CUAI 8기 논문 발표

의료 인공지능과 XAI

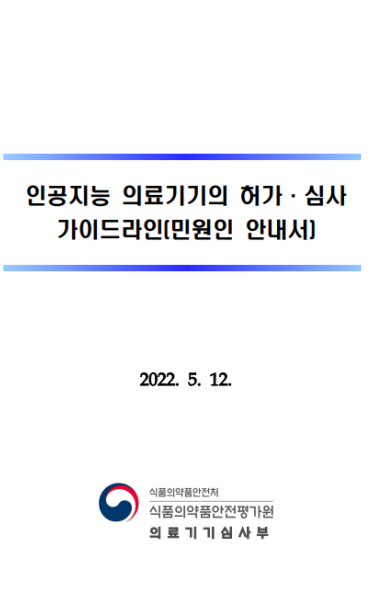
2025.03.11

발표자: 권하연

목차

1. 주제 선정 배경
   : 인공지능 의료기기 허가·심사 가이드라인
2. 논문리뷰
   : **설명 가능한 인공지능 기반 의료 AI 기술 연구 동향**

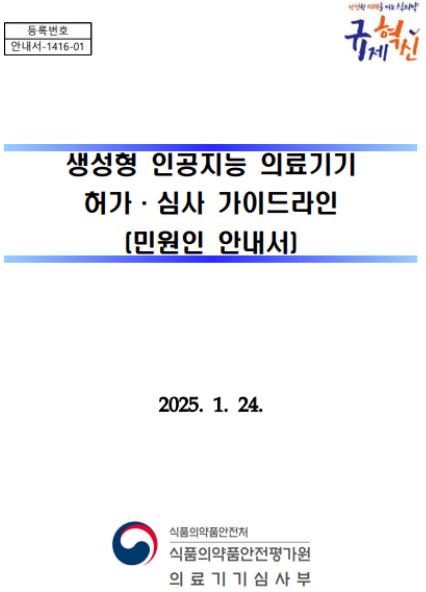
1. 주제 선정 배경


등록번호
안내서-1416-01

인공지능 의료기기의 허가 · 심사
가이드라인(민원인 안내서)

2022. 5. 12.

식품의약품안전처
식품의약품안전평가원
의 료 기 기 심 사 부


생성형 인공지능 의료기기
허가 · 심사 가이드라인
(민원인 안내서)

2025. 1. 24.

식품의약품안전처
식품의약품안전평가원
의 료 기 기 심 사 부


홈 > 의료기기·IT

'생성형 인공지능 의료기기' 허가 지침 세계 첫 제정

이정윤 기자 | 입력 2025.01.24 09:17 | 댓글 0

식약처, '독립형 디지털의료기기SW 사용적합성 가이드라인' 발간도

[의학신문·일간보사=이정윤 기자] 식품의약품안전처(처장 오유경)는 생성형 인공지능을 활용한 의료기기의 안전성·유효성 평가에 도움을 주고 제품화를 지원하기 위해 세계최초로 '생성형 인공지능 의료기기 허가·심사 가이드라인'을 제정·발간한다고 밝혔다.

생성형 인공지능 의료기기 개요

이번 '생성형 인공지능 의료기기 허가·심사 가이드라인'에서는 생성형 인공지능 의료기기에 해당하는 사례를 제시하고 허가신청서 작성 방법 및 제출자료에 대해 안내했다.

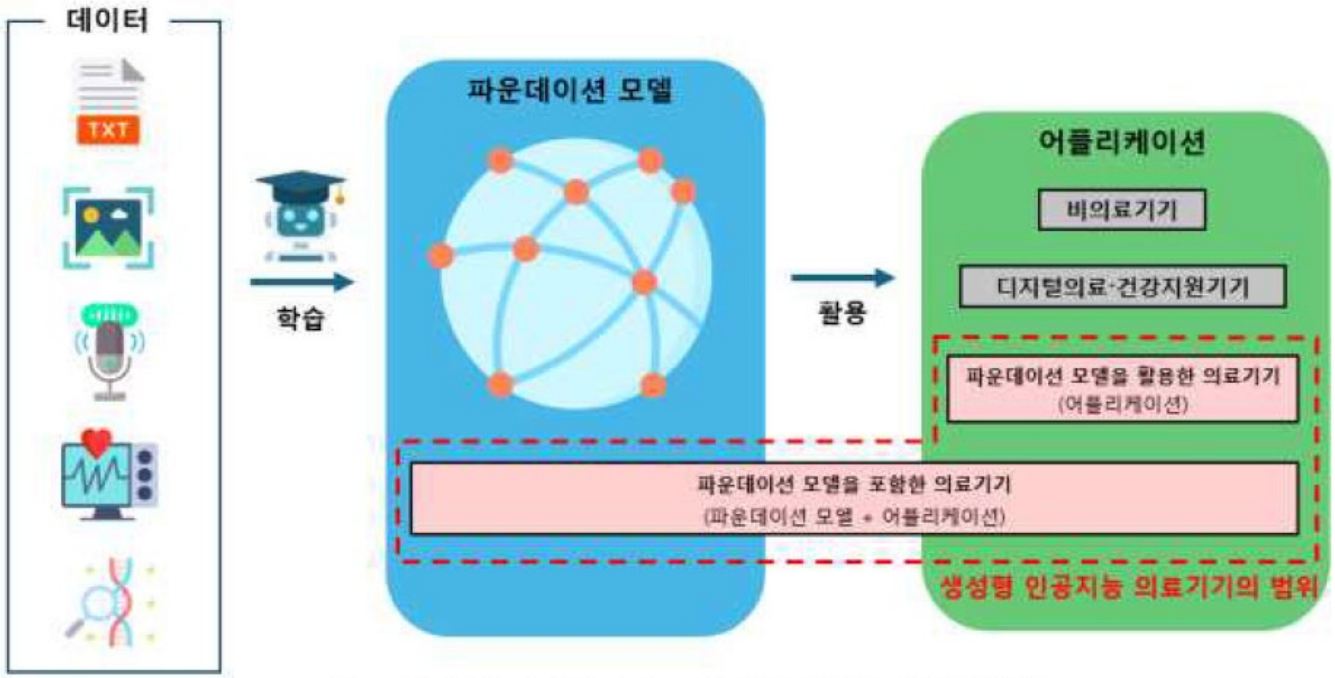
의료영상 판독, 진단 보조, 치

1. 주제 선정 배경

목 차

목 차

1. 주제 선정 배경



[그림. 생성형 인공지능 의료기기의 관리범위]

1. 주제 선정 배경

< 의료기기 예시 >

- 폐 CT 영상을 분석하여 폐암의 유무 또는 폐암의 진행상태(병기)를 자동으로 진단하는 소프트웨어
- 심전도 측정 결과를 이용하여 부정맥을 진단하거나 예측하는 소프트웨어
- 조직검사, 전자의무기록(EMR) 등 의료정보를 기반으로 특정 암의 발병확률을 계산하는 소프트웨어
- 피부병변 영상을 분석하여 피부암 유무를 진단하는 소프트웨어
- 혈당 데이터, 음식 섭취, 인슐린 주입 등 정보를 분석하여 저혈당증을 예측하는 소프트웨어
- 응급실에서 측정·통합한 생체신호를 분석하여 호흡곤란 등 응급상황을 예측하거나 알람 등 경고를 하는 소프트웨어
- 위 CT 영상 분석을 통해 이상 부위를 검출하여 표시해주는 스크리닝 소프트웨어
- 의료영상을 분석하여 혈류속도, 혈관직경 등 혈관 특정 부위의 정량적 수치를 제공하는 소프트웨어
- 의료데이터를 기반으로 방사선 치료계획을 수립하는 소프트웨어

1) 생성형 인공지능 의료기기에 해당하는 의료제품

< 예 시 >

- 흉부 엑스레이 영상과 대응되는 판독문을 생성형 인공지능으로 학습하여 영상을 분석하고, 검출할 수 있는 폐질환 관련 병변에 대한 판독문 초안을 작성해주는 소프트웨어
- 환자의 전자의무기록(EMR) 데이터를 대량으로 분석하여 맞춤형 치료계획을 생성해 주는 소프트웨어
- 환자의 음성녹음 파일을 분석하여 파킨슨병의 징후를 조기에 감지하고, 그 결과로 진료요약서를 작성해주는 소프트웨어
- 환자의 시퀀싱 데이터를 이용하여 특정 유전질환(유전자 변이 등)에 대한 발병 가능성을 예측하고, 이를 바탕으로 발생 위험 보고서를 생성하는 소프트웨어

2) 생성형 인공지능 의료기기에 해당하지 않는 의료제품

< 예 시 >

- 환자의 전자의무기록(EMR)데이터를 단순히 기록, 검색, 조회 및 요약 등을 하는데 사용하는 소프트웨어
- 음성으로 기록된 의료인의 진료내용을 텍스트로 변환하여 요약문서를 작성해주는 소프트웨어

1. 주제 선정 배경

일반적으로 기존 기계학습 가능 의료기기는 정교하게 설계되고 주의 깊게 모니터링되는 임상시험을 수행하여 1~2가지의 적응증에 대한 안전성과 유효성을 평가하였다. 또한, 추가하고자 하는 적응증이 있다면 임상시험을 통해 이를 입증하여야 했다.

이에 비해 생성형 인공지능 의료기기(GenAI Medical Devices)는 새로운 데이터를 지속적으로 학습함에 따라 생성형 인공지능 모델의 계산 능력이 끊임없이 변화하여 특정 적응증뿐만 아니라 잠재된 적응증까지도 평가하여야 하므로, 기존의 평가방법과 기준을 적용한다면 안전성과 유효성 평가에 막대한 비용과 시간을 필요로 하게 된다[5].

아울러, 생성형 인공지능 의료기기의 임상현장 도입에는 다음과 같은 문제점을 함께 고려하여야 한다.

첫째, 오류나 편향이 있는 데이터가 걸러지지 않고 학습될 가능성이 있어 부정확한 출력결과를 초래할 수 있다.

둘째, 생성형 인공지능 모델은 기존 인공지능 모델보다 설명가능성의 한계가 있으며, 결과의 신뢰성 확보가 중요하다.

셋째, 동일한 입력에도 일관성 없는 출력을 할 수 있는 가능성이 있어, 재현성과 신뢰성 검증이 추가적으로 필요하다.

[표 1. 생성형 인공지능 의료기기의 특성]

구 분	내 용
학습 (Learning)	사용목적 내에서 환자 건강에 영향을 줄 수 있는 구체적인 결과를 제공하기 위해 데이터를 축적(학습)할 수 있음
자율성 (Autonomy)	학습에 따라 임상의의 감독을 줄이거나 심지어 감독 없이도 프로세스 또는 결과를 수정할 수 있는 잠재력을 가지고 있음
설명불가능성 (Inexplicability)	정교한 계산 능력, 복잡한 통계, 크고 복잡한 데이터셋 등을 학습하므로, 출력값에 대한 근거(rationale)는 전문지식이 없는 개인뿐만 아니라, 잘 훈련된 임상의와 기타 의료진도 쉽게 이해하지 못할 수 있음

1. 주제 선정 배경

[표 2. 생성형 인공지능 의료기기 위해요인(Hazard) 예시]

구 분	위 해 요 인
성능 (Performance)	1. **설득력있는 환각(Hallucination)**: 인공지능 모델이 부정확하거나 편향되거나 의도하지 않은 출력을 생성하지만, 콘텐츠의 문법과 구조 등이 매우 단호하고 설득력이 있어 정확한 출력으로 오인될 수 있음
	2. **일관성이 없음(Inconsistency)**: 인공지능 모델 특성으로 인해 반복된 세션에서 동일 입력에 대한 출력이 일관되지 않을 수 있음(예. 동일한 흉부 X-ray 영상을 입력하나 매번 다른 판독문이 생성되어 출력됨 등)
	3. **연관성이 없음(Irrelevancy)**: 모호한 질문 간 차이 혹은 부적절한 질문의 맥락을 인지하지 못하고 연관성이 없거나 부정확한 답변 생성(예. 허혈성 뇌졸중과 출혈성 뇌졸중을 구분하지 못하고 잘못된 진단 결과를 출력; 흉부 X-ray 영상에 대한 판독문을 생성해야 하나 다른 부위의 X-ray 영상이 입력되어도 인지하지 못하고 결과를 출력 등)
	4. **불확실성 척도의 부재(No uncertainty indicator)**: 인공지능 모델 출력의 불확실성에 대한 정량적 척도를 제시함으로써 사용자가 해당 출력을 신뢰할 것인지에 대한 판단기준을 제시하지 못함 - 예1. 설명가능성(Explainability)[1] 측면: 사람이 이해할 수 있는 모델의 출력 결과에 대한 논리적 근거 미제시
	- 예2. 해석가능성(Interpretability)[2] 측면: 의료목적의 출력을 생성하는 원리를 뒷받침하는 대표적 참고문헌 또는 출처 미제시
데이터 품질 (Data Quality)	1. **데이터 오류(Incorrect data)**: 데이터값이 잘못되어(예. 입력 오류, 라벨링 오류 등) 있거나 적용할 수 없음(예. 데이터가 더 이상 환자의 상태를 대표하지 못함 등)
	2. **이상치 처리 오류(Incorrect handling of outliers)**: 인공지능 모델 학습 시 처리하면 안되는 이상치를 포함하거나 혹은 처리하여야 하는 이상치를 누락
	3. **불완전한 데이터(Incomplete data)**: 누락된 데이터(예. 연속형 데이터의 빈 공간 등)
	4. **주관적인 데이터(Subjective data)**: 객관적이며 정량적인 사실에 근거하지 않고 개인의 경험적 혹은 전문성에 영향을 받는 데이터(예. Lung-RADs를 활용한 폐 결절 분류에 있어 임상의 간 의견 차이 등)
	5. **일관되지 않은 데이터(Inconsistent data)**: 데이터는 출처 및 수집 시점 등에 따라 영향을 받을 수 있음(예. 감기 환자 데이터를 각각 여름과 겨울에 수집 등)
	6. **학습과 적용 데이터의 차이(Domain shifted data)**: 인공지능 모델 학습 시 사용한 데이터 품질이 실제 사용 시의 데이터 품질을 대표하지 못할 수 있음(예. 고해상도의 CT영상만을 학습한 인공지능 모델은 저해상도만 제공하는 병원에서 사용하기에 적합하지 않을 수 있음 등)
	7. **데이터 드리프트(Data drift)**: 환자집단과 의료관행 등은 시간이 지남에 따라 변화하여 더 이상 현재의 데이터를 대표할 수 없음 (예. 몇 년 전 폐렴 유병률 기반으로 학습한 인공지능 모델은 현시점의 폐렴 특성을 정확히 반영할 수 없음 등)
	8. **파편화된 데이터(Fragmented data)**: 인공지능 모델의 학습에 필요한 데이터가 하나의 형식 혹은 시스템으로 저장되지 않음 (예. 의료영상은 DICOM 형태로 저장되는 반면, 의료기록은 텍스트 형태로 저장됨 등)

1. 주제 선정 배경

편향 (Bias)	1. **선택편향(Selection bias)**: 누락된 데이터로 인해 발생 가능하며 ① 데이터가 무작위로 수집되지 않음 ② 데이터가 의도하는 환자집단의 특성과 일치하지 않음 ③ 개인정보 등의 문제로 수집된 데이터를 제외할 때 인구통계학적 분포가 고르지 않음 등의 원인으로 발생
	2. **중첩변수(Confounding variables)**: 실제로 입력과 출력 간 상관관계가 없지만 제3의 중첩변수로 인해 상관관계가 있는 것처럼 보일 수 있음(예. '단 음식 섭취'와 '체중 증가'는 직접적인 원인-결과처럼 보일 수 있으나 실질적으로는 두 가지 모두 '높은 혈당'으로 인한 결과일 수 있음 등)
	3. **비정규성(Non-normality)**: 모든 집단이 정규 분포를 갖지 않음에도 오히려 목적으로 하는 대상 집단이 정규 분포를 갖는다고 가정함으로써 발생 가능
	4. **대리 변수(Proxy variables)**: 원하는 데이터 요소를 수집할 수 없는 경우 대체 방식을 사용함으로써 발생 가능(예. '고위험 치료 관리'가 필요한 환자를 식별하는 알고리즘에 대해 '의료적 수요' 대신 '비용'을 데이터 요소로 수집할 경우, 의료비 지출이 상대적으로 낮은 가난한 사람들의 '의료적 수요'를 반영할 수 없음 등)
	5. **암시적 편향(Implicit bias)**: 개인의 정신 모델(mental model)에 기반하여 인지하지 못한 가정(예. 의사가 모든 환자를 동등하게 대하고자 하지만, 의사의 생각과 행동에 영향을 미칠 수 있는 특정 가설(가치관 등)을 갖고 있을 수 있음(예. 흑인 환자를 열등하게 생각해 약을 처방하지 않음))
	6. **집단 편향(Group attribution bias)**: 일반화의 오류(예. 한 집단에 대해서만 학습한 인공지능 모델이 똑같은 특성을 갖는 다른 집단에서는 정확히 동작하지 않을 수 있음 등)
	7. **실험자 편향(Experimental bias)**: 인공지능 모델이 실험자의 믿음과 일치하는 결과를 나타낼 때까지 학습시킴

사용자 (User)	1. **과잉 확신(Overconfidence)**: 인공지능 모델에 대한 사용자의 이전 경험으로 인해 해당 모델이 모든 상황에서 사용할 수 있다고 믿음
	2. **행동 실패(Failure to act)**: 사용자가 인공지능 모델을 신뢰하지 않고 무시함
	3. **인지된 위험(Perceived risk)**: 사용자가 실제보다 저위험으로 인식하여 인공지능 모델을 더 신뢰하거나 작업을 위임함
	4. **사용자 업무량, 시간 제약(User workload, Time constraints)**: 바쁜 사용자는 인공지능 모델을 더 신뢰하는 경향이 있음
	5. **자신감(Self-confidence)**: 사용자는 인공지능 모델이 '우월한 판단'을 한다고 믿고 따를 수 있음
	6. **사회적 신뢰 차이(Variation in social trust)**: 다른 사용자 집단(예. 전문성 혹은 국가가 다름)은 인공지능 모델에 대한 다양한 신뢰 수준을 갖고 있어 해당 신뢰에 대한 개발자의 가정이 전체 사용자 집단에 적용되지 않을 수 있음
적응형 시스템 (Adaptive System)	1. **연속 학습(Continuously learning)**: 인공지능 배포 이후에 지속적으로 데이터를 학습하는 것으로 품질이 낮은 데이터로 학습하면 시스템 성능이 저하될 수 있음
기타	1. **지식 부족(Lack of knowledge)**: 데이터의 의미와 맥락에 대한 이해없는 데이터 축적 (예. 일반적으로 천식은 고위험군으로 조기 치료되어 사망률이 낮아 사망률 기반 인공지능 모델은 해당 질병을 고위험군으로 분류하지 못함)

2. 논문리뷰

설명 가능한 인공지능 기반 의료 AI 기술 연구 동향 Top 0.5%

A Survey on Artificial Intelligence based Explainable Artificial Intelligence

저널정보

대한전자공학회
대한전자공학회 학술대회 | 학술대회자료
2021년도 대한전자공학회 하계종합학술대회 논문집
2021.06 | 2,403 - 2,406 (4page)

저자정보

김재현 (한양대학교)
김유신 (한양대학교)
이세종 (한양대학교)
안세영 (한양대학교)
노재원 (한양대학교)
저자 전체보기 >

이용수 **2,690** 내서재 **80**

설명 가능한 인공지능 기반 의료 AI 기술 연구 동향

김재현[1], 김유신[2], 이세종[3], 안세영[4], 노재원[5], 김종훈[6], 조성현*
한양대학교 인공지능융합학과[1], 한양대학교 컴퓨터공학과*[2,3,4,5,6]
바이오인공지능융합전공[1,2,3,4]
e-mail : {insam2802, hpwgg045, kingsaejong, tpdud1014, wodnjs1451, chopro, iproj2}@hanyang.ac.kr

A Survey on Artificial Intelligence based Explainable Artificial Intelligence

Jaehyeon Kim[1], Yushin Kim[2], Sejong Lee[3], Seyoung Ahn[4], Jaewon Noh[5], Jonghun Kim[6], Sunghyun Cho*
Dept. of Applied Artificial Intelligence, Hanyang University[1]
Dept. of Computer Science and Engineering, Hanyang University*[2,3,4,5,6]
Major in Bio-Artificial Intelligence[1,2,3,4]

Abstract

Artificial intelligence is being used in a variety of fields, including medicine. But their lack of interpretability and explainability stand as one of the main drawbacks. To solve this problem, explainable artificial intelligence has been studied in healthcare. In this paper, we summarize recent advances in explainable artificial intelligence technologies and introduce the use of explainable artificial intelligence studies in medicine.

I. 서론

인공지능 (Artificial intelligence, AI) 기술이 발전함에 따라 다양한 분야에서 활용되고 있다. 특히 의료 분야에서는 질병의 초기 발견 [1], 질환의 분류 [2], 최적화된 약물 용량 결정 [3] 등의 문제를 해결하려는 연구가 활발히 진행되고 있다. 하지만 AI 알고리즘의 복잡성으로 인해 블랙박스 문제가 발생한다. 블랙박스 문제 AI의 의사결정 근거 및 결정 도출 과정의 신뢰성을 보장할 수 없다. 이로 인해 강한 신뢰성을 필요로 하는 의료 분야의 특성상 AI의 의사결정을 전적으로 신뢰할 수 없다는 문제가 발생한다. 따라서 신뢰성을 보장하기 위해 의료 분야에 설명 가능한 AI (Explainable AI, XAI)를 적용하는 연구가 활발히 진행되고 있다. XAI 기술은 AI의 의사결정에 대해 해석적이고, 직관적이며 사람이 이해 가능한 설명을 제공한다. 이를 통해 AI의 의사결정에 대한 투명성 및 신뢰성을 확보할 수 있다.

II. 본론

본 장에서는 설명 가능한 AI의 대표적인 기법인 Class Activation Mapping (CAM) [4], Layer-wise Relevance Propagation (LRP) [5] 및 Local Interpretable Model Agnostic Explanations (LIME) [6]에 대하여 소개하고 해당 기술을 의료 분야에 활용한 연구에 대하여 기술한다.

2.1 CAM 기반 XAI 기술 활용 현황

CAM은 이미지 분류 모델에서 입력 데이터의 어떤 부분이 클래스를 예측하는데 영향을 주었는지 시각화하는 방법이다. CAM은 기존의 Convolution Neural Network (CNN)에서 사용하는 fully-connected layer 대신 global average pooling (GAP)을 사용한다. GAP는 convolution layer에서 분류하려는 클래스의 수만큼의 채널을 갖게 한 후 각 채널을 기준으로 평균 함을

2. 논문리뷰

1) 주제 살펴보기

I. 서론

인공지능 (Artificial intelligence, AI) 기술이 발전함에 따라 다양한 분야에서 활용되고 있다. 특히 의료 분야에서는 질병의 초기 발견 [1], 질환의 분류 [2], 최적화된 약물 용량 결정 [3] 등의 문제를 해결하려는 연구가 활발히 진행되고 있다. 하지만 AI 알고리즘의 복잡성으로 인해 블랙박스 문제가 발생한다. 블랙박스 문제 AI의 의사결정 근거 및 결정 도출 과정의 신뢰성을 보장할 수 없다. 이로 인해 강한 신뢰성을 필요로 하는 의료 분야의 특성상 AI의 의사결정을 전적으로 신뢰할 수 없다는 문제가 발생한다. 따라서 신뢰성을 보장하기 위해 의료 분야에 설명 가능한 AI (Explainable AI, XAI)를 적용하는 연구가 활발히 진행되고 있다. XAI 기술은 AI의 의사결정에 대해 해석적이고, 직관적이며 사람이 이해 가능한 설명을 제공한다. 이를 통해 AI의 의사결정에 대한 투명성 및 신뢰성을 확보할 수 있다.

II. 본론

본 장에서는 설명 가능한 AI의 대표적인 기법인 Class Activation Mapping (CAM) [4], Layer-wise Relevance Propagation (LRP) [5] 및 Local Interpretable Model Agnostic Explanations (LIME) [6]에 대하여 소개하고 해당 기술을 의료 분야에 활용한 연구에 대하여 기술한다.

2.1 CAM 기반 XAI 기술 활용 현황

2.2 LRP 기반 XAI 기술 활용 현황

2.3 LIME 기반 XAI 기술 활용 현황

2. 논문리뷰

1) 주제 살펴보기

Ⅲ. 결론 및 향후 연구 방향

본 논문에서는 XAI 기술을 활용한 의료 인공지능 연구들을 소개한다. 이를 통해 의료 인공지능에 XAI 기법을 적용하여 이미지, 수치 데이터 및 신호 데이터 등의 다양한 의료 데이터 형식에 따른 의사결정의 근거 및 결과 도출 과정을 설명할 수 있다. 그러나 아직 XAI의 설명 가능성에 대한 평가 기준이 명확하지 않다. 현재 설명 가능성에 대한 평가는 XAI가 제공한 정보을 사람이 개입하여 직접 평가해야 한다는 한계를 갖고 있다. 따라서 추후 XAI 모델 간의 설명 가능성 대해 비교할 수 있는 평가 기준을 설계하는 연구가 필요하다.

2. 논문리뷰

2) 내용 파악

표 1. 의료 분야에서 XAI 기술 활용 현황 비교

기법	확장성	연구	사용 분류 모델	분류 모델의 성능 수치 (%)	데이터 형식
CAM	CNN 모델	[7]	CNN	91.7 (Accuracy)	MRI Image
		[8]	ResNet	84.5 (Accuracy)	MRI Image
		[13]	CNN-CAM	98.5 (Accuracy)	Image
LRP	딥러닝 모델	[9]	LSTM	81.2 (Accuracy)	Numeric
		[10]	CNN	88.4 (Accuracy)	MRI Image
		[14]	CNN	88.0 (Accuracy)	MRI Image
		[15]	CNN	86.7 (Accuracy)	MRI Image
LIME	모든 모델	[11]	KNN, CNN	90.3 (F1-score)	Signal
		[12]	SVM, KNN, Random Forest, etc.	80.5 (Accuracy)	Numeric
		[16]	Random Forest	58.1 (Accuracy)	Numeric
		[17]	CNN	81.8 (Accuracy)	Image
		[18]	ResNet, VGG	89.0 (Accuracy)	Image
		[19]	SVM, XGBoost, Random Forest	91.9 (Accuracy)	Numeric

XAI는 크게 CAM, LRP 및 LIME 기반 기술로 나눌 수 있다. 표 1은 의료 분야에서 XAI 기술 활용 현황을 연구 현황에 따라 비교 정리한 표이다. CAM 기법은 이미지의 어떤 부분이 분류에 영향을 주었는지 설명할 수 있다. 하지만 CNN 기반 이미지 분류 모델에만 사용할 수 있기 때문에 확장성이 없다는 단점이 있다. LRP는 분해를 통해 계층별 기여도를 계산할 수 있다. CAM 기법과 달리 더욱 다양한 모델에 적용할 수 있지만 딥러닝 모델에만 적용 가능하기 때문에 딥러닝 모델에서 확장성을 가진다. LIME은 데이터의 어느 영역을 분류 근거로 사용했는지 설명한다. LIME은 위의 두 모델과 달리 모든 예측 모델에 적용할 수 있다. 따라서 모든 데이터 형식의 분류에 사용할 수 있기 때문에 모든 모델에 대하여 확장 가능성이 있다. 하지만 분류 모델의 복잡성이 미리 정의되어 있어야 한다는 단점이 있다.

2. 논문리뷰

2) 내용 파악

I. 서론

인공지능 (Artificial intelligence, AI) 기술이 발전함에 따라 다양한 분야에서 활용되고 있다. 특히 의료 분야에서는 질병의 초기 발견 [1], 질환의 분류 [2], 최적화된 약물 용량 결정 [3] 등의 문제를 해결하려는 연구가 활발히 진행되고 있다. 하지만 AI 알고리즘의 복잡성으로 인해 블랙박스 문제가 발생한다. 블랙박스 문제 AI의 의사결정 근거 및 결정 도출 과정의 신뢰성을 보장할 수 없다. 이로 인해 강한 신뢰성을 필요로 하는 의료 분야의 특성상 AI의 의사결정을 전적으로 신뢰할 수 없다는 문제가 발생한다. 따라서 신뢰성을 보장하기 위해 의료 분야에 설명 가능한 AI (Explainable AI, XAI)를 적용하는 연구가 활발히 진행되고 있다. XAI 기술은 AI의 의사결정에 대해 해석적이고, 직관적이며 사람이 이해 가능한 설명을 제공한다. 이를 통해 AI의 의사결정에 대한 투명성 및 신뢰성을 확보할 수 있다.

[1] Learning representations for the early detection of sepsis with deep neural networks (2017)

[2] Optimal feature-based multi-kernel SVM approach for thyroid disease classification (2018)

[3] Artificial intelligence in drug combination therapy (2019)

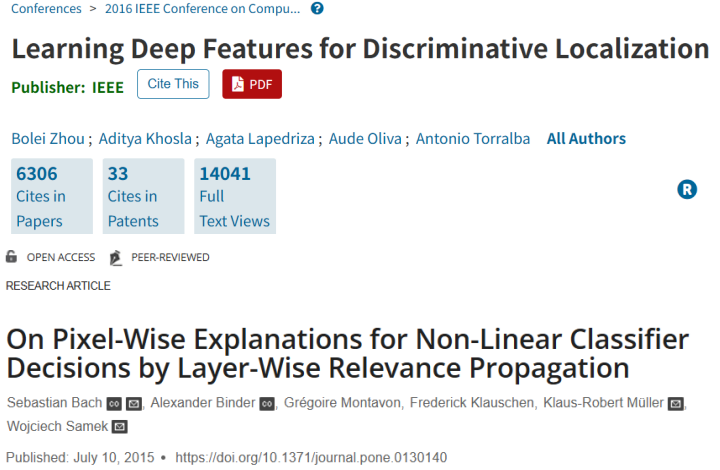
2. 논문리뷰

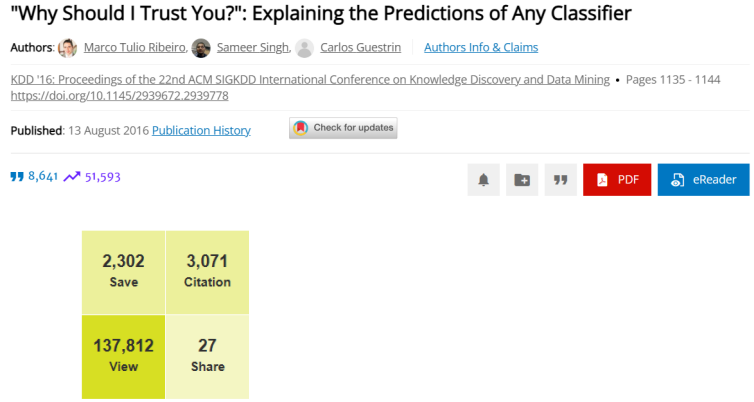
2) 내용 파악

II. 본론

본 장에서는 설명 가능한 AI의 대표적인 기법인 Class Activation Mapping (CAM) [4], Layer-wise Relevance Propagation (LRP) [5] 및 Local Interpretable Model Agnostic Explanations (LIME) [6]에 대하여 소개하고 해당 기술을 의료 분야에 활용한 연구에 대하여 기술한다.

[4] Learning deep features for discriminative localization (2016)
[5] On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation (2015)
[6] Why should i trust you?:Explaining the Predictions of Any Classifier (2016)





2. 논문리뷰

2) 내용 파악

	CAM	LRP	LIME
설명 방식	입력 데이터의 어떤 부분이 클래스를 예측하는데 영향을 주었는지 시각화		어느 영역을 분류의 근거로 사용했는지 설명
작동원리	CNN의 FC레이어 대신 Global Average Pooling 사용	분해를 통해 계층별 기여도 계산	입력 데이터를 변형하여 인식 단위로 쪼개어 데이터를 해석
활용 예시	[7] MRI기반 근위축증 이미지 분류 [8] 포도막 흑색종 분류	[9] 항호르몬 요법 영향 측정분류 [10] 다발성 경화증 진단	[11] 심혈관 질환의 치료와 예방을 위한 심장 박동 종류 분류 [12] Pima 당뇨병 분류 모델

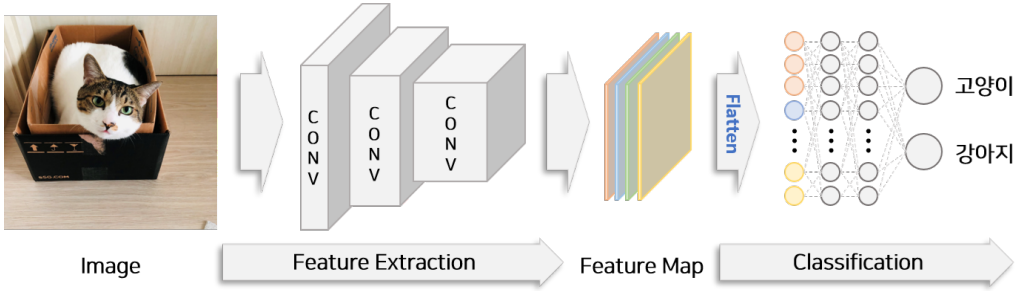
표 1. 의료 분야에서 XAI 기술 활용 현황 비교

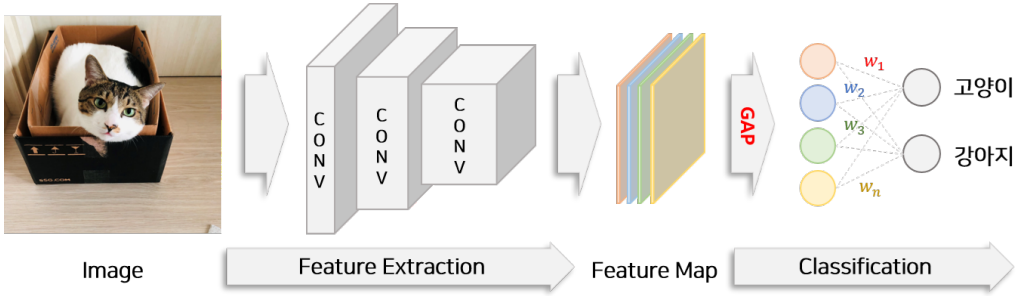
기법	확장성	연구	사용 분류 모델	분류 모델의 성능 수치 (%)	데이터 형식
CAM	CNN 모델	[7]	CNN	91.7 (Accuracy)	MRI Image
		[8]	ResNet	84.5 (Accuracy)	MRI Image
		[13]	CNN-CAM	98.5 (Accuracy)	Image
LRP	딥러닝 모델	[9]	LSTM	81.2 (Accuracy)	Numeric
		[10]	CNN	88.4 (Accuracy)	MRI Image
		[14]	CNN	88.0 (Accuracy)	MRI Image
		[15]	CNN	86.7 (Accuracy)	MRI Image
LIME	모든 모델	[11]	KNN, CNN	90.3 (F1-score)	Signal
		[12]	SVM, KNN, Random Forest, etc.	80.5 (Accuracy)	Numeric
		[16]	Random Forest	58.1 (Accuracy)	Numeric
		[17]	CNN	81.8 (Accuracy)	Image
		[18]	ResNet, VGG	89.0 (Accuracy)	Image
		[19]	SVM, XGBoost, Random Forest	91.9 (Accuracy)	Numeric

2. 논문리뷰

2) 내용 파악

	CAM (Class Activation Mapping)
설명 방식	입력 데이터의 어떤 부분이 클래스를 예측하는데 영향을 주었는지 시각화
작동원리	CNN의 Flatten 단계 Global Average Pooling 사용
활용 예시	[7] MRI기반 근위축증 이미지 분류 [8] 포도막 흑색종 분류





*GAP: 가로세로값을 모두 더해 하나의 특징을 추출하는 방법

2. 논문리뷰

2) 내용 파악

	CAM (Class Activation Mapping)
설명 방식	입력 데이터의 어떤 부분이 클래스를 예측하는데 영향을 주었는지 시각화
작동원리	CNN의 Flatten 단계 Global Average Pooling 사용
활용 예시	[7] MRI기반 근위축증 이미지 분류 [8] 포도막 흑색종 분류

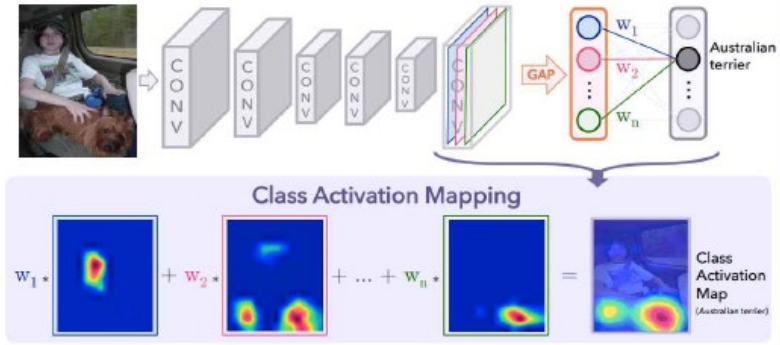
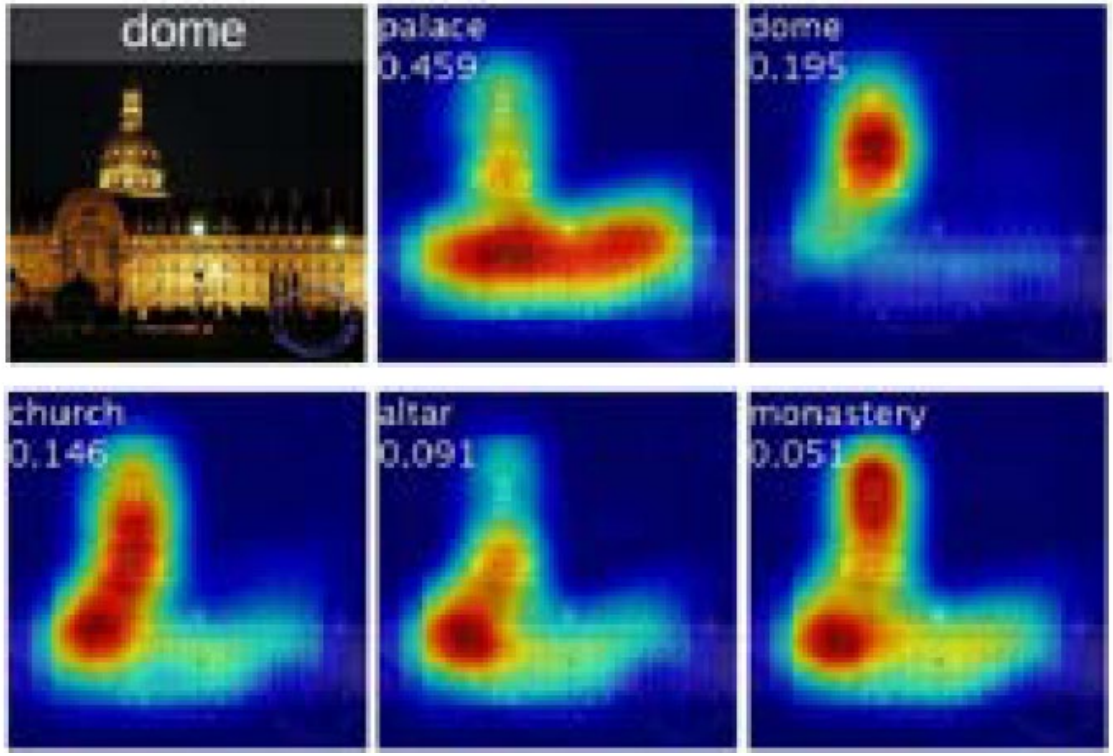


그림 1. CAM을 이용해 시각화된 heatmap

2. 논문리뷰

2) 내용 파악

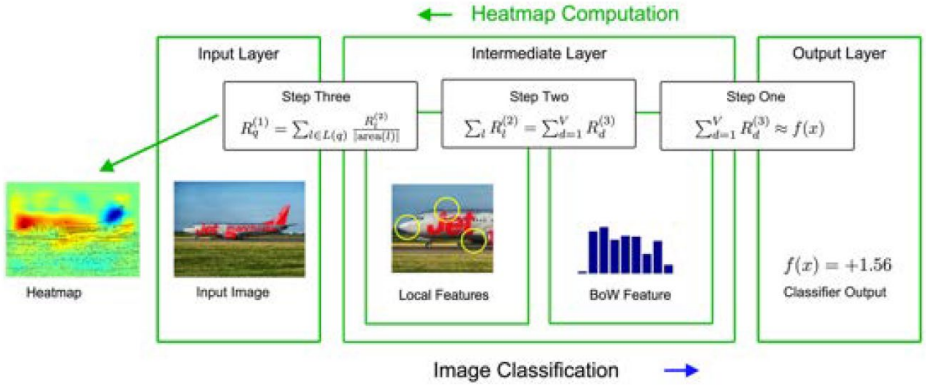
	CAM (Class Activation Mapping)



2. 논문리뷰

2) 내용 파악

	LRP (Layer-wise Relevance Propagation)
설명 방식	입력 데이터의 어떤 부분이 클래스를 예측하는데 영향을 주었는지 시각화
작동원리	분류 결과를 역순으로 탐지하며 분해 > 기여도 표시하여 모델 해석
활용 예시	[9] 항호르몬 요법 영향 측정분류 [10] 다발성 경화증 진단



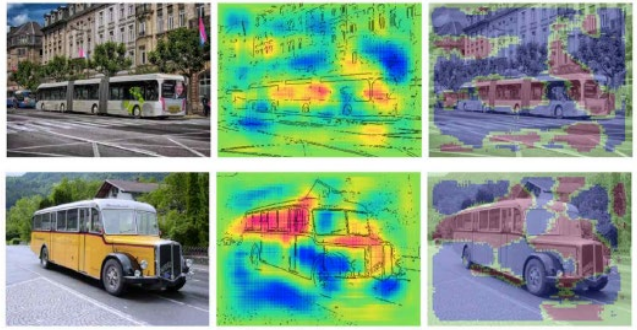
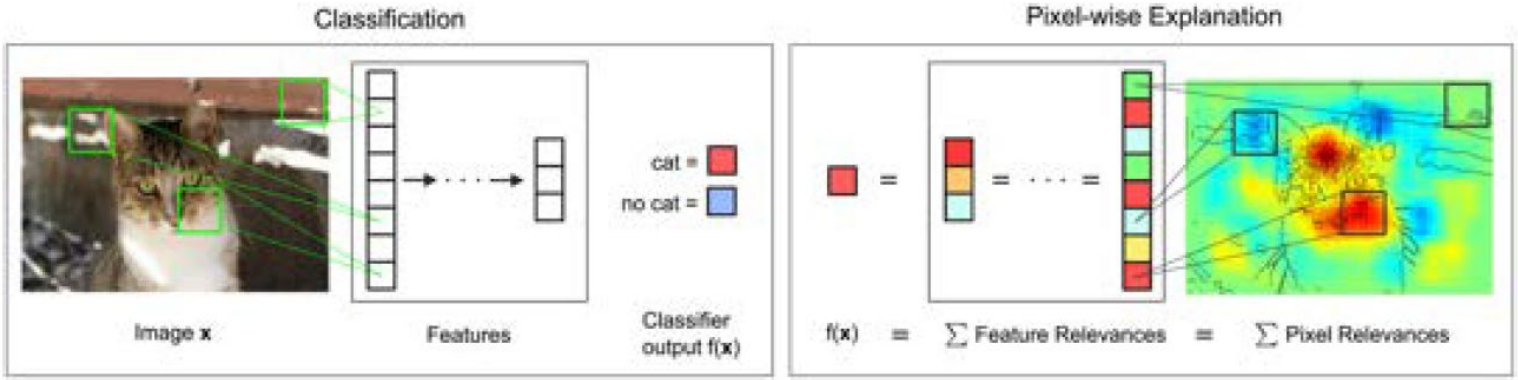


그림 2. LRP를 통해 시각화된 heatmap

2. 논문리뷰

2) 내용 파악

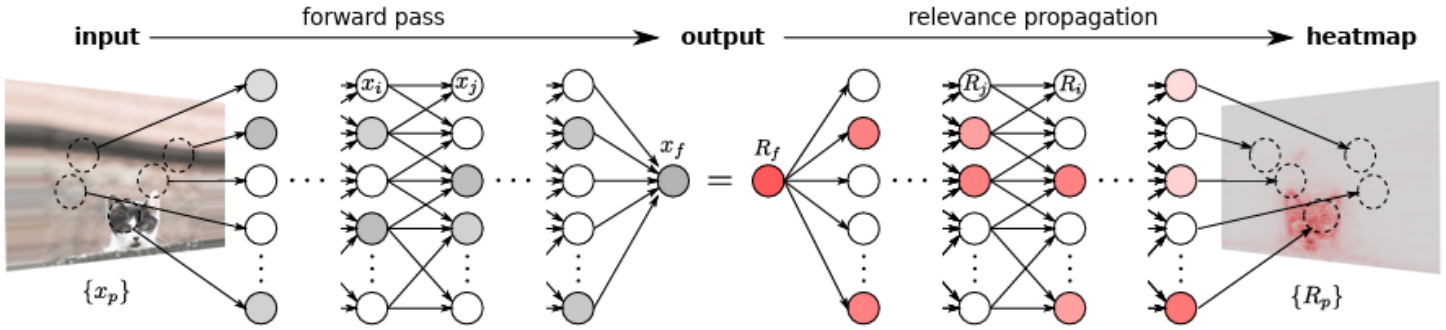
	LRP (Layer-wise Relevance Propagation)
설명 방식	입력 데이터의 어떤 부분이 클래스를 예측하는데 영향을 주었는지 시각화
작동원리	분류 결과를 역순으로 탐지하며 분해 > 기여도 표시하여 모델 해석
활용 예시	[9] 항호르몬 요법 영향 측정분류 [10] 다발성 경화증 진단



2. 논문리뷰

2) 내용 파악

	LRP (Layer-wise Relevance Propagation)
설명 방식	입력 데이터의 어떤 부분이 클래스를 예측하는데 영향을 주었는지 시각화
작동원리	분류 결과를 역순으로 탐지하며 분해 > 기여도 표시하여 모델 해석
활용 예시	[9] 항호르몬 요법 영향 측정분류 [10] 다발성 경화증 진단



2. 논문리뷰

2) 내용 파악

	LIME (Local Interpretable Model-agnostic Explanations)
설명 방식	어느 영역을 분류의 근거로 사용했는지 설명
작동원리	입력 데이터를 변형하여 인식 단위로 쪼개어 데이터를 해석
활용 예시	[11] 심혈관 질환의 치료와 예방을 위한 심장 박동 종류 분류 [12] Pima 당뇨병 분류 모델

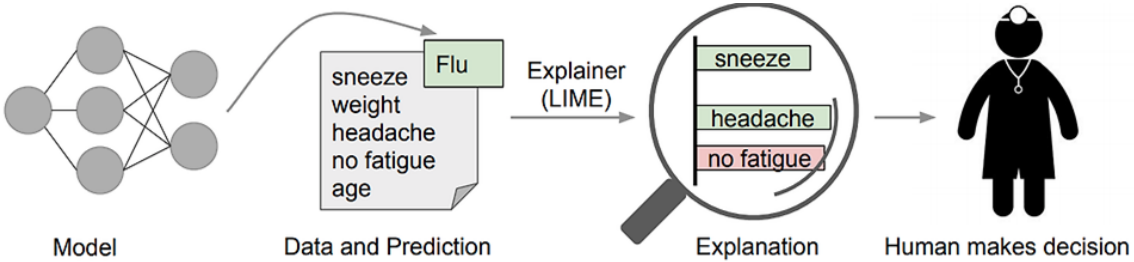
(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

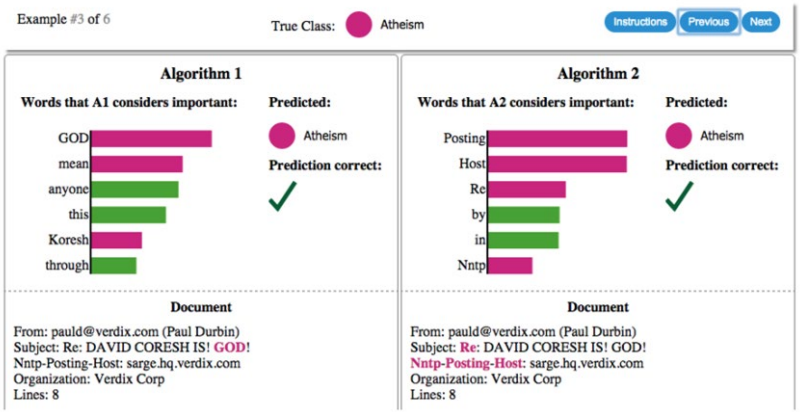
Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

2. 논문리뷰

2) 내용 파악

	LIME (Local Interpretable Model-agnostic Explanations)





2. 논문리뷰

3) 마무리

Ⅲ. 결론 및 향후 연구 방향

본 논문에서는 XAI 기술을 활용한 의료 인공지능 연구들을 소개한다. 이를 통해 의료 인공지능에 XAI 기법을 적용하여 이미지, 수치 데이터 및 신호 데이터 등의 다양한 의료 데이터 형식에 따른 의사결정의 근거 및 결과 도출 과정을 설명할 수 있다. 그러나 아직 XAI의 설명 가능성에 대한 평가 기준이 명확하지 않다. 현재 설명 가능성에 대한 평가는 XAI가 제공한 정보을 사람이 개입하여 직접 평가해야 한다는 한계를 갖고 있다. 따라서 추후 XAI 모델 간의 설명 가능성 대해 비교할 수 있는 평가 기준을 설계하는 연구가 필요하다.

[표 5. 거대언어모델이 적용된 의료기기의 성능 검증 지표 예시]

항 목		설 명
블루 (Bilingual Evaluation Understudy, BLEU)		모델이 생성한 글과 참조표준 글 사이의 연속적 단어 나열의 일치 정도를 정밀도에 중점을 두고 평가하는 지표
루지 (Recall-Oriented Understudy for Gisting Evaluation, ROUGE)		모델이 생성한 글과 참조표준 글 사이의 연속적 단어 나열의 일치 정도를 민감도에 중점을 두고 평가하는 지표
메테오 (Metric for Evaluation of Translation with Explicit ORdering, METEOR)		모델이 생성한 글과 참조표준 글 사이의 연속적 단어 나열의 일치 정도를 F1 점수에 중점을 두고 평가하는 지표
분류 모델 평가 지표	정밀도	참조표준 내 연속적으로 나열된 단어 중 모델이 생성한 글에 포함되는 단어의 수 또는 참조표준에 따라 진양성인 단어들과 모델이 생성한 결과를 비교하여 정밀도/민감도/정확도/F1 Score를 산출
	민감도	
	정확도	
	F1 점수	

감사합니다