

[인공지능 융합보안 팀 프로젝트 6조]

Poisoning Attack 수준 및 Calibration 종류에 따른 의료 이미지 데이터 분류 연구

A Study of Medical Image Data Classification on Poisoning
Attack Level and Calibration Types

미래융합기술공학과 220214014 이유림

미래융합기술공학과 220216027 김소연

학과 학번 이슬아

학과 학번 신희경 (발표자)

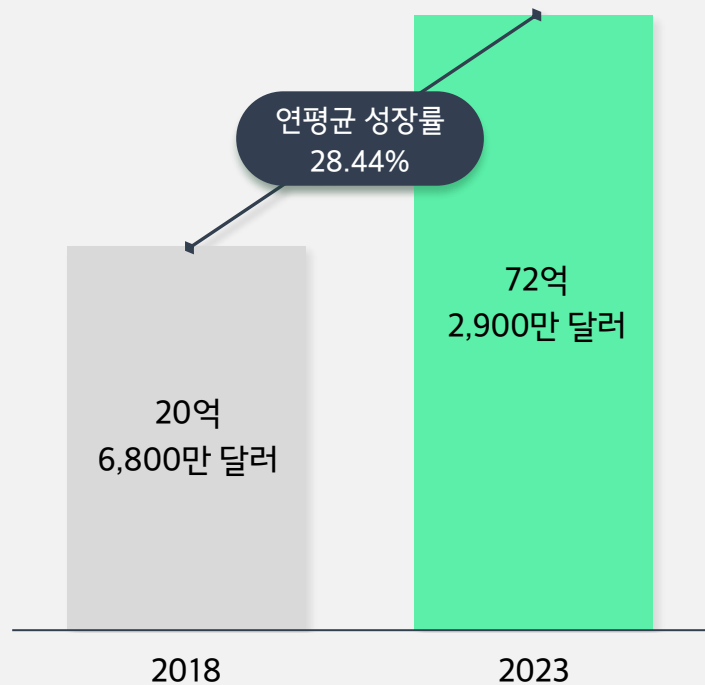
Contents

1. Introduction
2. Related Work
3. Experiment
 - 1) Dataset
 - 2) Setup
 - 3) Experimental Result
4. Conclusion

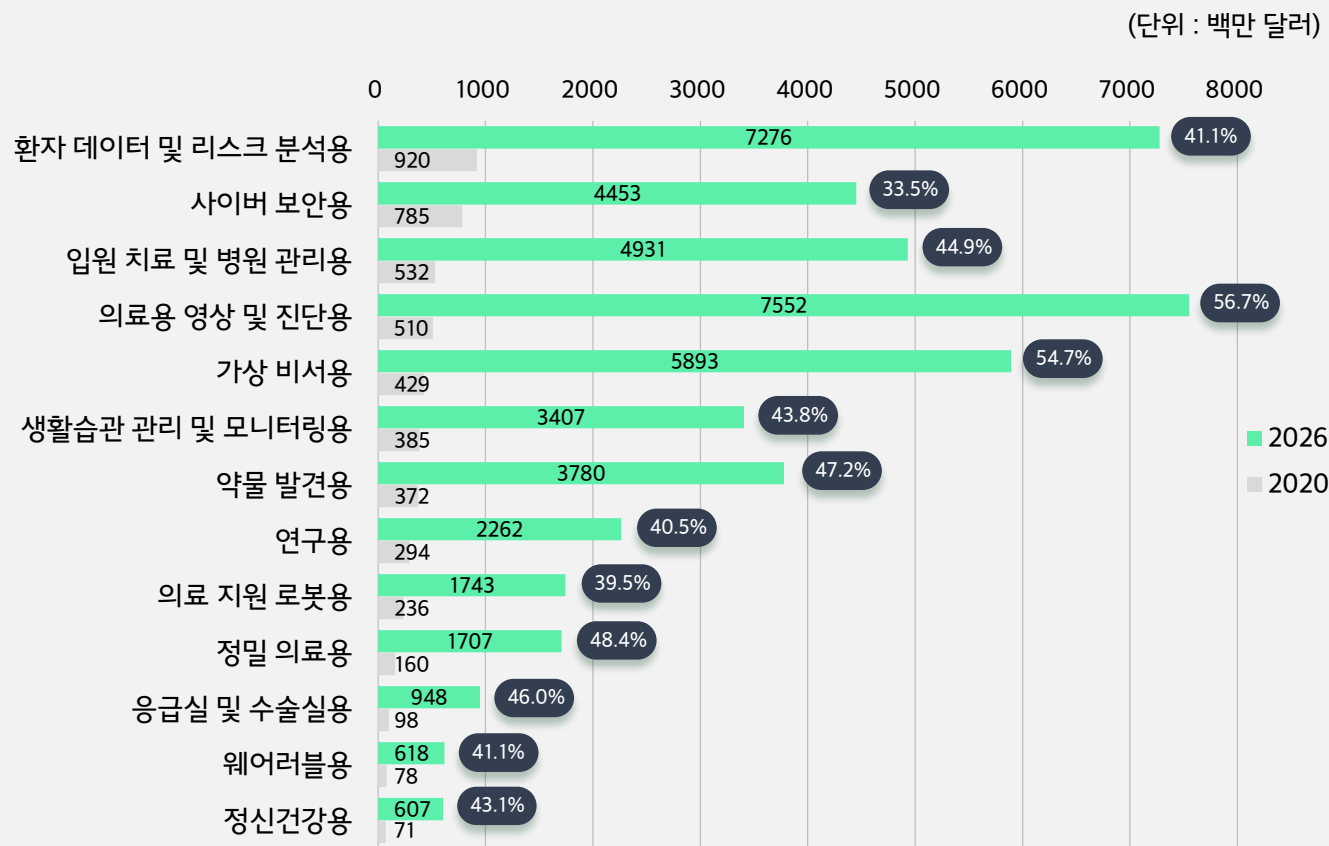
1. Introduction

❖ 연구 배경

○ 전 세계 의료 부문용 인공지능 시장 규모



[그림1] 글로벌 의료 부문용 인공지능 시장 규모 및 전망

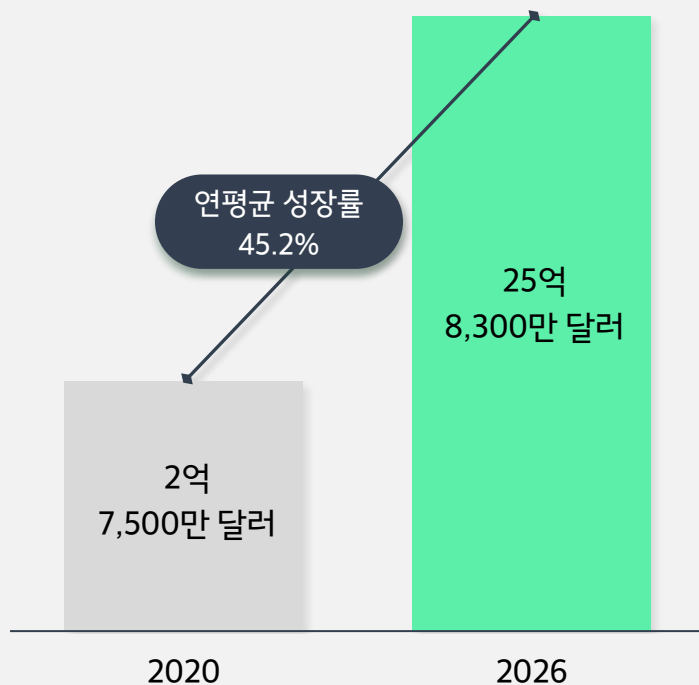


[그림2] 글로벌 의료용 인공지능 시장의 용도별 시장 규모 및 전망

1. Introduction

❖ 연구 배경

○ 국내 의료용 인공지능 시장 규모



[그림3] 국내 의료 분야 인공지능 시장 규모 및 전망

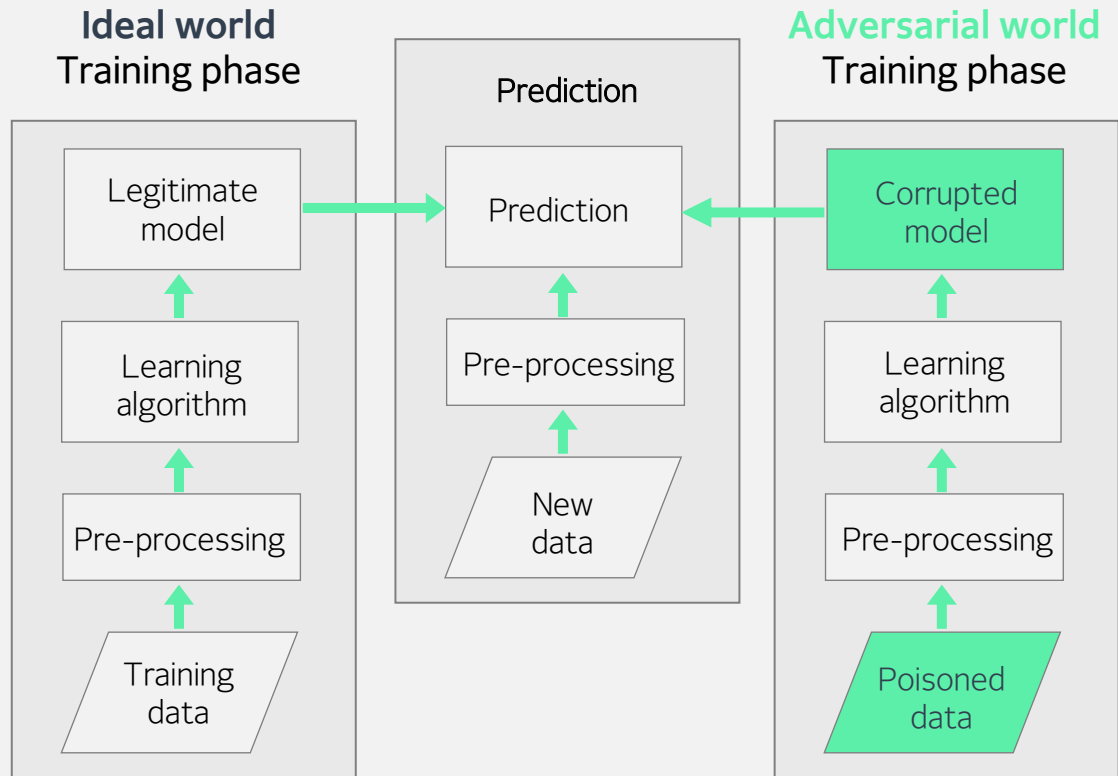
- 크고 복잡한 의료 데이터 세트 유입
- 의료 비용 절감에 대한 요구 증가
- 컴퓨팅 성능 향상 및 하드웨어 비용 감소
- 코로나(COVID-19) 대응을 위해 신약 개발, 영상 및 진단 분야에서 AI 기술의 잠재력 증대

➤ 다양한 요인으로 의료용 인공지능 시장이 향후 몇 년 동안 크게 발전할 것으로 예상

1. Introduction

❖ 연구 배경

○ 학습 단계에서의 보안상 취약점



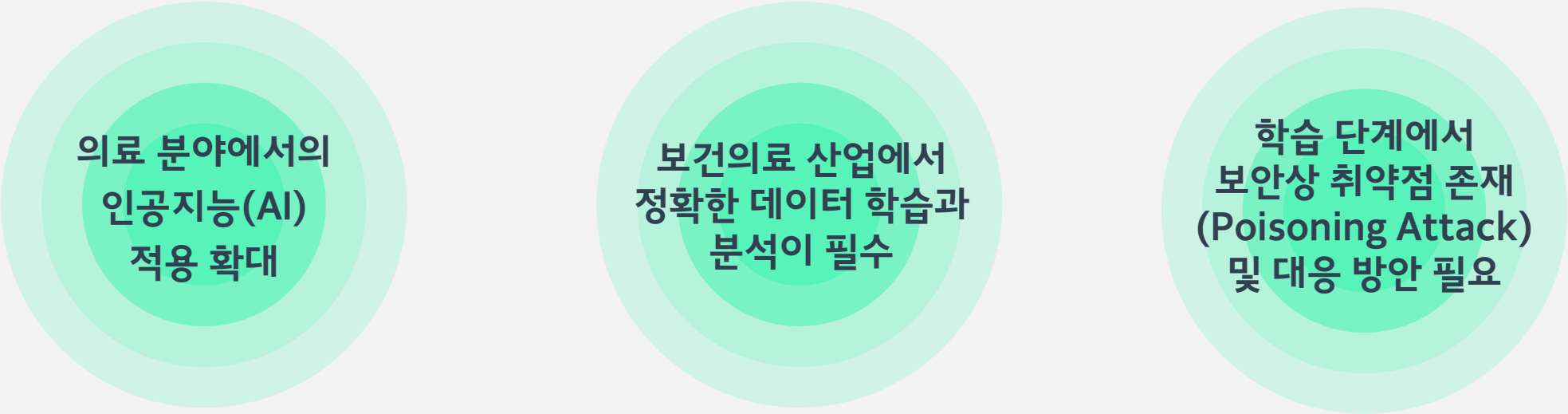
[그림4] System Architecture

○ Poisoning Attack

- 공격자가 의도적으로 학습 데이터에 영향을 주어 예측 모델 결과를 조작
- 다른 공격 기법과 달리 모델 자체를 공격하여 모델에 영향을 끼침
- 보건의료 산업에서 오진을 할 경우, 파급력이 크기 때문에 질병의 여부를 판단하거나 예측하는 모델은 정확한 데이터 학습과 분석이 필수적

1. Introduction

❖ 연구 필요성



의료 분야에서의
인공지능(AI)
적용 확대

보건의료 산업에서
정확한 데이터 학습과
분석이 필수

학습 단계에서
보안상 취약점 존재
(Poisoning Attack)
및 대응 방안 필요

- ✓ 딥러닝 모델 별 Poisoning Attack 수준에 따른 분류 성능 확인
- ✓ Label Smoothing과 Focal Loss 을 적용하여 성능 개선

3. Experiment

1) Dataset

○ 페럼 X-Ray 이미지 데이터 세트

- 원본 데이터 세트 구성
 - 총 샘플 수 : 5,863개
 - Train, Test, Val 3개 폴더로 구성
 - 각 폴더는 Normal(양성), Pneumonia(악성) 데이터로 구성

○ 본 논문에서 사용한 데이터 세트

- 데이터 세트 구성
 - 총 샘플 수 : 3,306개
 - Train : 양성 이미지 1341장, 악성 이미지 1341장
 - Test : 양성 이미지 312장, 악성 이미지 312장

- ✓ 딥러닝 모델 별 Poisoning Attack 분류 수준을 측정하고, Calibration을 적용하여 성능 개선을 파악하고자 원본 데이터 세트를 변경
- ✓ Poisoning Attack 분류 수준 측정을 위해, 전체 학습 이미지 중 무작위로 이미지를 선정하여 레이블 값을 악성-> 양성, 양성-> 악성으로 변경
- ✓ 변경한 레이블 값의 비율은 10%, 20%, 30%, 40%, 50%

3. Experiment

2) Set-up

calibration 종류에 따른 학습 개선도를 측정하기 위해 **Label smoothing**과 **focal loss**를 적용

이때, label smoothing 파라미터 값을 0.01, 0.1, 0.2로 높여가며 정확도와 손실률의 개선된 정도를 확인

=> label smoothing 파라미터 조정에 따른 학습 개선도 + label smoothing과 focal loss 두 기법의 성능 비교

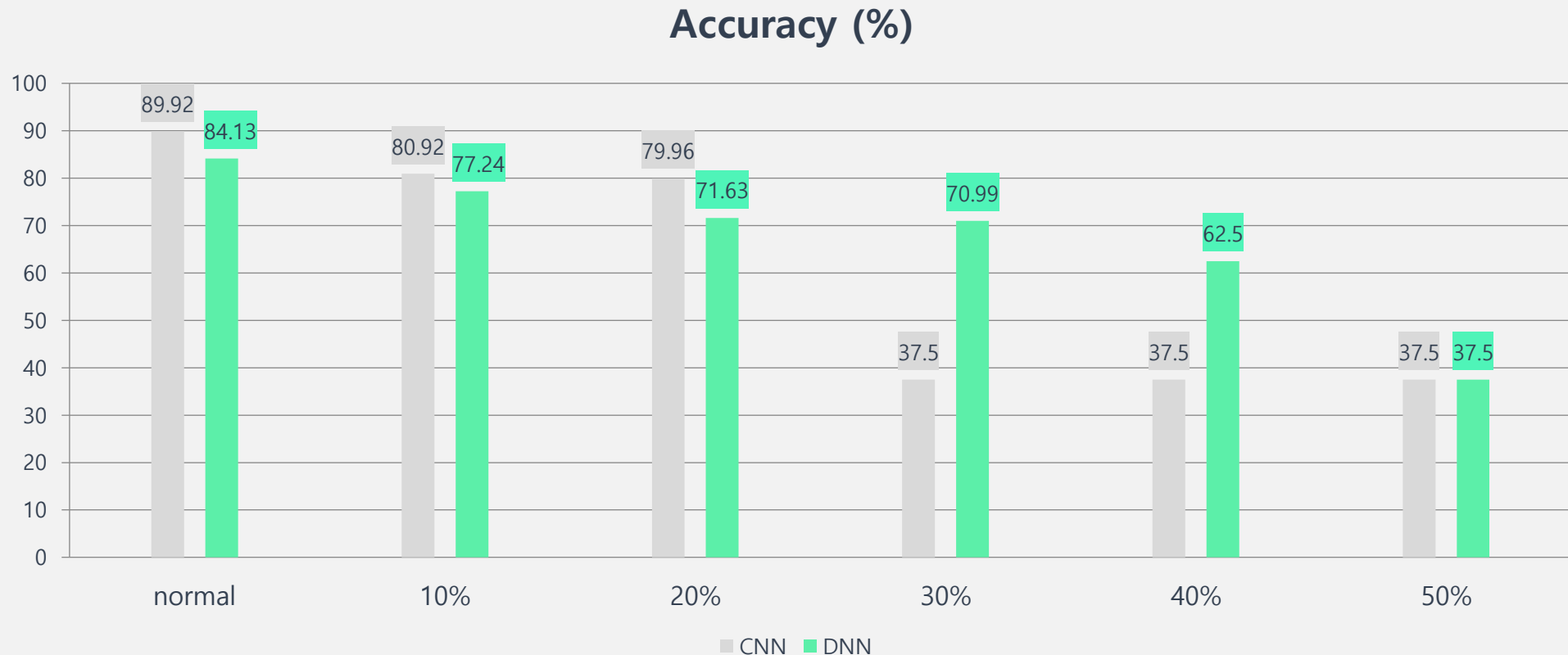
CNN(Convolution Neural Network), DNN(Deep Neural Network) 2개의 딥 러닝 모델 구현

	Framework	Batch size	Epoch	Loss function	Optimizer	Learning rate	Steps per epoch	Validation step
CNN	keras	4	30	Binary Cross entropy	RMSprop	1e-4	10	50
DNN	keras	16	30	Binary Cross entropy	Adam	0.1	x	x

3. Experiment

❖ Experimental result

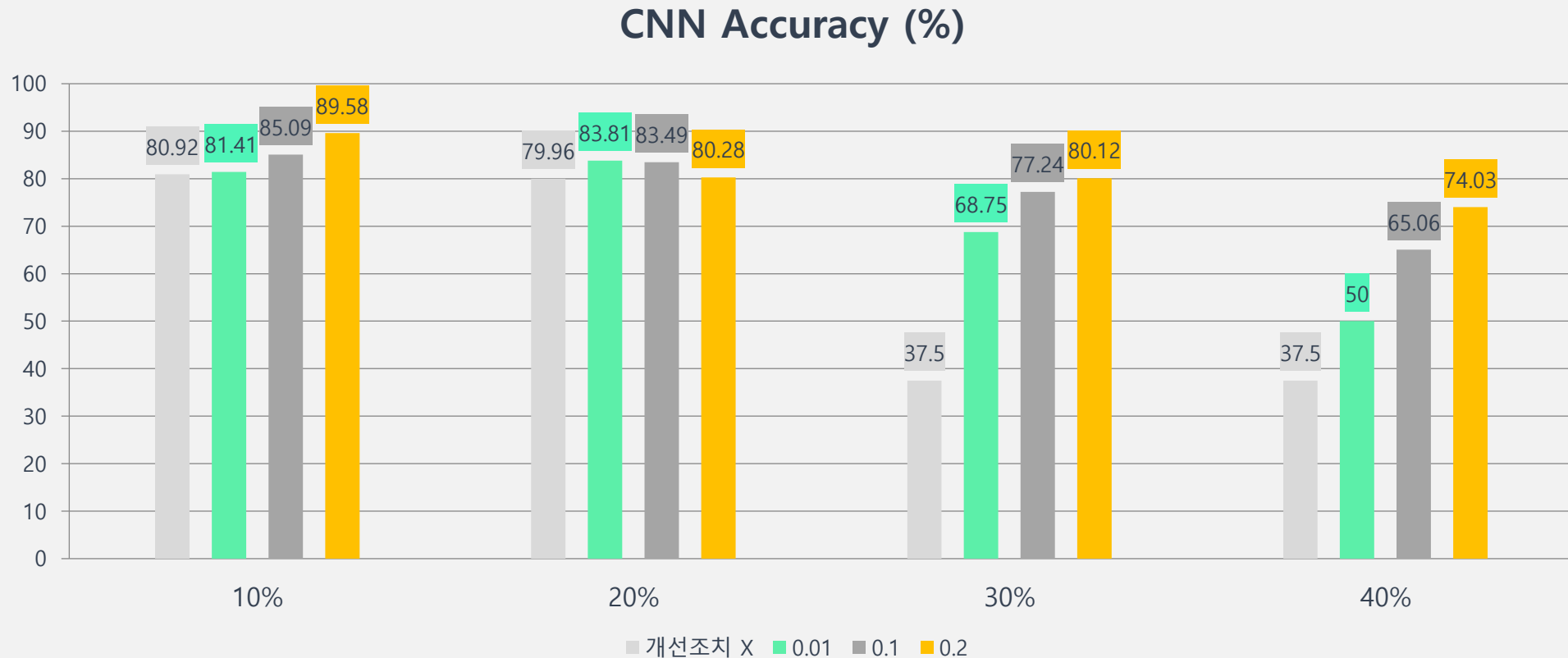
- Poisoning 정도에 따른 모델별 정확도
 - normal: poisoning을 하지 않았을 때의 모델 정확도
 - poisoning을 수행하였을 때 탐지 정확도가 안좋아짐
 - poisoning을 많이 할수록 탐지 정확도가 안좋아짐
 - poisonin을 적게할 경우 CNN이 공격에 대한 대응력이 더 좋고, poisoning을 많이 할 경우엔 DNN이 대응력이 더 좋음



3. Experiment

❖ Experimental result

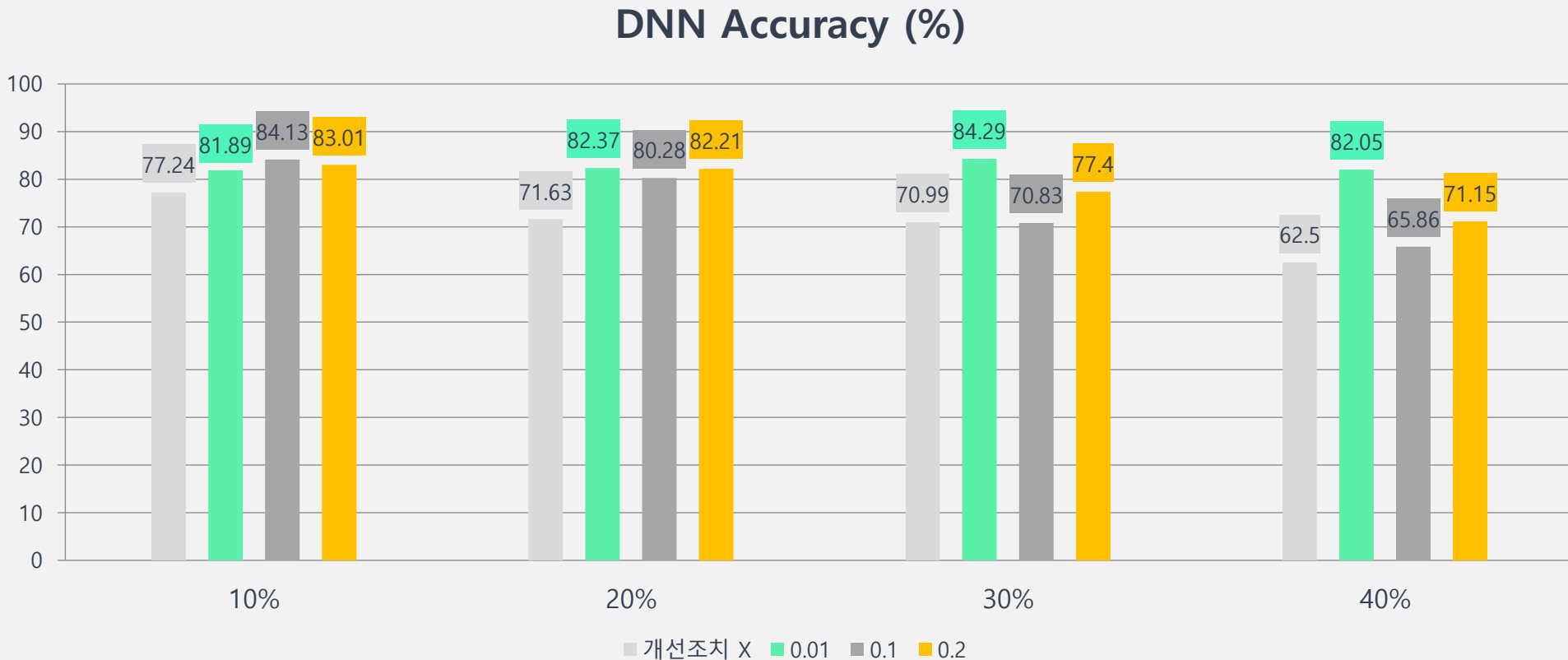
- label smoothing 정도에 따른 poisoning 모델별 정확도 개선 효과
 - calibration을 적용하였을 때 탐지 정확도가 향상됨
 - 각 poisoning level에서 라벨 스무딩 정도에 따른 탐지 정확도의 경향성이 있음
 - 라벨 스무딩을 적용했음에도 불구하고 poisoning level이 낮을수록 각 poisoning level의 최고 정확도가 높음 (89.58% > 83.71% > 80.12% > 74.03%)



3. Experiment

❖ Experimental result

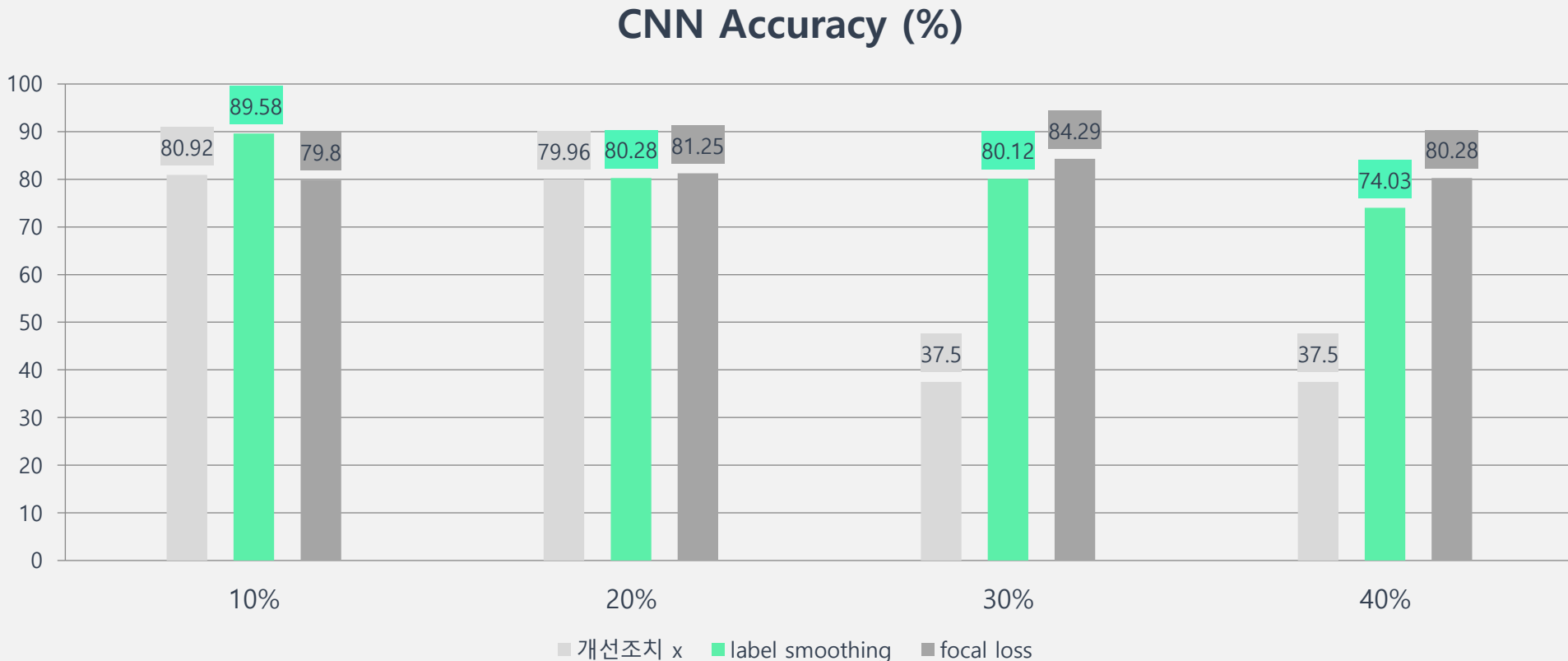
- label smoothing 정도에 따른 poisoning 모델별 정확도 개선 효과
 - 대부분 calibration을 적용하였을 때 탐지 정확도가 향상됨
 - 각 poisoning level 별 최고 탐지 정확도는 84.13%, 82.37%, 84.29%, 82.05%로, calibration을 적용하지 않았을 때 대비 평균 약 12.62%만큼 탐지 정확도 향상됨



3. Experiment

❖ Experimental result

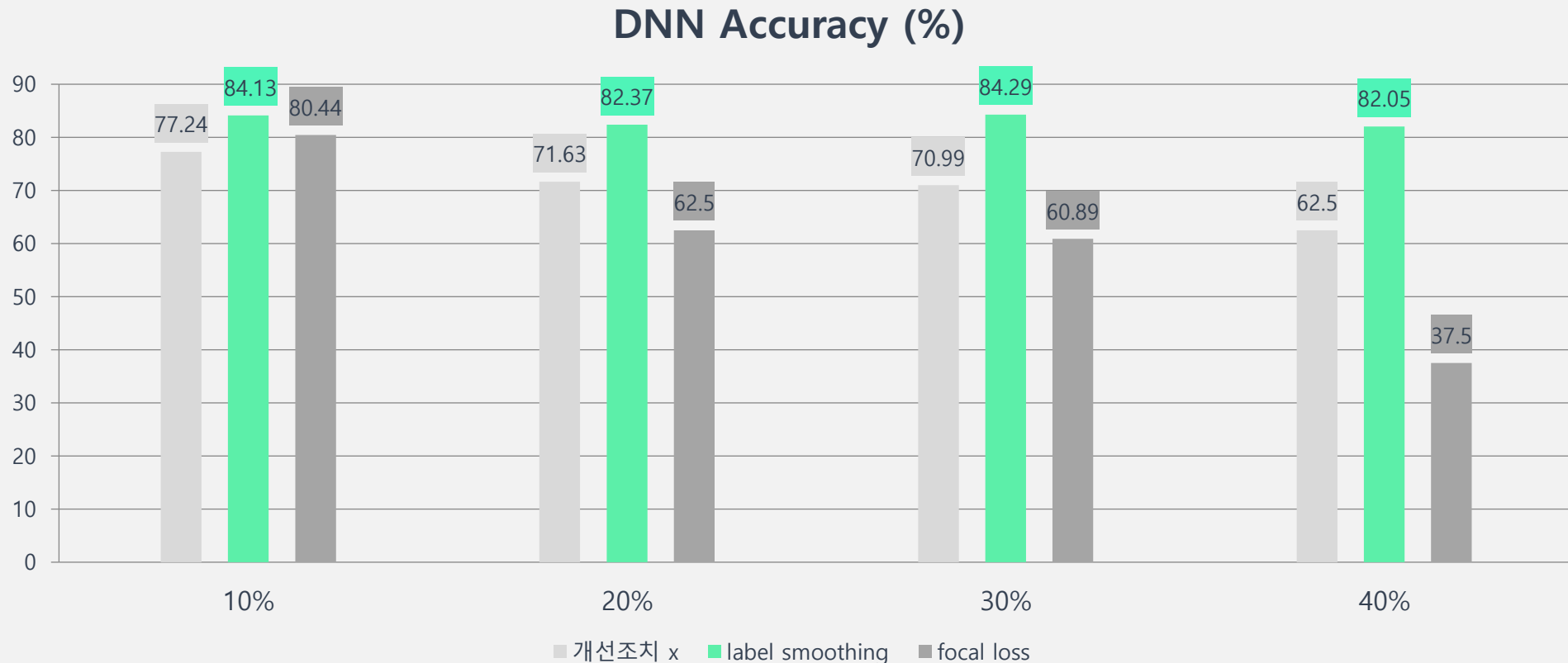
- calibration 종류 별 poisoning 모델별 정확도 개선 효과 비교
 - label smoothing의 경우, 탐지율이 가장 높을 때의 label smoothing 값을 사용하였음
 - 10%의 경우, label smoothing을 사용하는 것이 정확도 개선에 도움이 됨
 - 20~40%의 경우, focal loss를 사용하는 것이 정확도 개선에 도움이 됨
 - calibration을 적용하였을 때, 그렇지 않았을 때 대비 24.88%의 탐지율 개선 효과 O



3. Experiment

❖ Experimental result

- calibration 종류 별 poisoning 모델별 정확도 개선 효과 비교
 - label smoothing의 경우, 탐지율이 가장 높을 때의 label smoothing 값을 사용하였음
 - 모든 poisoning level에서 label smoothing을 사용하였을 때가 focal loss를 사용하였을 때보다 정확도 개선 효과가 좋음
 - 20~40%에서, focal loss를 사용하는 것보다 기본 이진분류에서 탐지율이 더 높음



4. Conclusion

중요 데이터에 대한 poisoning attack 발생 시 대응 가능성 확인

이미지 분류 성능이 좋은 딥러닝 모델 CNN, DNN 중 poisoning attack에 대응을 더 잘하는 모델 확인

calibration을 통해 poisoning attack에 대한 대응이 가능함을 증명

대표적인 calibration 방안인 label smoothing과 focal loss중 의료 데이터 poisoning attack에 대한 대응력 비교분석