극이 되고, 왼팔이 기록 전극이 됨. 0° 각도를 기록

데이터 전처리

- 잘못 탐지된 R 피크를 피크 보정을 통해 재 탐지 → 3th column 사용 (length : 357)

- 잘못 탐지된 경우

- ① R파와 T파가 동일할 경우, 동시 검출 문제
- ② R파보다 T파가 높을 경우, T파를 R파로 잘못 탐지하는 문제
- ③ 피크점이 거의 비슷해서 탐지가 어려운 문제



488,487.0	0.2886786832736596
489,488.0	0.2880288051654945
490,489.0	0.2877352586774743
491,490.0	0.2896210285389756
492,491.0	0.2932693547163599
493,492.0	0.2960740446251993
494,493.0	0.29555550144967
495,494.0	0.2917635674952222
496,495.0	0.2873636774205808
497,496.0	0.2852735323400199
498,497.0	0.2861095089429822
499,498.0	0.2877842295783067
500,499.0	0.2875418977113689
501,500.0	0.2845654707774232
502,501.0	0.2807791427241661
503,502.0	0.2790362467100814

Figure 1. ECG sample dataset

실험환경 구축 (1/2)

- 사용 모델 : 1D CNN (Convolutional Neural Network)
 - 1D CNN은 시계열 데이터 / 텍스트와 같은 1차원 데이터에 적합한 모델
- 시계열 데이터는 일반적으로 레이어가 깊어질 수록 성능 저하 문제가 발생하므로 4개의 실험환경 구축
 - ① 1개의 Convolution layer
 - ② 2개의 Convolution layer
 - ③ 3개의 Convolution layer
 - ④ 4개의 Convolution layer

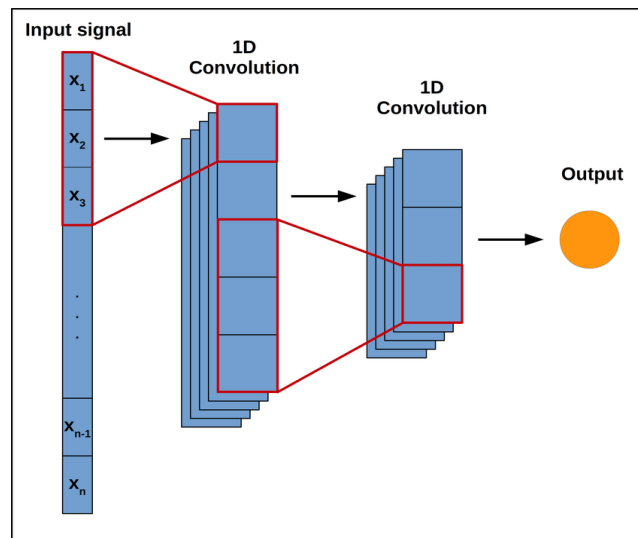
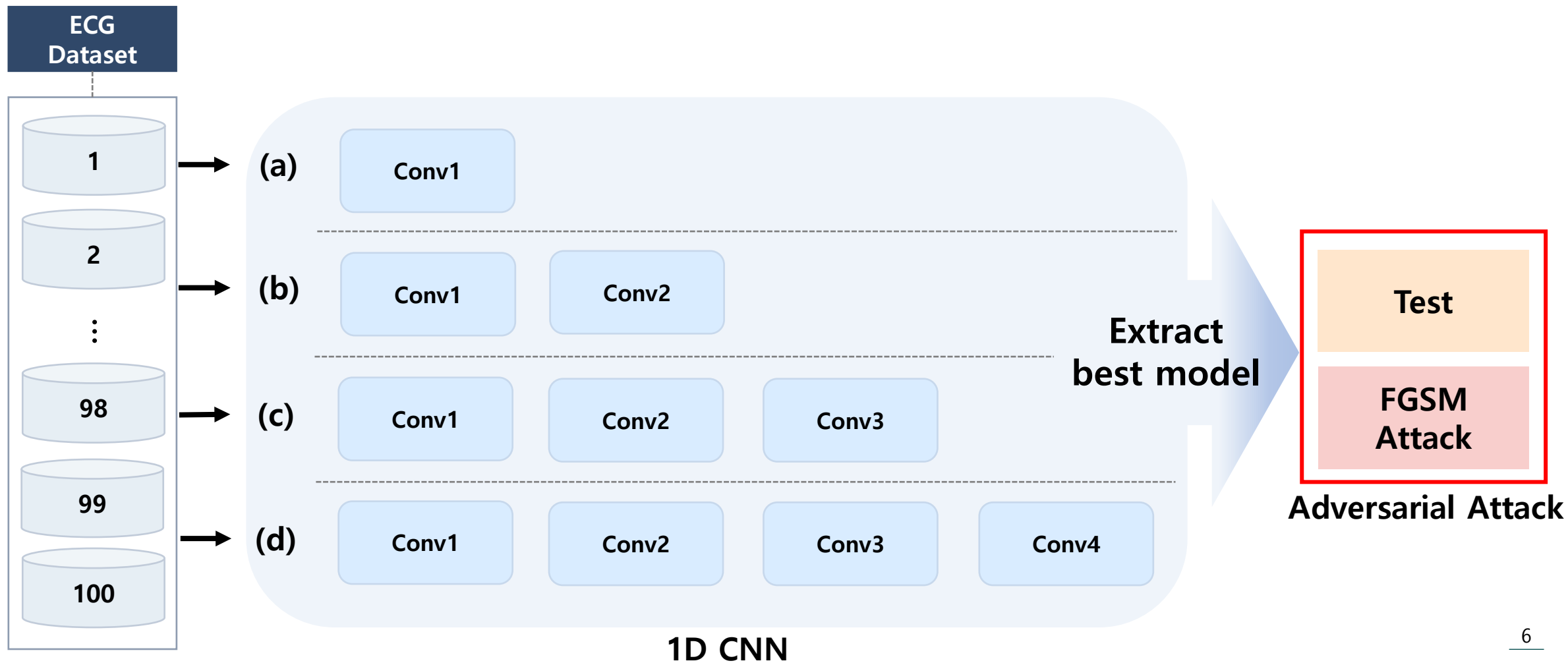


Figure 2. 1D CNN architecture

Architecture



실험환경 구축(2/2)

- Dataset → Training set : Validation set : Test set = 7 : 2 : 1
- 실험환경

CPU	Intel(R) Core(TM) i9-9900KF CPU
Memory	128GB RAM
GPU	NVIDIA TITAN RTX
OS	Windows 10

Table 2. Experiment setup

- 모델 훈련에 사용한 Hyperparameter

Objective function	Cross entropy loss
Optimizer function	Adam optimizer
Learning rate	0.0001
Batch size	64
Dropout	0.5
Epoch	100

Table 3. Hyperparameter

실험결과 (1/2) : Validation

- Conv1

Highest accuracy	96.92%
Average accuracy	81.97%
Accuracy Standard Deviation	25.32
Highest F1 score	97.62%
Average F1 score	81.32%
F1 score Standard Deviation	27.12

Table 4. Conv1 validation performance result

- Conv2

Highest accuracy	98.44%
Average accuracy	95.22%
Accuracy Standard Deviation	7.64
Highest F1 score	98.93%
Average F1 score	95.21%
F1 score Standard Deviation	8.23

Table 4. Conv2 validation performance result

실험결과 (2/2) : Validation + Test

• Conv3

Highest accuracy	99.25%
Average accuracy	96.30%
Accuracy Standard Deviation	7.39
Highest F1 score	99.34%
Average F1 score	96.72%
F1 score Standard Deviation	8.14

Table 4. Conv3 validation performance result

Accuracy	97.12%
F1 score	98.67%

Table 5. Test performance result

• Conv4

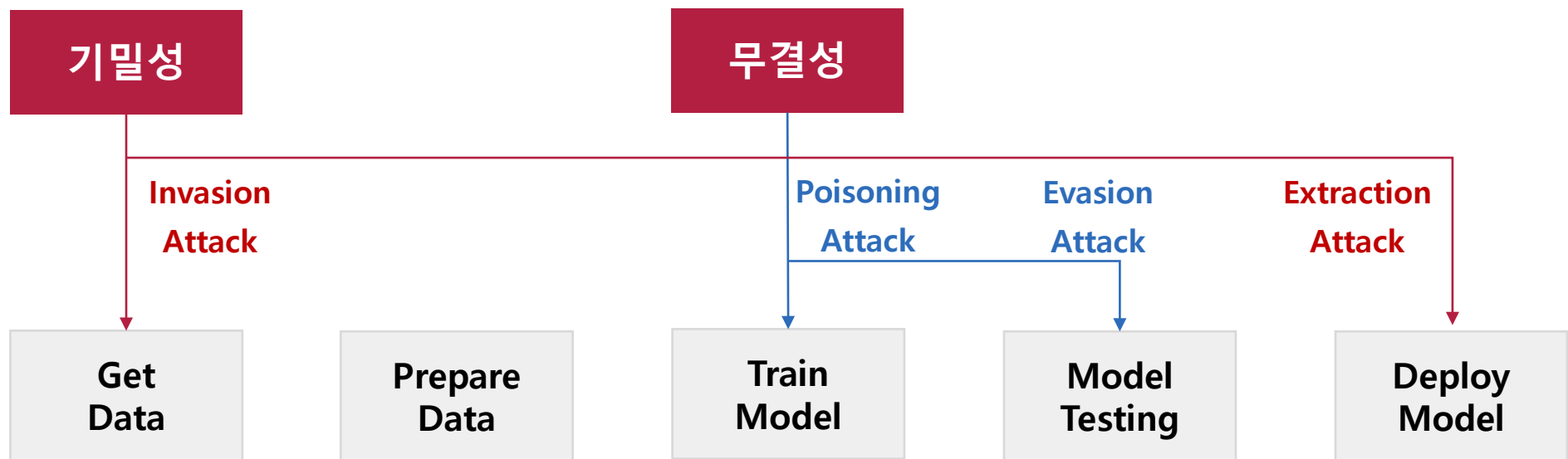
Highest accuracy	97.75%
Average accuracy	88.92%
Accuracy Standard Deviation	15.67
Highest F1 score	98.61%
Average F1 score	89.74%
F1 score Standard Deviation	16.23

Table 4. Conv4 validation performance result

Conv3에서 가장 높은 성능을 도출한 가중치 선택

적대적 공격

- 적대적 공격(Adversarial Attack) : 머신러닝 알고리즘에 내재하고 있는 취약점에 의해 적대적 환경에서 발생할 수 있는 보안 위험
- 적대적 공격의 3가지 유형 : **Inversion Attack**, **Poisoning Attack**, **Evasion Attack**, **Extraction Attack**
- 기밀성 : 인가된 사람/ 프로세스/ 시스템만이 정보에 접근할 수 있음
- 무결성 : 인가된 사람/ 프로세스/ 시스템만이 정보를 변경할 수 있음



Evasion Attack

- 입력 데이터에 최소한의 변조를 가해 머신러닝 모델을 속이는 기법
- 사람의 눈으로 식별하기 어려운 방식으로 이미지를 변조해 머신러닝 모델이 착오를 일으키게 만드는 방법
- 2018년 구글 리서치 그룹은 이미지 인식 머신러닝 알고리즘을 오작동 시킬 수 있는 'Adversarial Patch' 스티커를 발표
- Evasion attack이 실생활에 사용되면 보안 솔루션의 탐지 정책을 우회하거나, 교통 신호를 교란시켜 자율주행 차량의 오작동을 유발할 수 있음

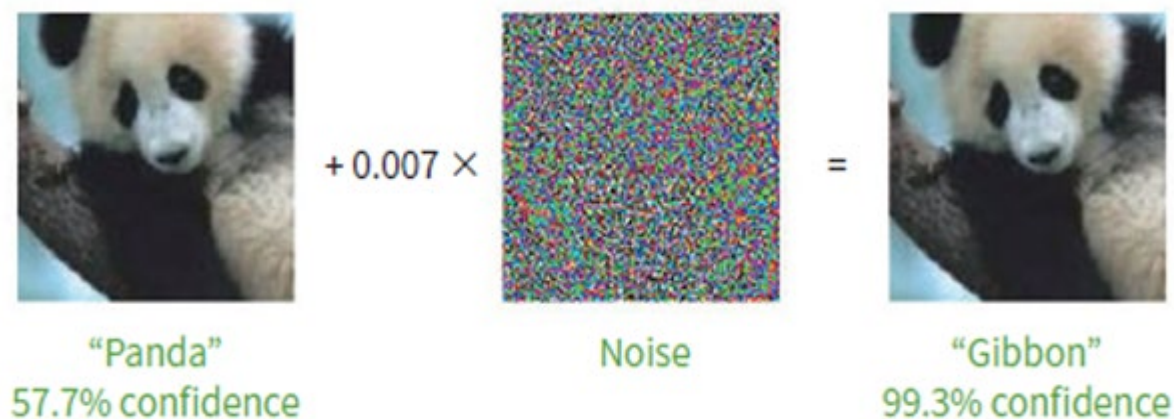


Figure 3. 팬더 이미지에 노이즈를 추가한 후 모델이 긴팔원숭이로 잘못 인식함

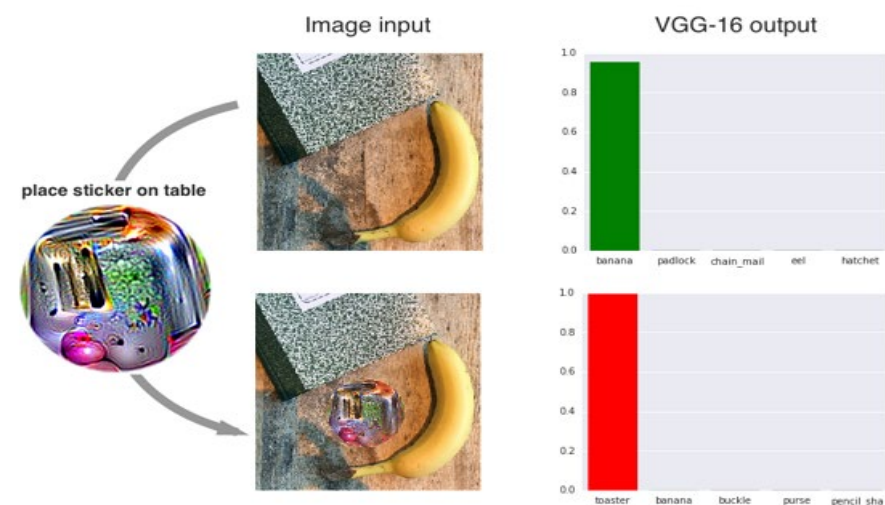


Figure 4. 토스터 그림과 적대적 스티커를 붙인 경우 결과 비교

FGSM

- FGSM (Fast Gradient Sign Attack) : 모델이 입력 데이터에 대한 gradient를 활용하여 신경망을 공격
- FGSM을 ECG 신호에 적용하는 방법
 - ① 공격 대상이 되는 ECG 신호를 선택
 - ② 선택한 데이터를 대상으로 forward pass를 수행하여 손실 값 계산
 - ③ 손실 값을 기반으로 backward pass를 수행하여 입력 데이터에 대한 gradient 얻음
 - ④ Gradient의 부호가 + 이면, 해당 신호의 각 시간 단계의 신호 값을 증가
- 이면, 해당 신호의 각 시간 단계의 신호 값을 감소
 - ⑤ 수정된 신호를 모델에 다시 입력하여 forward pass를 수행하여 오분류를 유도

실험결과 : FGSM

- Test performance result

Accuracy	97.12%
F1 score	98.67%

Table 5. Test performance result

- FGSM Attack performance result (epsilon : 0.05)

Accuracy	55.84%
F1 score	53.17%

Table 5. FGSM attack performance result



40% 이상의 성능 하락을 보임

Conclusion

- ECG 데이터는 3개의 conv에서 가장 높은 성능을 보임
 - FGSM 공격 시, 40% 이상의 공격 성공률을 보임
 - Center node와 local node 통신이 중요한 연합학습에 적대적 공격이 취약하다는 것이 입증됨
- 후속 연구로 해당 실험을 연합학습에 적용하여, 연합학습에서의 적대적 공격의 취약성을 검증하고자 함

Thank you

SaeBom Lee 이새봄
Department of Computer Engineering, Gachon University | Researcher

Tel. +82-31-750-8822 Mobile. +82-10-6641-9390
E-mail. dltoqha@gachon.ac.kr