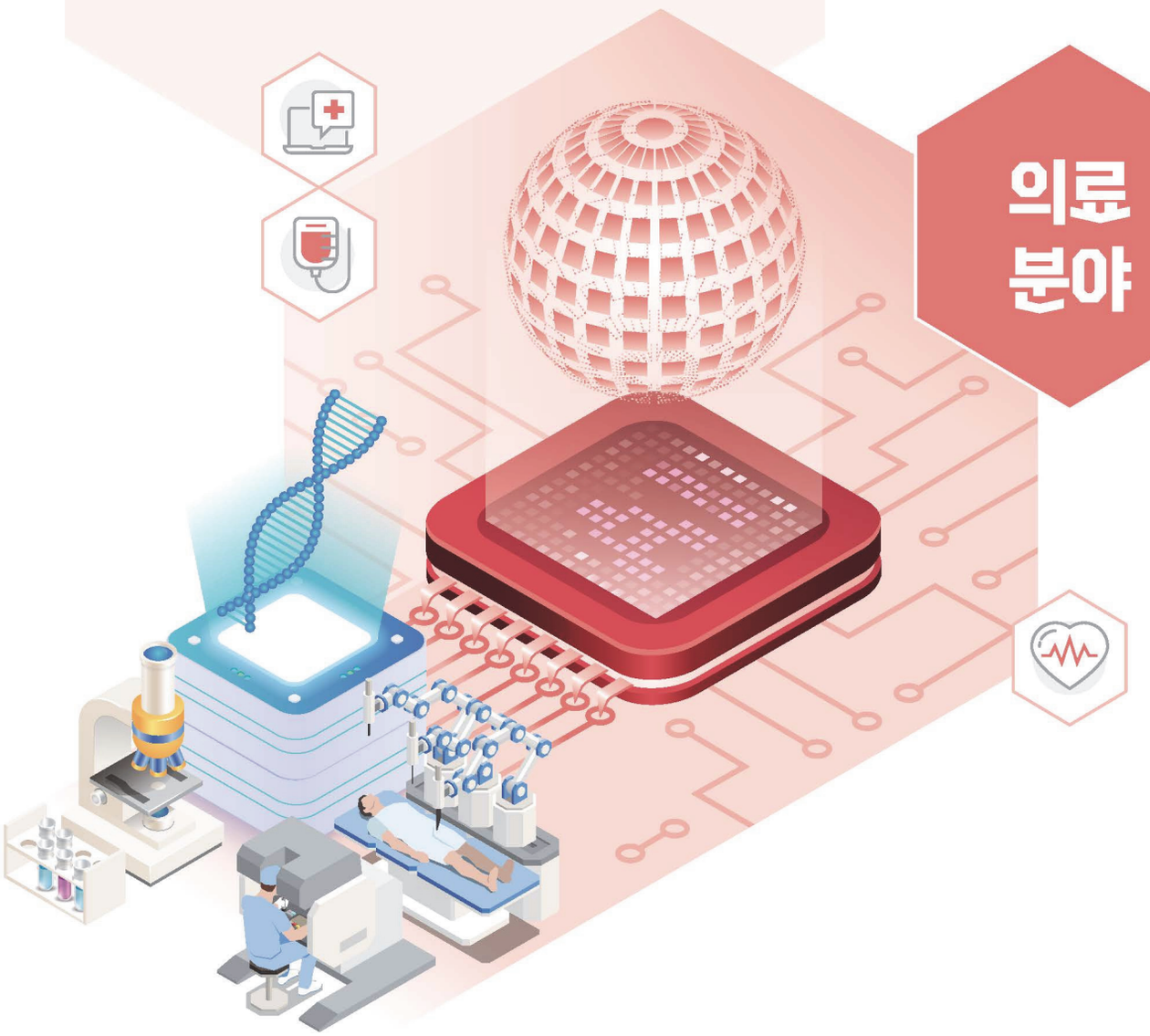
2023

신뢰할 수 있는 인공지능 개발 안내서

의료 분야



과학기술정보통신부
Ministry of Science and ICT

TTA 한국정보통신기술협회
Telecommunications Technology Association

Jinhan M&B

일러두기

- 본 안내서는 과학기술정보통신부 「AI 신뢰성 검증체계 고도화」 사업의 연구 결과로서 내용의 무단 전재를 금합니다.

- 또한, 안내서의 내용을 가공·인용하는 경우에는 반드시 '과학기술정보통신부·한국정보통신기술협회 《2023 신뢰할 수 있는 인공지능 개발 안내서 – 의료 분야》'의 출처를 밝혀 주시기 바랍니다.

- 본 안내서는 의료 인공지능 서비스 및 제품을 개발하는 과정에서 참고 자료로 활용할 수 있도록 편찬되었습니다. 본 안내서는 기업의 업무 환경과 상황, 개발 목적 등을 고려하여 필요하신 내용을 취사선택하여 활용하시기 바랍니다.

- 본 안내서의 의료·인공지능 동향 및 기술 정보는 2023년 2월 기준으로 서술되었습니다.

- 인공지능 신뢰성은 사회 구성원의 다양한 의견과 논의를 통해 합의와 공감대를 이루어야 하는 개념으로, 본 안내서가 이러한 담론의 수집과 논의의 장을 마련하는 촉매제가 되었으면 하는 바람입니다. 이를 위해 폭넓고 심도 있는 의견을 듣고 반영하고자 하오니, 많은 참여와 관심 부탁드립니다.

- 본 안내서는 한국정보통신기술협회가 운영하는 TrustOps 웹페이지(2023년 하반기 공개 예정)에도 콘텐츠가 공개되어 있으므로 참고하시면 더 편리하게 활용하실 수 있습니다.

- 의료 외 분야는 《2023 신뢰할 수 있는 인공지능 개발 안내서 – 일반 분야》를 참고해주시기 바라며, 특화될 서비스 분야는 점차 확대해나갈 예정입니다.

CONTENTS

안내서 활용을 위한 체크리스트

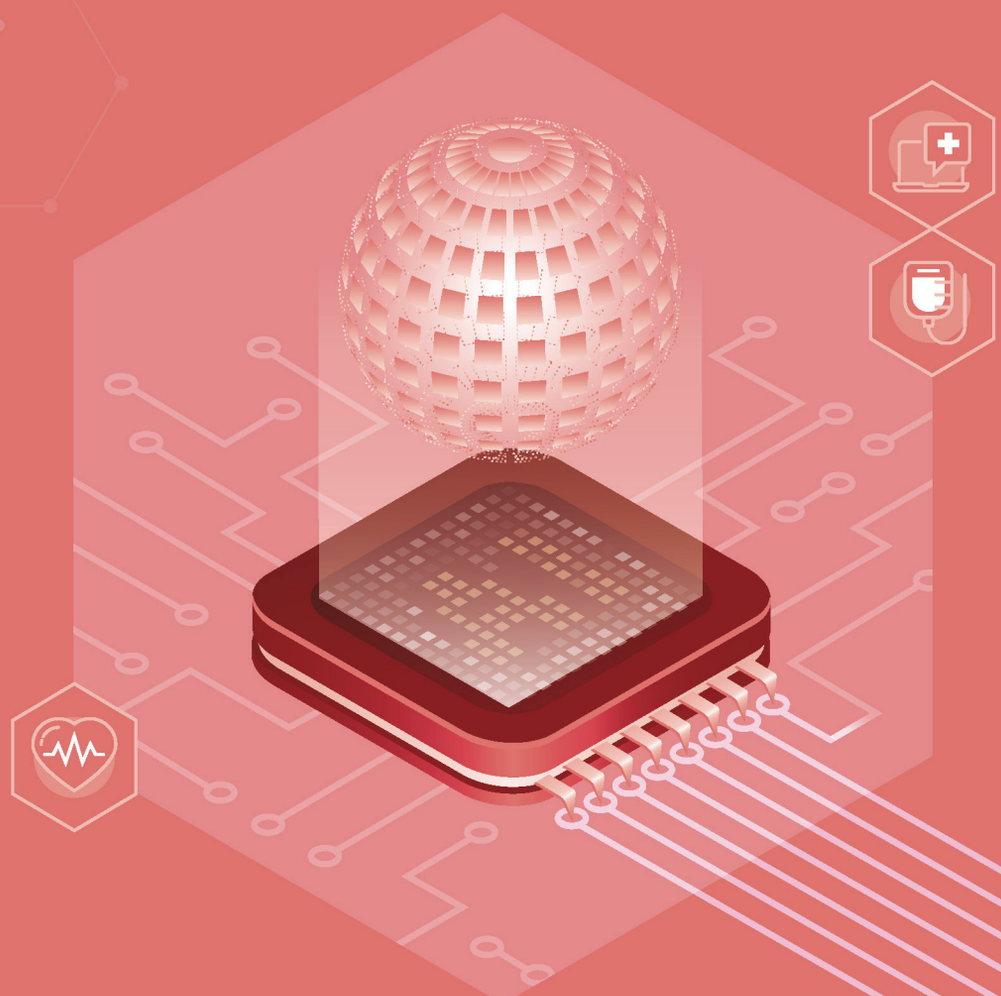
안내서 활용을 위한 체크리스트

생명주기	요구사항 및 체크리스트		Yes	No	N/A
1 계획 및 설계	요구사항 01	인공지능 시스템에 대한 위험관리 계획 및 수행			
	01-1	인공지능 시스템 생명주기에 걸쳐 나타날 수 있는 위험 요소를 분석하였는가?	☐	☐	☐
	01-1a	인공지능 시스템의 위험 요소를 도출하고 이의 파급효과를 파악하였는가?	☐	☐	☐
	01-2	위험 요소를 제거 및 방지하거나 영향을 완화하기 위한 방안을 마련하였는가?	☐	☐	☐
	01-2a	위험 요소 제거 방안을 도출하고 파급효과가 감소하였는지 확인하였는가?	☐	☐	☐
	요구사항 02	인공지능 거버넌스governance 체계 구성			
	02-1	인공지능 거버넌스에 대한 지침 및 규정을 수립하였는가?	☐	☐	☐
	02-1a	내부적으로 준수해야 할 인공지능 거버넌스에 대한 지침 및 규정을 마련하였는가?	☐	☐	☐
	02-2	인공지능 거버넌스를 위한 조직을 구성하고 인력 구성에 대해 검토하였는가?	☐	☐	☐
	02-2a	인공지능 거버넌스를 위한 조직을 구성하였는가?	☐	☐	☐
	02-2b	인공지능 거버넌스를 위한 조직은 충분히 훈련된 인력으로 구성하였는가?	☐	☐	☐
	02-3	인공지능 거버넌스 체계가 올바로 이행되고 있는지 감독하고 있는가?	☐	☐	☐
	02-3a	인공지능 거버넌스에 대한 내부 지침 및 규정 준수 여부를 감독하고 있는가?	☐	☐	☐
	02-4	인공지능 거버넌스 조직이 신규 및 기존 시스템의 차이점을 분석하였는가?	☐	☐	☐
	02-4a	신규 시스템과 기존 동일 목적의 의료 시스템을 비교하여 안전성과 효율성 확보가 가능한지 분석하였는가?	☐	☐	☐
	요구사항 03	인공지능 시스템의 신뢰성 테스트 계획 수립			
	03-1	인공지능 시스템의 특성을 고려한 테스트 환경을 설계하였는가?	☐	☐	☐
	03-1a	테스트 환경 결정 시 인공지능 시스템의 운영환경을 고려하였는가?	☐	☐	☐
	03-1b	가상테스트 환경이 필요한 인공지능 시스템의 경우, 시뮬레이터를 확보하였는가?	☐	☐	☐
	03-2	인공지능 시스템의 테스트 설계에 필요한 협의 체계를 구성하였는가?	☐	☐	☐
	03-2a	인공지능 시스템의 기대 출력을 결정하기 위한 협의 체계를 구성하였는가?	☐	☐	☐
	03-2b	설명가능성 및 해석가능성 확인을 위한 사용자 평가단을 구성하였는가?	☐	☐	☐
2 데이터 수집 및 처리	요구사항 04	데이터의 활용을 위한 상세 정보 제공			
	04-1	데이터의 명확한 이해와 활용을 지원하는 상세한 정보를 제공하는가?	☐	☐	☐
	04-1a	정제 전과 후의 데이터 특성을 설명하였는가?	☐	☐	☐
	04-1b	학습 데이터와 메타데이터metadata를 구분하고 각 명세자료를 확보하였는가?	☐	☐	☐
	04-1c	보호변수protective attribute의 선정 이유 및 반영 여부를 설명하였는가?	☐	☐	☐
	04-1d	라벨링 작업자를 위해 교육을 시행하고 작업 가이드 문서를 마련하였는가?	☐	☐	☐

생명주기	요구사항 및 체크리스트		Yes	No	N/A
2 데이터 수집 및 처리	04-2	데이터의 출처는 기록 및 관리되고 있는가?	☐	☐	☐
	04-2a	신뢰할 수 있는 출처로부터 제공되는 데이터셋을 사용하였는가?	☐	☐	☐
	04-2b	오픈소스 데이터셋을 활용하는 경우, 출처를 명시하였는가?	☐	☐	☐
	요구사항 05	**데이터 강건성 확보를 위한 이상abnormal 데이터 점검**			
	05-1	이상 데이터의 식별 및 정상 여부를 점검하였는가?	☐	☐	☐
	05-1a	전체 학습용 데이터 분포를 시각화하여 발생 가능한 오류들을 확인하였는가?	☐	☐	☐
	05-1b	학습 데이터 이상값 식별 기법을 적용하였는가?	☐	☐	☐
	05-2	데이터 공격에 대한 방어 수단을 강구하였는가?	☐	☐	☐
	05-2a	데이터 중독poisoning, 회피evasion 등 공격에 대한 방어 대책을 마련하였는가?	☐	☐	☐
	요구사항 06	**수집 및 가공된 학습 데이터의 편향 제거**			
	06-1	데이터 수집 시, 인적·물리적 요인으로 인한 편향 완화 방안을 마련하였는가?	☐	☐	☐
	06-1a	인적 편향을 제거하기 위한 절차적, 기술적 수단을 적용하였는가?	☐	☐	☐
	06-1b	데이터의 다양성 확보를 위해 이기종 수집 장치를 활용하였는가?	☐	☐	☐
	06-1c	하드웨어로 인해 발생할 수 있는 데이터의 편향을 점검하였는가?	☐	☐	☐
	06-2	학습에 사용되는 특성feature을 분석하고 선정 기준을 마련하였는가?	☐	☐	☐
	06-2a	보호변수 선정 시 충분한 분석을 수행하였는가?	☐	☐	☐
	06-2b	편향을 발생시킬 수 있는 특성의 영향력을 완화하였는가?	☐	☐	☐
	06-2c	데이터 전처리 시 특성이 과도하게 제거되었는지 검토하였는가?	☐	☐	☐
	06-3	데이터 라벨링 시, 발생 가능한 편향을 확인하고 방지하였는가?	☐	☐	☐
	06-3a	데이터 라벨링 기준을 명확히 수립하고 작업자에게 제공하였는가?	☐	☐	☐
	06-3b	다양한 데이터 라벨링 작업자를 섭외하기 위해 노력하였는가?	☐	☐	☐
	06-3c	다양한 데이터 라벨링 검수자를 확보하기 위해 노력하였는가?	☐	☐	☐
	06-4	데이터의 편향 방지를 위한 샘플링을 수행하였는가?	☐	☐	☐
	06-4a	편향 방지를 위한 샘플링 기법을 적용하였는가?	☐	☐	☐
3 인공지능 모델 개발	요구사항 07	**오픈소스 라이브러리의 보안성 및 호환성 확보**			
	07-1	오픈소스 라이브러리의 안정성을 확인하였는가?	☐	☐	☐
	07-1a	활성화된 오픈소스 라이브러리를 사용하였는가?	☐	☐	☐
	07-2	오픈소스 라이브러리의 위험 요소는 관리되고 있는가?	☐	☐	☐
	07-2a	사용 중인 오픈소스 라이브러리의 라이선스 준수사항을 이행하였는가?	☐	☐	☐
	07-2b	사용 중인 오픈소스 라이브러리의 호환성 및 보안취약점을 확인하였는가?	☐	☐	☐

생명주기	요구사항 및 체크리스트	Yes	No	N/A
3 인공지능 모델 개발	**요구사항 08** 인공지능 모델의 편향 제거			
	08-1 모델 편향을 제거하는 기법을 적용하였는가?	☐	☐	☐
	08-1a 개발하려는 모델에 맞게 편향제거 기법을 선택하였는가?	☐	☐	☐
	08-1b 편향성 평가 및 모니터링을 위한 정량적 지표를 선정하고 관리하는가?	☐	☐	☐
	요구사항 09 인공지능 모델 공격에 대한 방어 대책 수립			
	09-1 모델 추출 공격(model extraction attack)에 대한 방어 방안을 수립하였는가?	☐	☐	☐
	09-1a 모델 추출 공격에 대비하는 방어 기법을 적용하였는가?	☐	☐	☐
	09-2 모델 회피 공격(model evasion attack)에 대한 방어 방안을 수립하였는가?	☐	☐	☐
	09-2a 모델 회피 공격에 대비하는 방어 기법을 적용하였는가?	☐	☐	☐
	요구사항 10 인공지능 모델 명세 및 추론 결과에 대한 설명 제공			
	10-1 사용자가 모델 추론 결과의 도출 과정을 수용할 수 있도록 근거를 제공하는가?	☐	☐	☐
	10-1a XAI(eXplainable AI 기술 적용이 가능한 경우, 인공지능 모델의 추론 결과를 설명하기 위한 기법 적용에 대해 검토하였는가?	☐	☐	☐
	10-1b XAI 기술 외에도 수용가능한 모델 추론 결과의 근거를 제공하는가?	☐	☐	☐
	10-2 인공지능 모델 상세 문서를 통해 모델의 명세를 투명하게 제공하는가?	☐	☐	☐
	10-2a 시스템 개발 과정과 모델 작동 방식에 대한 세부 정보가 설명된 문서를 작성하였는가?	☐	☐	☐
	10-3 필요 시, 인공지능 모델 추론 결과에 대한 설명을 제공하는가?	☐	☐	☐
	10-3a 모델 추론 결과에 대한 설명이 필요한지 검토하였는가?	☐	☐	☐
	10-3b 사용자에게 인공지능 모델 추론 결과에 대한 설명을 제공하였는가?	☐	☐	☐
4 시스템 구현	**요구사항 11** 인공지능 시스템 구현 시 발생 가능한 편향 제거			
	11-1 소스 코드 및 사용자 인터페이스로 인한 편향을 제거하기 위해 노력하였는가?	☐	☐	☐
	11-1a 데이터 접근 방식 구현과정 등 소스 코드에서의 편향 발생 가능성을 확인하였는가?	☐	☐	☐
	11-1b 사용자 인터페이스(user interface) 및 상호작용(interaction) 방식으로 인한 편향을 확인하였는가?	☐	☐	☐
	요구사항 12 인공지능 시스템의 안전 모드 구현 및 문제발생 알림 절차 수립			
	12-1 공격, 성능 저하 및 사회적 이슈 등의 문제 발생 시 대응 가능한 안전 모드를 적용하는가?	☐	☐	☐
	12-1a 문제 상황에 대한 예외 처리 정책이 마련되어 있는가?	☐	☐	☐
	12-1b 인공지능 시스템의 보안 강화를 위한 보안 기법을 적용하였는가?	☐	☐	☐
	12-1c 인공지능 시스템의 불확실성을 완화하기 위해 사람을 포함한 의사결정 방식을 고려하였는가?	☐	☐	☐
	12-1d 예상되는 사용자 오류에 대한 안내 및 대응을 제공하는가?	☐	☐	☐

생명주기	요구사항 및 체크리스트		Yes	No	N/A
4 시스템 구현	12-2	인공지능 시스템에서 문제가 발생할 경우, 시스템은 이를 운영자에게 전달하는 기능을 수행하는가?	☐	☐	☐
	12-2a	편견, 차별 등 윤리적 문제에 대한 알림 절차를 수립하였는가?	☐	☐	☐
	12-2b	시스템 성능 저하를 평가하기 위한 지표 및 절차를 설정하고 알림 절차를 수립하였는가?	☐	☐	☐
	요구사항 13	**인공지능 시스템의 설명에 대한 사용자의 이해도 제고**			
	13-1	인공지능 시스템 사용자의 특성(user characteristics)과 제약사항을 분석하였는가?	☐	☐	☐
	13-1a	사용자 특성에 따른 세부 고려사항을 분석하였는가?	☐	☐	☐
	13-2	사용자 특성에 따른 충분한 설명을 제공하는가?	☐	☐	☐
	13-2a	사용자 특성에 따른 설명 평가의 기준을 수립하였는가?	☐	☐	☐
	13-2b	사용자 특성에 맞는 적합한 용어를 사용하였는가?	☐	☐	☐
	13-2c	사용자의 구체적인 행동과 이해를 이끌어낼 수 있도록 명확한 표현을 사용하였는가?	☐	☐	☐
	13-2d	설명이 필요한 위치와 타이밍은 적절한가?	☐	☐	☐
	13-2e	사용자 경험을 평가할 수 있는 다양한 사용자 조사 기법을 활용하였는가?	☐	☐	☐
5 운영 및 모니터링	요구사항 14	**인공지능 시스템의 추적가능성 및 변경이력 확보**			
	14-1	인공지능 시스템의 의사결정에 대한 추적 방안을 수립하였는가?	☐	☐	☐
	14-1a	인공지능 시스템의 의사결정에 대한 기여도 추적 방안은 확보하였는가?	☐	☐	☐
	14-1b	인공지능 시스템의 의사결정 추적을 위한 로그 수집 기능을 구현하였는가?	☐	☐	☐
	14-1c	지속적인 사용자 경험 모니터링을 위해 사용자 로그를 수집 및 관리하고 있는가?	☐	☐	☐
	14-2	학습 데이터의 변경 이력을 확보하고, 데이터 변경이 미치는 영향을 관리하였는가?	☐	☐	☐
	14-2a	데이터 흐름 및 계보(lineage)를 추적하기 위한 조치를 마련하였는가?	☐	☐	☐
	14-2b	데이터 소스 변경에 대한 모니터링 방안을 확보하였는가?	☐	☐	☐
	14-2c	데이터 변경 시, 버전관리를 수행하였는가?	☐	☐	☐
	14-2d	데이터 변경 시, 이해관계자를 위한 정보를 제공하는가?	☐	☐	☐
	14-2e	신규 데이터 확보 시, 인공지능 모델의 성능평가를 재수행하였는가?	☐	☐	☐
	요구사항 15	**서비스 제공 범위 및 상호작용 대상에 대한 설명 제공**			
	15-1	인공지능 서비스의 올바른 사용을 유도하기 위한 설명을 제공하는가?	☐	☐	☐
	15-1a	서비스의 목적과 목표에 대한 설명을 제공하는가?	☐	☐	☐
	15-1b	서비스의 한계와 범위에 대한 설명을 제공하는가?	☐	☐	☐
	15-2	상호작용의 대상을 명확히 설명하는가?	☐	☐	☐
	15-2a	사용자가 인공지능과 상호작용하고 있다는 사실을 사용자에게 명확히 설명하였는가?	☐	☐	☐



PART 1

개요

01 안내서 발간 배경 및 목적

01 안내서 발간 배경 및 목적

의료 산업 분야에서는 정보통신기술ICT, Information and Communication Technology의 진보로 인해 첨단 기술과 융합한 새로운 개념의 의료기기가 등장하고 있다. 특히, 인공지능 기술과의 융합은 진료 기록 또는 의료기기에서 측정된 생체 측정 정보, 의료 영상, 유전 정보 등 환자의 데이터를 통해 학습한 인공지능 모델을 활용하여 의료진의 진료 과정에서 빠른 의사 결정을 보조하거나 신약 개발 기간 단축, 개인 맞춤형 건강 관리 서비스를 제공할 수 있는 효과적인 수단으로써 활용이 기대된다.

이러한 의료용 소프트웨어SaMD, Software as Medical Device, 인공지능을 활용한 진단·치료의 자동화, 디지털 치료 등 새로운 기술의 도입은 기존 의료기기 산업을 규정하던 영역 확장의 필요성을 불러일으켰으며, 미국과 중국 등 주요국은 혁신 제품의 개발을 촉진하는 다양한 지원책을 마련하고 있다. 국내에서도 2019년 「의료기기산업 육성 및 혁신의료기기 지원법」 제정을 통한 법적 근거 마련, '디지털 헬스케어 규제 자유특구' 지정 등을 통해 첨단 의료기기 산업이 발전하도록 지원 방안을 마련하고 있다.

이처럼 의료 분야에 인공지능 활용을 촉진하기 위한 다양한 노력이 수행되고 있지만, 한편으로는 인공지능 활용에 대한 우려의 시선도 있다. 이는 인공지능 알고리즘이 어떤 과정을 통해 결과를 도출하는지 알지 못하는 블랙박스black-box적인 한계와 함께 학습에 사용되는 의료 데이터에 포함된 민감한 개인 정보 이슈, 사고 발생 시 환자의 신체·재산에 직접적인 영향을 미치는 위험 등 다양한 사회적·윤리적 문제 발생의 가능성으로 인해 시스템을 신뢰할 수 없다는 이유로 발생된다.

국제 사회에서는 이러한 의료 인공지능의 부정적 영향은 완화하고, 긍정적 효과를 극대화할 수 있도록 다양한 대응 방안을 마련하고 있다. 세계보건기구WHO, World Health Organization에서는 《Generating evidence for artificial intelligence based medical devices: A framework for training validation and evaluation》('21)을 발간하여 의료기기로서 인공지능 기반 소프트웨어를 개발하는 프레임워크를 제시하고, 미국 식품의약국FDA, Food and Drug Administration에서는 《Artificial intelligence and machine learning (AI/ML) software as a medical device action plan》('21)을 발간하여 의료기기로서 소프트웨어의 안전성과 효과를 유지하고자 한다. 국내에서도 식품의약품안전처에서 〈인공지능 의료기기의 허가·심사 가이드라인〉('22) 발표해 이러한 위험을 완화하려는 노력을 수행 중이다.

이러한 노력은 비단 의료 분야뿐만 아니라 인공지능 산업 전반을 아우르는 문헌들에서도 언급되고 있다. 유럽연합EU, European Union의 《High-level expert group on artificial intelligence》('19)에서는 사람의 생명을 다루는 의료분야를 대상으로 신뢰할 수 있는 인공지능의 전반적 요구사항을 준수하길 요구하고 있다. 또한, 유네스코의 《Recommendation on the Ethics of Artificial Intelligence》('21)에서는 인공지능 기기를 활용한 의료 지원 시, 편견을 완화하기 위한 인간의 감독 보장, '도메인 전문가' 정의 시 전문가, 환자, 간병인 등의 포함, 의학적 모니터링 시 개인 정보 보호 등에 특별한 주의를 기울이도록 하고 있다.

그러나, 지금까지 나온 의료 분야의 인공지능 신뢰성 원칙, 정책, 표준 등은 주로 윤리 또는 프로세스 관점에서 추상적인 항목을 제시하거나, 임상시험, 인허가 관점에서 고려해야 하는 사항을 제시하고 있어 인공지능 시스템을 개발하는 실무 현장에서 활용하기는 어렵다. 특히 중소기업은 인력과 연구 개발 투자 여력이 제한적이어 직접 신뢰성 요구사항을 도출하고 검증체계를 마련하기 어려워 본 개발 안내서는 이러한 현실적인 문제점을 해결하고자 작성되었다. 미국과 유럽 등 선진국과 국제기구가 발표한 권고안, 가이드, 표준, 사례 및 연구 자료 등을 참고하였으며, 주요 항목에 대한 개발 요구사항과 검증항목으로 자율적 점검이 가능할 것이다.

02 의료 인공지능 신뢰성 동향

의료 분야 개발자 및 기획자 등 인공지능 제품·서비스 개발 실무자는 본 개발 안내서에 제시된 항목을 참고하여 최소한의 신뢰성을 확보하며, 의료 분야에서 신뢰성을 다룰 때 어떤 부분이 중요한지 이해할 수 있을 것이다. 특히, 개발 과정에서 전문 의료진과 협업이 수행되어야 하는 필요성과 고려 사항도 도움을 얻을 수 있을 것이다. 나아가 본 개발 안내서의 내용을 바탕으로 의료 인공지능 서비스에 적합한 요구사항과 검증방법을 마련함으로써 신뢰성 높은 의료 서비스를 개발할 수 있을 것이다. 본 개발 안내서를 통해 우리나라 의료 인공지능 관련 기업과 기관이 더욱 신뢰성 있는 의료 인공지능 기술 또는 시스템 및 서비스를 확보할 수 있는 기초 자료가 되길 희망한다.

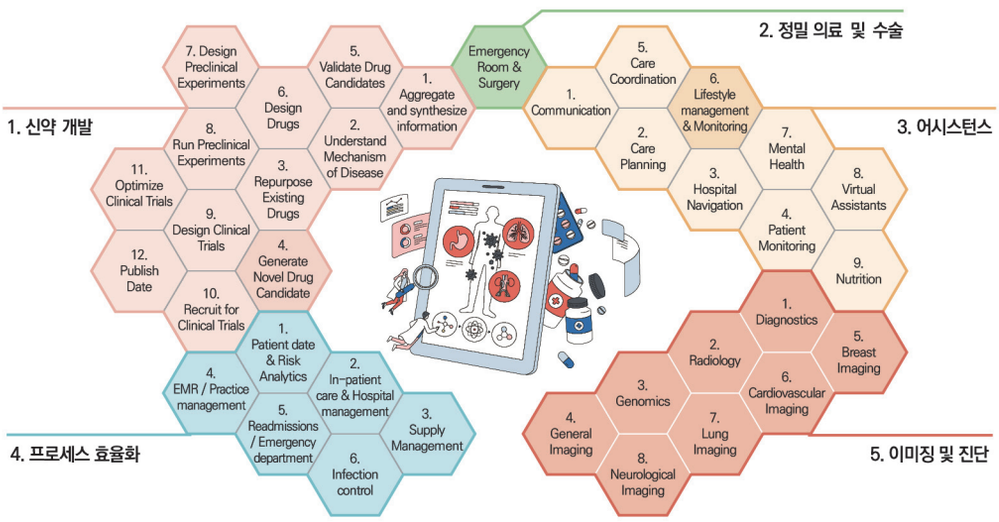
02 의료 인공지능 신뢰성 동향

현재 의료 인공지능의 활용에 대해 세계의 주요 국가와 표준 관련 기구, 기술 단체들은 인공지능의 신뢰성을 확보하고자 각기 상황에 맞는 방안을 제시하고 있다. 본 절에서는 의료 인공지능이 활용되는 영역과 인공지능을 활용하며 발생하는 문제점을 알아보고, 국내외에서 진행 중인 관련 정책 및 연구 동향을 알아보고자 한다.

2.1 의료 인공지능 활용 영역

의료 인공지능은 의료 데이터를 학습하고 특정 패턴을 인식해 진단 또는 예측하거나 환자에게 적합한 맞춤 치료 방법을 제공하는 기술로, 환자의 상태를 실시간 모니터링하는 등 의료 정보에 대한 접근성 향상에 기여한다. 또한, 의료 인공지능에 의해 표준적 치료 중심에서 진단·예방 등 맞춤형 치료로 의사의 지식 경험 기반 진료에서 데이터 기반의 정밀도 높은 진료로 의료 패러다임의 전환을 촉진하고 있다. 따라서, 초기의 의료 인공지능은 영상 의료 데이터를 인공지능 기술에 적용하여 질병 진단을 보조하는 수준이었으나, 최근에는 질병의 예후 예측, 치료·처방 등 의료 서비스 전 주기 과정에 다양하게 활용될 전망이다. 본 절에서는 심층학습deep learning 기반 인공지능 기술을 이용한 의료 인공지능의 기술과 활용 영역에 관해 설명하고자 한다. 해당 사례들과 그 외 문헌 및 사례들을 참고하면 의료 인공지능의 활용 영역은 크게 다섯 가지로 나누어 볼 수 있다. 해당 영역은 신약 개발, 정밀 의료 및 수술, 어시스턴스, 프로세스 효율화, 이미징 및 진단을 포함한다[1].

▼ 의료분야 인공지능 활용영역[1]



① **신약 개발**: 신약 개발의 과정은 일반적인 산업의 제품 개발 과정과 달리 물질의 검증과 전임상시험, 임상시험 등의 복잡한 단계를 거친다. 인공지능은 이러한 과정의 전 단계에 걸쳐 다양하게 활용할 수 있다. 신약 개발은 12가지로 활용이 세분화되는데, 정보 결합 및 합성, 질병 기전의 이해, 기존 약물의 용도 변경, 신약 후보 물질 생성, 신약 후보 물질 검증, 신약 설계, 전임상시험 설계, 전임상시험 수행, 임상시험 설계, 임상시험을 위한 환자 모집, 임상시험의 최적화 서비스, 데이터 출판 등이 이에 해당한다.

② **정밀 의료 및 수술**: 다양한 정보 혹은 빅데이터를 기반으로 개별 특성에 적합한 방식을 찾아내는 정밀 의료 영역에 활용이 가능하며, 수술 로봇을 통해 직접 수술을 진행하거나 수술 방식에 대한 안내 등의 역할 구분이 가능하다. 따라서 해당 영역에서는 주로 빅데이터나 로보틱스 분야 기술과 인공지능 기술이 결합한 형태로 활용된다.

③ **어시스턴스**: 어시스턴스는 주로 환자와 의료진 사이의 소통을 원활하게 하고, 건강 관리 전반이나 체계적인 치료를 관리하는 일련의 모든 활동을 포함한다. 예를 들어 커뮤니케이션 기능은 의사나 간호사를 통해 상담받기 힘든 다양한 정보 혹은 어려운 진료 기록들을 해석하여 환자와의 소통을 강화할 수 있다. 또한 환자의 모니터링이나 담당 의료진 및 병원 전반의 안내, 건강 관리 기능의 구현도 하여 의료진과 환자의 간극을 줄이고, 환자는 맞춤형 관리를 받을 수 있도록 하고, 의료인은 진단과 치료에 집중할 수 있는 환경을 제공해 준다. 어시스턴스는 활용이 9가지로 세분화되는데, 의료 관련 전문가와 소통 가능한 플랫폼, 건강 관리 계획을 수립해 주는 플랫폼, 병원 방문 전부터 진료실까지 실내 길 찾기 제공, 환자 건강 상태 실시간 모니터링, 환자의 건강관리를 위해 의료관계자들이 협력할 수 있도록 하는 플랫폼, 웨어러블 기기를 활용해 생활 습관을 관리해 주고 모니터링하는 플랫폼, 정신 건강 관리 관련 서비스 제공, 채팅을 통해 가상의 의료진이 증상을 진단, 식이 요법 및 생활 습관 관련 서비스 제공 등이 이에 해당한다.

④ **프로세스 효율화**: 프로세스의 효율화는 의료 경영과 행정 활동에 도움을 주는 것으로 건강 검진 자료, 환자의 치료 자료 등을 효율적으로 관리해 주는 영역을 의미한다. 특히, 단순한 재고 관리와 데이터 누적 등 회계 관리 업무뿐만 아니라 데이터 분석을 통해 환자의 위험을 관리하고, 위급상황 발생 시 동선을 통제 및 관리하거나 감염을 통제해 보건 위생을 관리하는 등 추론과 판단이 가능한 의료 경영의 영역을 모두 포함한다. 프로세스 효율화 분야의 활용은

6가지로 세분화되는데, 환자 데이터 및 위험 분석, 입원 환자의 건강 관리, 병원 관리, 재고 관리, 전자 의무 기록EMR, Electronic Medical Record 데이터와 실무 관리, 응급환자 발생 시 실시간 응급실 정보 제공, 감염 통제 등이 이에 해당한다.

⑤ **이미징 및 진단**: 이미징 및 진단 영역은 기존의 이미지 관련 의료기기에 판독과 진단 기능이 부가되는 것을 의미한다. 이미징 영역은 크게 진단방사선 이미지, 심혈관 진단, 유방 진단과 폐, 뇌 등 장비를 통해 판단하는 영역으로, 인공지능은 의료진이 육안으로 판독하기 힘든 영역의 표식을 찾아내는 정밀 판독과 기존의 판독 자료와 질환 관계 등을 분석하여 해당 이미지를 통해 질환을 진단하는 것으로 나눌 수 있다. 이미징 및 진단 분야의 활용은 8가지로 세분화할 수 있는데, 환자 데이터를 분석하여 진단 결과 제시, 방사선 이미지 분석을 통한 진단, 게놈 분석을 통한 진단, 일반적인 진단, 유방 진단, 심혈관진단, 폐 진단, 뇌 진단 등이 이에 해당한다.

2.2 의료 인공지능 이슈 사례

의료 인공지능의 개발은 현재 진행형이며 아직은 의료 기관에서 활발하게 도입되지 않은 것이 현실이다. 이러한 상황은 환자의 생명을 직접적으로 다루는 분야의 특성상 신뢰성이 확실하게 보장되지 못하면 새로운 기술이 쉽게 도입되기 어려운 문제에서 기인한다. 한국개발연구원KDI, Korea Development Institute에서 2021년에 수행한 '디지털 헬스케어에 대한 국민 인식 조사'에 따르면, AI 헬스케어에 대한 주요 우려 사항으로 '환자와 의사 간 정서적 교감을 어렵게 함(3.68/5)', '오작동으로 인한 의료 사고 위험이 큼(3.52/5)', '진단 결과를 신뢰할 수 없음(2.98/5)' 등이 제시되었다. 이는 의료 인공지능의 도입에 따른 효율성 향상은 긍정적으로 평가하나, 실제 활용에서 결과에 대한 믿음을 가지기 어렵다고 인식함을 알 수 있다.

AI 헬스케어	구분	각 의견에 대한 동의 정도를 5점 만점으로 점수화 한 평균값 (① 전혀 동의하지 않는다 ← ③ 보통이다 → ⑤ 매우 동의한다)
기대	진료 프로세스의 효율성 향상	3.91
	개인별 질병 예측 및 예방	3.78
	정밀한 진단 및 치료	3.66
우려	환자와 정서적 교감의 어려움	3.68
	오작동으로 인한 의료 사고 위험	3.52
	결과(진단)에 대한 신뢰성 부족	2.98

향후 의료 인공지능이 본격적으로 도입될 때 발생할 수 있는 주요 이슈로는 편향 문제, 보안 문제, 책임 소재 문제 등이 있다. 우선 편향 문제는 인종 다양성의 원인으로 해외에서 주로 언급되며, 이미 구축되었거나 현재 구축 중인 대부분의 의료 데이터가 선진국의 백인 남성 환자들을 위주로 구성되어 데이터셋 자체의 공평성이 확보되지 못한다는 데서 발생한다. 백인 데이터 위주로 학습된 인공지능 모델은 흑인이나 아시아인에게 임상 적용될 때 진단 정확도가 현저히 떨어질 수 있으며, 같은 맥락에서 남성 데이터 위주로 학습된 인공지능 모델은 여성 환자에 대한 임상적 유효성이 낮게 나타날 수 있다. 또한, 데이터 편향 이외에도 알고리즘 및 모델 편향의 문제도 발생할 수 있다. 예를 들어, 알고리즘 내 변수 선정에 오류가 생기면 진단이나 처방에 있어서 특정 부류의 환자들에 대한 특혜 또는 차별이 발생할 수 있다.

두 번째로, 보안 문제는 의료 데이터가 기본적으로 민감한 개인의 데이터이며, 관리 미비로 인해 유출될 시에는 심각한 사생활 침해를 유발한다는 점에서 제기된다. 인공지능 알고리즘의 성능 향상을 위해 수많은 개인에게서 의료 데이터를 수집해 활용하는데, 이때 데이터의 기본적 사용에 대한 환자의 동의를 받았더라도 서버에 저장되고 운용되는 방대한 양의 데이터가 제삼자에 의해 해킹되거나 유출될 가능성은 항상 존재한다. 따라서 의료 데이터를 관리하거나 사용하는 주체에 대한 법적 규제가 마련되어 있으며, 그러한 규정이 제대로 준수되는지 지속적인 감독을 요구한다.

마지막으로, 책임 소재 문제는 미래에 의료 인공지능이 일선 의료 기관에 도입되고 활용되기 전에 가장 핵심적으로 고려되어야 할 이슈로 꼽힌다. 아직은 의료 상황의 복잡성으로 인해 명확한 가이드라인이나 법적 규제가 없는 상태이다. 의료진이 의료 인공지능을 보조 또는 직접적으로 활용하여 의료 행위를 수행할 때 의료 사고가 발생한다면 각 상황의 책임 주체는 누구인지 불분명한 상황이다. 또한, 의료 인공지능의 오작동이나 응급 상황 대처 미숙으로 인해 문제가 발생했을 시에도 책임 소재가 모호할 수 있다. 이러한 문제를 해결하기 위해 의료계, 법조계 등의 학제 간 협업을 통해 관련 논의를 심화시킬 필요가 있으며, 모든 의료 사고 상황을 고려하여 책임 소재를 명확하게 분류하는 일은 향후 가장 중요한 쟁점이 될 것이라고 예상된다.

2.3 의료 인공지능 신뢰성 정책 및 연구 동향

환자의 생명을 다루고 민감한 의료 데이터를 취급하는 특성으로 인하여 의료 인공지능의 윤리적 사용에 대한 필요성이 지속해서 제기되고 있다. 국내를 비롯한 세계보건기구WHO, 미국, 일본 등 주요 국가들은 인공지능의 신뢰성 확보를 의료 인공지능의 사회적·산업적 수용과 발전의 전제 조건으로 정의하고 신뢰성 확보 정책을 추진 중이다. 또한 산업계 및 학계에서도 의료 인공지능의 신뢰성을 확보해 관련 기술을 개발하고, 의료 인공지능 산업을 발전시키는 노력이 활발하게 수행되고 있다. 세계보건기구는 《의료 분야 인공지능 윤리와 거버넌스 지침서》를 통해 의료 인공지능 사용 시 발생할 수 있는 윤리 및 거버넌스 문제를 파악하고, 올바른 사용을 위한 핵심 원칙인 자율성, 안전성, 투명성, 책무성, 형평성, 지속가능성 등을 제시하였다. 이 외에도 미국, 유럽, 일본 등 주요국에서는 의료 인공지능의 신뢰성을 확보하는 데 필요한 정책과 규제를 본격적으로 마련하고 있다. 특히 국내에서는 관련 가이드라인 제정 및 개정을 통해 의료기기 인허가 제도를 지속해서 정비하고, 임상시험 기준을 마련하고 있다. 한편, 민간 부문에서는 인공지능 신뢰성 확보를 위한 가이드라인 등의 연구를 수행하고, 자율적으로 인공지능의 신뢰성을 점검하고 확보할 수 있는 환경을 조성하고자 노력하고 있다.

▼ 주요국 의료 인공지능 신뢰성 관련 정책 동향

국가	주요 정책(연도)	특징
세계 보건기구WHO	• 의료 분야 인공지능 윤리와 거버넌스 지침서('21)	• 인공지능을 의료 분야에 윤리적으로 사용하기 위한 6가지 핵심 원칙(자율성, 안전성, 투명성, 책무성, 형평성, 지속가능성) 설명
미국	• 인공지능 기반 의료기기를 위한 규제 프레임워크('19) • 인공지능 및 기계학습을 기반으로 한 의료기기 소프트웨어 규제('21)	• 안전하고 효과적인 인공지능 기술이 적용 및 발전하도록 제품 수명 전주기 감독을 기반으로 하는 접근법 및 향후 규제 방향성 제시
유럽연합EU	• 신뢰할 수 있는 인공지능을 위한 윤리지침('19) • 인공지능 법안AI Act, Artificial Intelligence Act('21, '24 내 적용 예정) • 의료 분야의 인공지능('22)	• 임상 환경에서 인공지능을 활용하면서 발생할 수 있는 위험, 편향, 기타 사회 윤리적 영향 등을 분석하여 대응책 제시 및 유형별 규제 마련

국가	주요 정책(연도)	특징
일본	• 공적 인증제 도입('17) • 미래 투자 전략(건강수명연장)('17) • 인공지능 기반 소프트웨어 의료기기에 대한 규정 완화 계획('22)	• AI-인간 협업을 위한 안전성, 신뢰성을 바탕으로 보건 및 의료 간병 분야에서 ICT, AI, 빅데이터 등의 기술 혁신을 추진
한국	• 지능정보사회 윤리 가이드라인('18) • 보건의료 데이터·인공지능 혁신전략('21) • 인공지능 의료기기 국제 공통 가이드라인('22) • 보건의료 데이터 활용 가이드라인('22) • 인공지능 의료기기의 허가·심사 가이드라인('22)	• 보건 의료 데이터의 활용과 의료 인공지능 기술의 규제 적용 범위, 규정 등을 정의

▼ 국내외 주요 산·학·연 의료 인공지능 신뢰성 연구 동향

국가	기관명	활동 및 내용
한국	대한민국의학한림원	• 의료 인공지능 개발 및 사용 가이드라인('22)을 발간하여 모델 성능평가, 신뢰성과 안전성 확보를 위해 고려해야 하는 요건을 제시
	한국보건산업진흥원	• 보건의료빅데이터의 표준화와 품질평가 보고서('19) 및 의료 인공지능의 신뢰성과 안전성 보고서('21) 등을 발간하여 의료 데이터 및 인공지능 관련 연구 동향을 파악하고 의료 인공지능 활용을 위한 신뢰성 요건을 제시
싱가포르	싱가포르 국립대학	• 사회를 위한 보건의료 분야 인공지능 활용 가이드('21)를 발간하여 다양한 사용자 및 이해관계자 입장에서 의료 인공지능 활용 시 고려해야 할 논점들을 정리
미국	스탠포드 대학교	• 의료 인공지능 센터AIMI, Artificial Intelligence in Medicine and Imaging를 운영하며 방대한 분량의 의료용 오픈소스 데이터셋 공유

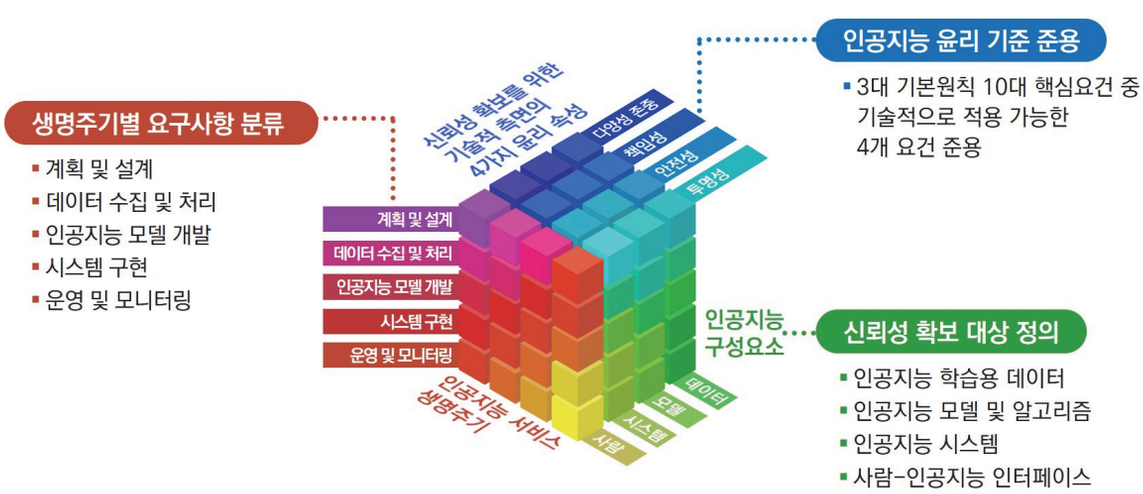
03 안내서 마련 과정

발간 배경에서 밝힌 바와 같이, 국내외 의료 분야에서 인공지능 활용 시 고려해야 할 여러 윤리 지침, 원칙 및 규제적 접근이 이루어지나, 의료 인공지능 개발 시 기술적 관점에서 고려 가능한 상세한 구현 방법론을 정리한 사례는 없었다. 따라서 본 안내서에서는 의료 인공지능 개발 현장에서 데이터 과학자, 모델 개발자, 전문 의료진 등 이해관계자들이 실무 관점에서 신뢰성 확보에 참고할 수 있는 지침서 성격의 자료를 만들고자 했다. 이를 위해, 2021년 1월부터 모든 산업 분야를 아우를 수 있는 일반 분야 안내서를 마련하기 시작했으며, 이를 기반으로 2022년에는 의료 분야에 특화된 안내서를 마련하였다. 의료 분야의 안내서 마련 과정에서 학계 및 산업계 전문가와 실무자들을 대상으로 의견 수렴을 진행하였다. 또한, 의료 인공지능 관련 서비스를 제공하는 기업과 협업해 안내서의 현장 적용과 컨설팅 공동 연구를 진행하여 케이스 스터디를 마련하고, 피드백을 받는 과정을 거쳐 실무 활용도를 높이고자 했다.

3.1 인공지능 신뢰성 프레임워크 적용

안내서 개발 과정 중 가장 우선적으로 신뢰성 확보를 위해 어떤 요소들이 실무적으로 고려되어야 하는지 탐색해보았고, 그 결과 세 가지 설계 요소를 도출하여 안내서에 반영하였다. 각 설계 요소는 요구사항과 검증항목 마련 시 모두 반영되었으며, 이러한 접근법을 아래 그림과 같이 매트릭스 형태로 체계화하여 '인공지능 신뢰성 프레임워크'로 정의하였다. 이 프레임워크는 의료 분야뿐만 아니라, 일반 분야 및 타 산업 분야에도 동일하게 적용된다.

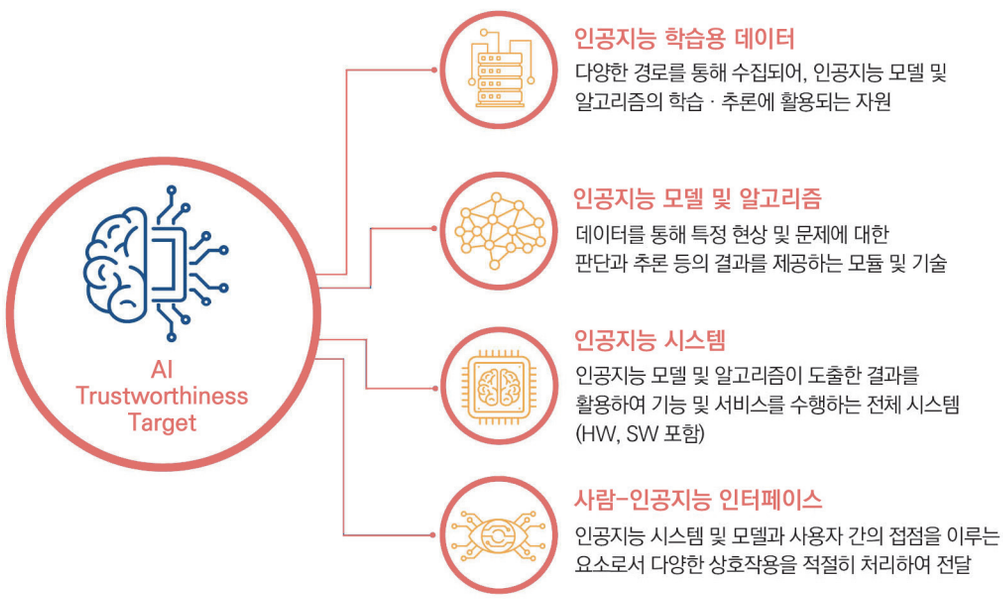
▼ 인공지능 신뢰성 프레임워크



03 안내서 마련 과정

첫 번째는 인공지능 구성 요소이다. 인공지능을 구성하는 4가지 요소는 학습과 추론 기능을 수행하는 인공지능 모델 및 알고리즘, 인공지능 학습용 데이터, 실제 기능을 구현할 시스템, 사용자와 상호작용하기 위한 인터페이스가 있다. 각 구성 요소들은 개별적으로 또는 통합적으로 인공지능 서비스의 생명주기에 따라 개발, 검증 및 운영된다. 따라서 구성 요소별 신뢰성 확보 방안을 고민하고, 각 요소에 따른 요구사항과 검증항목을 제시하고자 했다. 각 요소에 대한 신뢰성 확보 방안은 다음과 같다.

▼ 인공지능 서비스 구성 요소



인공지능 서비스 구성 요소	신뢰성 확보 방안
인공지능 학습용 데이터	인공지능 학습 및 추론 과정에 활용하는 데이터를 대상으로 편향성 등이 배제되었는지 검증
인공지능 모델 및 알고리즘	인공지능이 모델 및 알고리즘에 따라 안전한 결과를 도출하며, 이에 대한 설명이 가능한지, 악의적인 공격에 강건한지 등을 검증
인공지능 시스템	인공지능 모델 및 알고리즘이 적용된 전체 시스템을 대상으로 인공지능이 추론한 대로 작동하는지, 인공지능이 잘못 추론한 경우의 대책이 존재하는지 등을 검증
사람-인공지능 인터페이스	인공지능 시스템 사용자·운영자 등이 인공지능 시스템의 동작을 쉽게 이해할 수 있으며, 인공지능 오작동 시 사람에게 알리거나 제어권을 이양하는지 등을 검증

두 번째, 인공지능 서비스 생명주기는 첫 번째에서 살펴본 인공지능 서비스 구성 요소들을 구현하고 운영하는 일련의 절차를 말한다. 기존 소프트웨어 시스템에서 다루는 공학 프로세스나 생명주기와 비슷하나, 인공지능 특성상 데이터 처리 및 모델 개발 단계가 별도로 필요하며, 이외의 단계에서도 주요 활동에 대한 정의가 조금씩 달라진다. 현재 인공지능 혹은 인공지능 서비스의 생명주기는 다수의 문헌에서 6~8가지 단계로 구분한다. 대표적으로 OECD와 ISO/IEC에서 제시한 생명주기가 있는데, 본 안내서는 두 기구에서 제시한 생명주기를 대표성 있는 사례로 참고하여, 실무자들이 쉽게 활용할 수 있도록 각 생명주기 단계의 성격과 활동을 왜곡하지 않는 선에서 아래와 같이 5단계로 정리하였다.

▼ 인공지능 서비스 생명주기별 주요 활동

생명주기 단계	주요 활동
1. 계획 및 설계	- 인공지능 시스템 관리 감독 조직 및 방안 마련 - 인공지능 시스템 위험 요소 분석 및 대응 방안 마련
2. 데이터 수집 및 처리	- 데이터 품질 확보,데이터 사용자의 이해를 위한 정보 제공 방안 마련 - 데이터 라벨링 및 데이터셋 특성feature 문서화 - 인공지능 모델 구축을 위한 데이터셋 마련
3. 인공지능 모델 개발	- 비즈니스 목적에 따른 인공지능 모델 구현 - 구현된 인공지능 모델 확인 및 검증 - 인공지능 모델 튜닝, 데이터 분석, 추가로 필요한 데이터 수집 - 인공지능 모델에 대한 성능평가
4. 시스템 구현	- 문제 발생 대비 안전모드 구현 및 알림 절차 수립 - 인공지능 시스템 검증 및 사용자 설명에 대한 평가
5. 운영 및 모니터링	- 시스템 모니터링 및 인공지능 모델 재학습을 통한 성능 보장 - 모델 편향 탐지, 공평성, 설명가능성 등 시스템 신뢰성 모니터링 - 치명적 문제 발생 시 해결 방안 마련

인공지능 서비스의 생명주기 단계는 반복적·순환적인 성격을 띠지만, 반드시 순차적인 것은 아니다. 본 개발 안내서는 이해를 돕기 위해 1단계부터 5단계까지 순차적인 것처럼 설명했으나, 실제 데이터를 수집하고 가공하거나 모델을 개발, 운영하는 과정에서는 순서가 달라질 수 있다.

세 번째, 인공지능 신뢰성에 필요한 요건을 정의하고자 '인공지능 윤리기준'의 10대 핵심요건을 준용하여 기술적 관점에서 필요한 요구사항과 검증항목으로 '다양성 존중', '책임성', '안전성', '투명성'을 도출했다.

EC, OECD, IEEE 및 ISO/IEC 등의 국제기구는 인공지능 신뢰성의 하위 속성들을 세분화해 제시하고 있다. 특히, ISO/IEC 24028:2020 – Overview of trustworthiness in artificial intelligence는 신뢰성 확보에 필요한 고려 사항의 형태로 키워드를 제공한다. 여기에는 투명성, 통제가능성, 강건성, 복구성, 공평성, 안전성, 개인정보보호, 보안성 등이 포함되나, 키워드 간 관계나 신뢰성과의 연관성은 정의되지 않았다. 이처럼 관점에 따라 유사해 보이지만 조금씩 다른 용어들이 여러 문헌에서 제각각 달리 정의되고 있으며, 아직 합의된 속성 분류나 정의는 없는 상황이다. 이에, 앞서 언급한

03 안내서 마련 과정

EC, OECD, IEEE, ISO/IEC 등 여러 기구에서 제시한 속성과 키워드를 종합적으로 분석하고, 국내 학계·연구계·산업계 전문가의 의견을 수렴해 합의점을 모색했다. 이처럼 폭넓은 의견 공유 과정을 거쳐 인공지능 신뢰성 속성을 도출한 후, 이를 국가 인공지능 윤리기준의 10대 요건에 대응시켜서 기술적 측면에서 다룰 만한 요건을 최종 선정하였다. 각 요건에 대한 정의는 아래와 같다.

▼ 인공지능 신뢰성 요건

신뢰성 요건	정의
다양성 존중	인공지능이 특정 개인이나 그룹에 대한 차별적이고 편향된 관행을 학습하거나 결과를 출력하지 않으며, 인종·성별·연령 등과 같은 특성과 관계없이 모든 사람이 평등하게 인공지능 기술의 혜택을 받을 수 있는 것 - 관련 속성: 공평성·공정성fairness, 정당성justice - 관련 키워드: 편향bias, 차별discrimination, 편견prejudice, 다양성diversity, 평등equality - 국제표준(ISO/IEC TR 24027:2021 – Bias in AI systems and AI aided decision making)에서는 공평성을 정의하지 않는다. 공평성은 복잡하고 문화·세대·지역 및 정치적 견해에 따라 다양하여 사회적으로나 윤리적으로 일관되게 정의하기 힘들기 때문이다.
책임성	인공지능이 생명주기 전반에 걸쳐 추론 결과에 대한 책임을 보장하는 메커니즘이 마련된 것 - 관련 속성: 책무성responsibility, 감사가능성auditability, 답변가능성answerability - 관련 키워드: 책임liability - 국제표준(ISO/IEC TR 24028:2020)에서의 정의: 엔터티entity의 작업이 해당 엔터티에 대해 고유하게 추적될 수 있도록 하는 속성
안전성	인공지능이 인간의 생명·건강·재산 또는 환경을 해치지 않으며, 공격 및 보안 위협 등 다양한 위험에 대한 관리 대책이 마련되어 있는 것 - 관련 속성: 보안성security, 강건성·견고성robustness, 성능보장성reliability, 통제가능성·제어가능성controllability - 관련 키워드: 적대적 공격adversarial attack, 회복탄력성resilience, 프라이버시privacy - 국제표준(ISO/IEC TR 24028:2020)에서의 정의: 용인할 수 없는 위험risk으로부터의 자유
투명성	인공지능이 추론한 결과를 인간이 이해하고 추적할 수 있으며, 인공지능이 추론한 결과임을 알 수 있는 것 - 관련 속성: 설명가능성explainability, 이해가능성understandability, 추적가능성traceability, 해석가능성interpretability - 관련 키워드: 설명가능한 인공지능XAI, eXplainable AI, 이해도comprehensibility - 국제표준(ISO/IEC TR 29119-11:2020 – Guidelines on the testing of AI-based systems)에서의 정의: 시스템에 대한 적절한 정보가 관련 이해관계자에게 제공되는 시스템의 속성

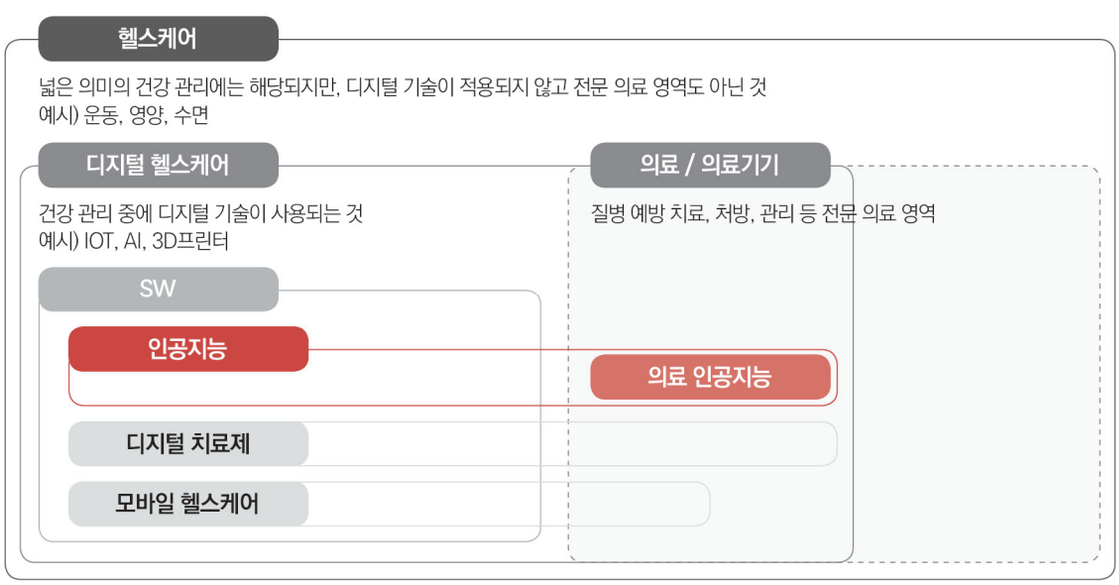
위와 같이 인공지능 신뢰성 확보를 위한 다양한 속성들이 있으며, 각 신뢰성 속성들에 대한 정의를 파악하는 것뿐만 아니라 신뢰성 속성 간의 상호 의존 관계 역시 중요하게 고려되어야 한다. 예를 들어, 인공지능 서비스에 대한 과도한 투명성 요구는 프라이버시 관련 위험을 초래할 수 있다. 또한, 설명가능성만으로는 투명성을 보장하기에 부족하지만, 설명가능성은 투명성을 확보하기 위한 중요한 요소 중 하나이다. 따라서, 인공지능 신뢰성 속성에 대한 충분한 이해를 바탕으로 인공지능 서비스를 제공하는 것이 중요하며, 해당 인공지능 서비스가 고려한 속성에 대해 적절하게 이행하는지 지속해서 검토해야 한다.

3.2 의료 분야 주요 고려사항 반영

본 안내서는 기술적 관점에서 상세한 방법론과 함께 개발 과정에서 전문 의료진과의 협업 필요성을 제시함으로써, 의료 인공지능 제품·서비스 개발 현장에서 실무자가 신뢰성 확보에 참고할 수 있는 실무 지침서 성격의 자료를 지향한다. 따라서, 본 안내서는 일반 분야에서 다루는 구성 요소 및 생명주기를 바탕으로, 인공지능 신뢰성 확보 시 고려되어야 할 요소들을 의료 분야에 특화하여 정리했다.

첫 번째, 본 안내서에서 신뢰성 확보 대상으로 다룰 의료 인공지능의 범위는 의료 인공지능 활용이 가능한 모든 범위를 포함하지는 않는다. 본 안내서에서 정의하는 의료와 의료 인공지능은 식품의약품안전처의 〈인공지능 의료기기 허가·심사 가이드라인〉을 참조하여 '질병을 진단·예측하는 임상 결정 지원CDS, Clinical Decision Supporting 소프트웨어나 의료 영상 진단 보조CAD, Computer-Aided Detection/Diagnosis 소프트웨어' 등 기계학습 가능 의료기기MLMD, Machine Learning-enabled Medical Devices를 대상으로 신뢰성을 확보하기 위한 요소를 다룬다. 즉, 본 안내서는 '진단'과 '처방' 활동에 직간접적으로 활용되는 인공지능을 대상으로 하며, 본문의 원활한 이해를 돕고자 필요시 헬스케어 범위의 사례를 일부 수록하였다.

▼ 의료 인공지능 범위



두 번째, 의료 인공지능 서비스를 구성하는 4가지 구성 요소는 아래와 같은 범위를 고려하였다. 본 안내서에서는 의료 인공지능의 주요 범위를 의료진의 진단 및 처방에 보조적으로 활용되는 의료용 소프트웨어나 기기로 설정하고, 의료 데이터 수집을 위해 환자가 사용할 수 있는 의료용 소프트웨어나 기기까지 포함하였다. '사람-인공지능 인터페이스'에서도 주 사용자인 의료진과 환자로 구분하여 접근 방향성을 제시하였다.

03 안내서 마련 과정

▼ 의료 분야 인공지능의 서비스 구성요소

구성요소	설명
학습용 데이터 (의료용 데이터, 학습 데이터셋)	진료 기록 또는 의료기기로 측정된 생체 측정 정보, 의료 영상, 유전 정보 등 질병을 진단 또는 관리하거나 예측하고자 사용되는 다양한 의료 데이터
모델 및 알고리즘 (인공지능)	의료진의 진단 및 처방에 보조적으로 활용되거나, 이를 위한 환자의 의료 데이터를 수집하여 처리하는 인공지능 모델 및 알고리즘
인공지능 시스템	의료 인공지능 또는 인공지능이 탑재된 의료기기
사람-인공지능 인터페이스	인공지능이 분석·판단·예측한 내용을 사용자(의료진 및 환자)에게 안내하는 수단 및 방법

세 번째, 의료 분야 인공지능 서비스 생명주기는 아래와 같은 활동을 추가로 고려하였다. 환자의 안전과 직결되는 의료 분야의 특성상 시스템 전반에 대한 관리 감독과 위험 대응 방안의 마련을 중시하였으며, 의료 분야에서 궁극적으로 요구되는 임상시험 및 의료기기 인허가와 연계될 수 있도록 관련 가이드라인 및 제도를 소개하였다. 특히, 모든 생명주기 단계에서 의료 기관 및 의료진과 긴밀한 협업이 필수적임을 강조하였다.

▼ 의료 인공지능 서비스 생명주기별 주요 활동

생명주기 단계	주요 활동
1. 계획 및 설계	- 의료 인공지능 시스템 관리 감독 조직 및 방안 마련 - 의료 인공지능 시스템 위험 요소 분석 및 대응 방안 마련 - 의료 인공지능 시스템 개발에 필요한 각종 사전 승인 절차 확인
2. 데이터 수집 및 처리	- 의료 데이터셋 품질 확보, 데이터 사용자의 이해를 돕는 정보 제공 - 의료 데이터 수집 및 처리를 위한 의료기관과 협업 - 의료 데이터 편향을 완화하는 관련 사례 참고 및 구체적인 대응 방안 마련
3. 인공지능 모델 개발	- 임상 목적에 따른 의료 인공지능 모델 구현 및 검증 - 의료 인공지능 모델에 대한 성능평가 및 임상시험 방안 마련 - 의료 인공지능 모델 편향을 완화하는 대응 방안 마련
4. 시스템 구현	- 의료 인공지능 시스템의 진단 및 추론 결과에 대한 설명가능성 제시 방안 마련 - 문제 발생 대비 안전모드 구현 및 알림 절차 수립 - 의료 인공지능 시스템 검증 및 사용자(의료진 및 환자) 설명에 대한 평가
5. 운영 및 모니터링	- 의료 인공지능 시스템 모니터링 및 모델 재학습을 통한 성능 보장 - 모델 편향 탐지, 공평성, 설명가능성 등 시스템 신뢰성 모니터링 방안 마련 - 치명적 문제 발생 시 해결 방안 마련

3.3 요구사항 및 검증항목 도출

다음 단계로 의료 분야의 인공지능과 관련한 구체적인 요구사항과 검증항목을 도출했다. 우선 표준화기구, 기술 단체, 국제기구, 주요 국가 정부에서 의료 분야 인공지능의 신뢰성을 확보하고자 발표한 정책, 권고안 그리고 표준을 기반으로 준수해야 할 기술적 요구사항을 도출하고 구체화하여 제시했다. 또한, 의료기기 위험관리 관련 표준인 ISO 14971:2019 – Application of risk management to medical devices 및 인공지능 신뢰성 관련 표준인 ISO/IEC TR 24028:2020에서 다루는 내용들을 주의 깊게 살펴보았다. 이와 함께 FDA 《Artificial Intelligence/Machine Learning(AI/ML)-Based Software as a Medical Device(SaMD) action plan》('21.1), WHO 《Ethics and governance of artificial intelligence for health: WHO Guidance》('21.6) 등 주요 해외 문헌과 보건복지부 〈보건의료 데이터 활용 가이드라인〉('22.5), 식품의약품안전처 〈인공지능 의료기기의 허가·심사 가이드라인〉('22.5), 〈인공지능 의료기기 임상시험방법 설계 가이드라인〉('22.7) 등 국내외에서 의료 인공지능의 신뢰성을 확보할 목적으로 발표된 문헌들을 검토했다. 이러한 과정을 거쳐 개발 안내서에 중요한 내용은 반영하고 중복된 내용은 제거하거나 축약했다. 해당 참고문헌은 다음과 같다.

▼ 인공지능 신뢰성 관련 주요 참고문헌

기관명	발간연월	권고 및 표준안 명
대한민국	2022.05	보건의료 데이터 활용 가이드라인
	2022.05	인공지능 의료기기의 허가·심사 가이드라인
	2022.07	인공지능 의료기기 임상시험방법 설계 가이드라인
미국	2021.01	Artificial Intelligence/Machine Learning(AI/ML)-Based Software as a Medical Device(SaMD) Action Plan
세계보건기구 (WHO)	2021.06	Ethics and Governance of Artificial Intelligence for Health: WHO Guidance
국제표준화기구 (ISO/IEC)	2018.02	ISO 31000:2018, Risk management – Principles and Guidelines
	2019.12	ISO 14971:2019, Medical devices – Application of risk management to medical devices
	2020.05	ISO/IEC TR 24028:2020, Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence
	2021.11	ISO/IEC TR 24027:2021, Information technology – Artificial Intelligence (AI) – Bias in AI systems and AI aided decision making
	2022.04	ISO/IEC 38507:2022, Information technology – Governance of IT – Governance implications of the use of artificial intelligence by organizations
	2023.02	ISO/IEC 23894:2023, Information technology – Artificial intelligence – Guidance on risk management

이를 통해 최종 도출한 요구사항은 아래 표와 같으며, 인공지능 윤리의 핵심 요건에 대응시킨 결과도 함께 표시했다.

▼ 인공지능 신뢰성 확보를 위한 기술적 요구사항과 윤리 요건 매칭 결과

요구사항	다양성 존중	책임성	안전성	투명성
요구사항 01 인공지능 시스템에 대한 위험관리 계획 및 수행		✓		✓
요구사항 02 인공지능 거버넌스 체계 구성	✓	✓	✓	✓
요구사항 03 인공지능 시스템의 신뢰성 테스트 계획 수립			✓	✓
요구사항 04 데이터의 활용을 위한 상세 정보 제공		✓		✓
요구사항 05 데이터 강건성 확보를 위한 이상 데이터 점검			✓	
요구사항 06 수집 및 가공된 학습 데이터의 편향 제거	✓	✓		✓
요구사항 07 오픈소스 라이브러리의 보안성 및 호환성 확보		✓	✓	
요구사항 08 인공지능 모델의 편향 제거	✓			
요구사항 09 인공지능 모델 공격에 대한 방어 대책 수립			✓	
요구사항 10 인공지능 모델 명세 및 추론 결과에 대한 설명 제공		✓		✓
요구사항 11 인공지능 시스템 구현 시 발생 가능한 편향 제거	✓			
요구사항 12 인공지능 시스템의 안전 모드 구현 및 문제발생 알림 절차 수립		✓	✓	✓
요구사항 13 인공지능 시스템의 설명에 대한 사용자의 이해도 제고				✓
요구사항 14 인공지능 시스템의 추적가능성 및 변경이력 확보		✓		✓
요구사항 15 서비스 제공 범위 및 상호작용 대상에 대한 설명 제공		✓		✓

3.4 현장 적용 및 전문가 의견 수렴

신뢰성 확보를 위한 요구사항을 도출한 후에는 각 항목을 기술적 타당성, 효용성 및 포괄성 등의 관점에서 검토한 후 고도화했다. 각각의 세부 검증항목이 요구사항에 해당하는 내용이 맞는지(타당성), 개발 현장에서 실무적으로 활용 가능한 내용인지(효용성), 검증을 위한 내용들이 과거부터 지금까지 연구 내용을 폭넓게 포함하는지(포괄성) 확인했다. 이를 위해 의료 인공지능 개발에 참여한 경험이 있는 전문 의료진, 의료 인공지능 분야 전문가가 참여하여 직접 검토하고 자문했으며, 다양한 검토 의견을 수렴하여 반영했다. 의료 인공지능 분야 전문가에는 기업의 기획자, 개발 프로젝트 리더, 교수, 의료진 등 산업계 및 학계의 연구자 등 분야를 가리지 않고 다양한 의견을 수렴하였다. 또한, 의료 인공지능 관련 서비스를 제공하는 기업과 협업해 안내서의 현장 적용과 컨설팅 공동 연구를 진행하여 케이스 스터디를 마련하고 피드백을 받는 과정을 거쳐 실무 활용도를 높이고자 했다.

04 안내서 활용 대상

본 안내서는 의료 분야의 인공지능 서비스를 구현하는 과정에 직간접으로 관련되거나 영향을 주는 모든 조직과 개인을 포함한 이해관계자가 참고할 수 있도록 작성되었다. 주요 대상은 특히 업무상 기술적 관점에서 신뢰성과 관련된 시스템 기획자, 시스템 엔지니어, 데이터 공급자, 데이터 과학자, 인공지능 모델 개발자 등이다. 또한, 전문 지식이 필요한 분야의 특성상 개발 및 운영 과정에서 전문 의료진과 협업이 필요하며, 이에 대한 대상은 의료 인공지능 서비스에 직간접적으로 관련된 전문의, 간호사, 이사진, 의료정보관리자 등이 있을 수 있다. 이들이 의료 분야 인공지능 생명주기의 각 단계마다 의료 분야의 인공지능 신뢰성을 확보하기 위해 검토해야 할 주요 요구사항은 다음과 같다.

▼ 의료 인공지능 생명주기 단계별 신뢰성 확보 요구사항

생명주기 단계	주요 행위자	주요 요구사항
1. 계획 및 설계	• 시스템 기획자 • 비즈니스 결정권자 • 품질 관리자 • 시스템 운영자 • 전문 의료진	– 인공지능 시스템 전체 생명주기에 걸친 신뢰성 확보 요구사항 검토 및 적용 방안 수립 – 의료 인공지능 시스템 생명주기별 위험 요소 식별 및 대응 방안 검토 – 의료기기 테스트와 인허가에 필요한 거버넌스 및 프로세스 검토
2. 데이터 수집 및 처리	• 데이터 과학자 • 데이터 공급자 • 인공지능 모델 개발자 • 전문 의료진	– 학습 데이터 확보 과정에서 발생할 수 있는 데이터 오류 및 편향에 대한 관리 방안 확보 – 학습 데이터 수집 시나리오 설계 및 시나리오별 정제 기준 마련
3. 인공지능 모델 개발	• 인공지능 모델 개발자 • 시스템 엔지니어 • 데이터 과학자 • 전문 의료진	– 학습 모델의 편향적인 출력이나 공격에 대한 대응 방안 수립 – 학습 모델의 출력에 대한 설명 및 해석 방안 제공
4. 시스템 구현	• 시스템 엔지니어 • 인공지능 모델 개발자 • 품질 관리자 • 전문 의료진	– 인공지능 시스템 개발 시 발생 가능한 편향이나 오류에 대한 대응책 마련 – 사용자(의료진 및 환자)별 안전 관리 및 오류 대응 방안 수립 – 인공지능 서비스가 도출한 결과에 대해 사용자 친화적인user-friendly 설명 제공
5. 운영 및 모니터링	• 시스템 엔지니어 • 시스템 운영자 • 인공지능 모델 개발자 • 비즈니스 결정권자 • 전문 의료진	– 인공지능 시스템 문제 발생 시 원인 추적을 통한 대응 방안 마련 – 서비스 목적 및 한계를 사용자에게 설명하여 오남용 예방

요구사항별로 대표 행위자와 협력 대상을 상세하게 대응시킨 결과는 아래와 같다. 요구사항별 협력 대상은 의료 분야의 인공지능 개발 안내서를 활용하는 서비스 및 기업 환경에 따라 상이할 수 있으므로 참고 사항으로 활용되길 바란다.

04 안내서 활용 대상

▼ 인공지능 신뢰성 확보를 위한 요구사항별 활용 권장 대상

요구사항	대표 행위자	협력 대상
요구사항 01 인공지능 시스템에 대한 위험관리 계획 및 수행	• 시스템 기획자	• 비즈니스 결정권자 • 시스템 엔지니어 • 시스템 운영자 • 인공지능 모델 개발자 • 전문 의료진
요구사항 02 인공지능 거버넌스 체계 구성	• 시스템 기획자	• 비즈니스 결정권자 • 시스템 운영자 • 전문 의료진
요구사항 03 인공지능 시스템의 신뢰성 테스트 계획 수립	• 품질 관리자	• 시스템 기획자 • 시스템 엔지니어 • 비즈니스 결정권자 • 전문 의료진
요구사항 04 데이터의 활용을 위한 상세 정보 제공	• 데이터 과학자	• 데이터 공급자 • 전문 의료진 • 인공지능 모델 개발자
요구사항 05 데이터 강건성 확보를 위한 이상 데이터 점검	• 데이터 과학자	• 데이터 공급자 • 인공지능 모델 개발자 • 전문 의료진
요구사항 06 수집 및 가공된 학습 데이터의 편향 제거	• 데이터 공급자	• 데이터 과학자 • 전문 의료진 • 인공지능 모델 개발자
요구사항 07 오픈소스 라이브러리의 보안성 및 호환성 확보	• 인공지능 모델 개발자	• 시스템 엔지니어
요구사항 08 인공지능 모델의 편향 제거	• 인공지능 모델 개발자	• 데이터 과학자 • 시스템 엔지니어
요구사항 09 인공지능 모델 공격에 대한 방어 대책 수립	• 인공지능 모델 개발자	• 시스템 엔지니어
요구사항 10 인공지능 모델 명세 및 추론 결과에 대한 설명 제공	• 인공지능 모델 개발자	• 데이터 과학자 • 시스템 엔지니어 • 시스템 운영자 • 전문 의료진
요구사항 11 인공지능 시스템 구현 시 발생 가능한 편향 제거	• 시스템 엔지니어	• 시스템 운영자 • 인공지능 모델 개발자
요구사항 12 인공지능 시스템의 안전 모드 구현 및 문제발생 알림 절차 수립	• 시스템 엔지니어	• 시스템 운영자 • 인공지능 모델 개발자 • 품질 관리자
요구사항 13 인공지능 시스템의 설명에 대한 사용자의 이해도 제고	• 시스템 엔지니어	• 시스템 기획자 • 시스템 운영자 • 인공지능 모델 개발자 • 비즈니스 결정권자 • 전문 의료진
요구사항 14 인공지능 시스템의 추적가능성 및 변경이력 확보	• 시스템 엔지니어	• 인공지능 모델 개발자 • 데이터 과학자
요구사항 15 서비스 제공 범위 및 상호작용 대상에 대한 설명 제공	• 시스템 엔지니어	• 시스템 기획자 • 시스템 운영자 • 인공지능 모델 개발자 • 비즈니스 결정권자 • 전문 의료진

05 안내서 활용 방법

본 안내서는 범용성을 갖추고자 인공지능의 신뢰성 관점에서 기술적 고려가 필요한 요구사항 및 검증항목을 포괄적으로 수록하였다. 따라서, 기업 내부의 기술 역량, 제품 특성 등을 고려하여 적절한 요구사항과 검증항목을 선택하여 적용하고, 기업에서 제공 중인 서비스의 환경에 맞게 신뢰성 확보를 위한 참고 자료로써 활용하길 바란다. 더불어, 인공지능 신뢰성을 확보하려면 기술적 측면 외에도 윤리, 개인정보보호 등 법적·제도적 측면도 함께 요구된다. 그러므로 본 안내서를 활용하기에 앞서 인공지능 윤리적 고려사항 점검을 위한 〈인공지능 윤리기준 실천을 위한 자율점검표〉와 개인정보보호의 준수 여부 점검을 위한 〈인공지능(AI) 개인정보보호 자율점검표〉 등을 선행적으로 검토할 것을 권고한다. 특히, 의료 인공지능 제품·서비스는 의료 현장에 사용되기 전 임상시험 및 인허가 절차가 반드시 수행되므로 식약처의 〈인공지능(AI) 의료기기 임상시험방법 설계 가이드라인〉, 〈인공지능 의료기기의 허가·심사 가이드라인〉을 검토해야 한다. 또한, 인공지능에 해당하는 속성뿐만 아니라 기존 소프트웨어 시스템에 적용되는 전통적 속성도 적용되었는지 확인이 필요하다. 따라서, 안내서에 기술된 내용 외에도 시스템 성능, 보안 등 품질 관점의 검증 절차도 반드시 병행되어야 할 것이다.

안내서는 다음과 같은 절차로 활용할 수 있다.

① **의료 인공지능 서비스 위험 영향 분석**: 의료 분야에서 인공지능 서비스의 기획 또는 도입을 고려하거나 운용 중인 인공지능 서비스의 점검을 원할 때, 서비스의 활용 목적과 범위, 사고 위험 및 사고 발생 시 사회적 파급도를 분석하여야 한다. 특히, 의료 인공지능은 잘못된 예측이나 작은 오류가 환자의 신체·생명·재산에 치명적인 피해를 줄 수 있다. 따라서 영향 분석 과정에서 전문 의료진의 참여는 필수적이며, 이 외에도 비즈니스 결정권자, 기획자, 개발자 및 시스템 운영자 등이 함께 논의에 참여하여 다양한 관점에서 검토를 수행할 것을 권장한다.

② **요구사항 선정**: '①'의 분석 내용을 토대로 개발 안내서 요구사항과 세부 요구사항 본문을 참고하여 인공지능 서비스에서 신뢰성 확보를 위해 필요한 요구사항을 선정한다. 의료 인공지능은 환자의 신체·생명·재산에 직접적인 영향을 미치므로, 가능한 모든 요구사항을 선정할 것을 권장한다. 이 과정에서 전문 의료진 및 개발자 등 요구사항별 활용 권장 대상(대표 행위자 및 협력 대상)을 협의해야 하며, 만약 불필요하다고 판단된 요구사항은 'N/A'를 표시하여 점검 대상에서 제외할 수 있다.

③ **자가 점검 수행**: '②'에서 선정한 요구사항은 세부 요구사항 및 검증항목 본문을 참고하여 충족 여부를 점검한다. 이 과정에서 본 개발 안내서의 본문에 소개된 기술 및 기법 예시를 참고하여 요구사항을 충족하지 못할 때 이를 해결할 만한 수단 또는 기술이 있는지 확인해 볼 것을 권고한다. 각 요구사항의 대표 행위자가 주도하여 협력 대상과 함께 검증항목의 충족 여부를 판단하는 데 필요한 절차서, 코드, 분석 자료 등의 관련 산출물을 확인하고, 테스트나 측정이 필요한 항목은 해당 활동을 수행한다. 검증항목에 따라 충족 여부를 정성적으로 평가할 수 있으나, 이는 '①'에서 분석한 서비스 영향 정도를 고려하여 대표 행위자와 협력 대상자가 협의하여 충족 여부를 판단할 수 있다.

PART 2

요구사항 및 검증항목

1. 계획 및 설계
2. 데이터 수집 및 처리
3. 인공지능 모델 개발
4. 시스템 구현
5. 운영 및 모니터링

목차

목차

목차

목차

01 계획 및 설계

책임성 투명성

요구사항 01 인공지능 시스템에 대한 위험관리 계획 및 수행

대표행위자 | 시스템 기획자 협력 대상 | 비즈니스 결정권자 시스템 엔지니어 시스템 운영자 인공지능 모델 개발자 전문 의료진

- 의료 인공지능 시스템 도입·운영 시 의료기기 위험관리 체계를 기반으로 예측 성능 저하, 보안 및 개인정보 이슈 등 위험 요소를 사전에 인식하고, 위험의 크기(심각성 및 파급효과)를 분석하여 대응 방안을 마련한다.

※ 의료기기 위험관리 : 의료기기의 위험을 분석, 평가, 통제 및 모니터링하는 업무에 대해 관리 정책, 절차 및 실무의 체계적 적용을 통하여, 위험을 허용이 가능한 수준으로 관리하는 안전 관리 시스템[2]

① 위해 요인 및 위해 상황 식별 → 의도된 사용 목적 파악, 위해 요인 식별, 위험산정
② 관련 위험산정 및 평가 → 위험 수용성 결정
③ 위험 통제 → 대안 분석 및 실행, 잔여 위험 평가, 위험/이득 분석
④ 통제 수단 효과성 감시 → 위험관리 프로세스 영향 및 적정성 평가

01-1 인공지능 시스템 생명주기에 걸쳐 나타날 수 있는 위험 요소를 분석하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

의료 인공지능 시스템에서 발생할 수 있는 다양한 위험 요소 분석 시 본 항목을 고려하여 만족 여부를 판단하십시오.

- 위험관리는 위험 인식identification, 위험 분석analysis, 위험 평가evaluation, 위험 대응treatment으로 구분한다. 신뢰성 확보를 위해 이러한 네 가지 활동을 생명주기 단계별로 지속적·반복적으로 수행함으로써 위험을 제거 및 방지하여야 한다. ISO 31000:2018 – Risk management – Principles and Guidelines에는 위험관리에 대한 개념 및 정의와 전체적인 흐름이 소개되어 있다.
- 다만, 인공지능의 신뢰성을 확보하는 과정에서 방해가 될 수 있는 위험 요소를 인식, 분석 및 평가하는 방법론은 기존의 소프트웨어 및 하드웨어 기반 시스템과는 상이할 수 있으므로 이 점을 고려해야 한다. ISO/IEC 24028:2020와 ISO/IEC 23894:2023 – Guidance on risk management에서는 인공지능 신뢰성 관점에서 살펴보아야 할 위험 요소의 분류가 제공되며, 의료기기 위험관리 관련 표준인 ISO 14971도 참고할 수 있다.
- 의료 분야의 경우 인공지능의 기술적 한계로 인해 나타나는 불확실성과 설명 불가능성 등의 문제, 거짓 양성false-positive이나 거짓 음성false-negative 등 진단 착오, 개인 정보나 생체정보의 유출 및 외부 공격에 관한 보안 문제 등으로 인명 피해가 발생할 수 있는 분야 특성상 세심한 위험 분석이 이루어질 필요가 있다.

- 국제의료기기당국자포럼IMDRF, International Medical Device Regulators Forum과 미국 식품의약국FDA 및 유럽연합EU에서는 의료용 소프트웨어SaMD의 위험 분류를 정의하고 있다. 해당 정의에는 의료 인공지능 시스템의 위험도를 등급별로 구분하여 정의하고 있으며, 각 등급에 따라 신고나 승인을 획득하는 절차가 상이하다[3].

01-1a 인공지능 시스템의 위험 요소를 도출하고 이의 파급효과를 파악하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템의 위험 요소는 소프트웨어 및 하드웨어 기반 시스템에서 발생할 수 있는 요소와는 다르다. 소프트웨어의 결함 및 오류, 하드웨어의 노후화 및 마모 등과 달리 데이터 기반 분석의 특성으로 나타날 수 있는 편향, 설명 미제공, 모델에 대한 공격 등의 위험 요소를 도출해야 한다.
- 이러한 위험 요소의 분류와 주요 내용은 ISO/IEC 24028:2020와 ISO/IEC 23894:2023에 제시되어 있으며, 이 외에도 01-2a의 '의료 인공지능 시스템의 생명주기 단계별 발생 가능 이슈와 대응방안 예시'를 참고하여 인공지능 생명주기 별로 발생할 수 있는 이슈들을 고려할 수 있다.
- 인공지능 알고리즘을 도입한 의료기기를 개발할 때는 다음과 같은 항목을 고려하고, 이로 인한 파급효과(예: 인명 피해)를 파악해야 한다. 위험은 잠재적인 피해의 원인이므로 의료기기 제품은 가능한 모든 위험을 식별해야 한다[4].
  - ✓ 인공지능 시스템을 통한 진단 보조의 임상 목적 부합 정도
  - ✓ 시스템의 예상 사용자 분석
  - ✓ 기존에 인공지능 시스템을 활용한 적이 없던 분야인지, 아니라면 신규 시스템과 차별점
  - ✓ ISO 14971:2019에 따른 의료기기 위험 식별 예
    - 위험 요인: 전자기 에너지Electromagnetic Energy(ESD)
    - 예측 가능한 사건의 순서
      1. 정전기로 대전된 환자가 인슐린 주입 펌프에 닿음
      2. ESD로 인해 인슐린 펌핑 및 펌프 경보가 실패함
      3. 환자에게 인슐린이 전달되지 않음
    - 위험 상황: 혈당 수치가 높은 환자에게 알려지지 않은 인슐린 투약 실패
    - 최악의 파급효과: 사망
- 진단이나 임상적 의사결정 보조 시에 의료진과 환자에게 악영향을 끼칠 가능성이 있다면 다음과 같은 관련 자료 및 정보의 제공이 필요할 수 있다.
  - ✓ 입력 데이터의 종류(CT 영상 등)와 출력 데이터 및 정보(병변 유무, 병변 위치, 병변 심각도, 진단 정확도 등)
  - ✓ 시스템의 민감도와 특이도 등 시스템이 야기할 수 있는 효과와 사전에 분석한 위험 요소[5]

- 인공지능 알고리즘을 활용한 의료기기 생명주기에서 추가로 고려가 필요한 위험 요소는 다음과 같다.
  - ✓ (임상시험) 임상시험방법 설계 가이드라인(참고)을 활용하여 위험 요소를 최소화한 설계 방법 준용
  - ✓ (현장 운용) 시스템을 직접 활용하는 의료진이나 진단 결과를 받는 환자 등 상황, 활용 및 수혜 대상에 따른 이해관계 충돌 가능성을 고려한 충분한 사전 분석 및 대책 마련
  - ✓ (보안) 사이버 보안 관련 가이드라인[6] 및 도구를 활용하여 개인 정보 유출, 랜섬웨어, 외부의 공격 등 예상되는 위험 요소와 그에 대한 관리 방안을 분석하여 제시
- 위와 같이 위험 요소를 도출한 이후 다양한 환경이나 상황에 따른 관리 방안과 파급효과를 분석해야 하며, 분석 결과를 준용하여 인공지능 시스템 생명주기 전반에 걸쳐 반복적으로 주기적인 추가 분석과 모니터링을 시행하여야 한다.

참고 SaMD 위험 분류 프레임워크[7]

- IMDRF에서 분류한 SaMD의 위험도는 I, II, III, IV의 등급으로 관리된다. '진단 시 SaMD가 제공하는 정보의 중요성'과 '의료 상황 또는 상태'를 종합적으로 고려하여 등급이 결정된다.
  - 만약 특정 의료기기의 사용처 및 용도가 다양하여 여러 분류 체계에 걸쳐 있을 때는 가장 높은 등급에 해당하는 위험관리 준용
  - 업데이트나 수정으로 인하여 SaMD가 변경되면 위험도 평가 재수행
  - 한 의료기기 내에 여러 의료 인공지능 시스템이 복합적으로 사용되면 각각에 대한 위험 분류를 독립적으로 관리
- 인공지능 시스템을 포함한 의료기기가 어떤 위험 분류 등급에 해당하는지 세부 기능을 평가한 후 관리하여야 한다. 다음은 SaMD 위험 분류 프레임워크 내 등급에 대한 예시이다.
  - Ⅰ. (위험 영향력 하) 심각하지 않거나 다소 심각한 질병이나 증상에 대한 임상 정보 제공, 임상 관리
  - Ⅱ. (위험 영향력 중) 심각하지 않거나 다소 심각하거나 위독한 질병이나 증상에 대한 임상 정보 제공, 임상 관리, 치료 또는 진단
  - Ⅲ. (위험 영향력 상) 다소 심각하거나 위독한 질병이나 증상에 대한 임상 관리, 치료 또는 진단
  - Ⅳ. (위험 영향력 최상) 위독한 질병이나 증상에 대한 치료 또는 진단

의료 상황 또는 상태	진단 시 SaMD가 제공하는 정보의 중요성		
	치료 또는 진단	임상 관리	임상 정보 제공
위독critical	Ⅳ	Ⅲ	Ⅱ
심각serious	Ⅲ	Ⅱ	Ⅰ
심각하지 않음non-serious	Ⅱ	Ⅰ	Ⅰ

참고 인공지능 의료기기 임상시험방법 설계 가이드라인(식품의약품안전처)[8]

- 시험 데이터셋 선정 방법 및 수집
  - 시험 데이터셋은 임상시험을 위해 수집하는 의료 데이터로서 진료 기록 또는 기존의 임상시험 과정 중 발생한 환자 데이터 등이 해당한다.
  - 시험 데이터셋은 품질, 수집 방법, 종류 등에 따라 후향적 임상시험의 임상 유효성 평가 결과에 영향을 미칠 수 있다.
  - 시험 데이터셋의 선정은 매우 중요한 과정으로, 임상시험 설계 시에는 선정 기준 및 제외기준을 명확히 하고, 수립된 기준에 따라 데이터를 선정하여야 한다.
  - 선정 기준 및 제외기준은 의료기기의 적응증에 따라 질환군, 질환의 빈도, 성별 등 목표 집단target population을 반영하여야 한다.
  - 또한 시험 데이터셋은 의료기기의 개발 과정 동안 사용된 학습 데이터셋과의 독립성이 유지되어야 하고, 편향이 발생하지 않도록 모집된 데이터 집단에서 무작위 배정하여 추출할 것을 권고한다.
  - 시험 데이터셋의 수는 대상 질병, 임상시험의 목적, 임상시험 평가 변수, 검정력 등을 고려하여 후향적 임상시험에 적합한 통계학적 방법에 따라 산출할 수 있다.
  - 이러한 시험 데이터셋의 수는 1차 유효성 평가 변수의 종류, 임상 결과의 기대치 정도, 비교 대상 및 방법 등에 따라 달라질 수 있으며, 적절한 통계가설을 설정한 후 피험자 데이터 수 산출 공식을 적용하여 산정하도록 한다.
- 피험자 동의
  - 피험자 동의는 의료기기 임상시험 실시기준 등 관련 규정(의료기기법 시행규칙 제24조 제1항 제4호)에 따라 임상시험을 시작하기 전, 피험자에게 동의를 받고 이를 문서화하여야 한다. 또한, 피험자 동의서 서식, 피험자 설명서 및 그 밖의 문서화 정보는 임상생명심사위원회IRB, Institutional Review Board의 승인을 얻어야 한다.
  - 그러나 기존의 의료용 데이터를 활용하는 후향적 임상시험에서는 피험자의 동의를 받는 것이 현실적으로 어렵거나 피험자에게 미치는 위험이 매우 낮을 수 있다.
  - 따라서 후향적 임상시험에서 피험자의 동의를 면제하는 것은 충분한 검토가 필요하며, 임상시험심사위원회의 승인 여부에 따라 피험자 동의 면제를 고려할 수 있다.

01-2 위험 요소를 제거 및 방지하거나 영향을 완화하기 위한 방안을 마련하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

01-1 에서 위험 요소를 분석한 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 01-1 에서 분석된 위험 요소별로 대응 방안을 마련하여야 한다. 이에 해당하는 대응 방안은 위험 요소의 원인을 제거하여 인명 피해 및 사고를 미연에 방지하거나, 사고로 인한 파급효과 및 부정적 영향을 최소화하는 수단 등이 이에 해당한다.
- 대응 방안이란, 구현 및 운영 방식 등의 절차, 소프트웨어 및 하드웨어 기능, 모델 학습 기법 및 전략 등 기술적으로 적용할 수 있는 모든 방법을 의미한다. 이에 대해 01-2a 의 대응 방안 예시를 참고할 수 있다. 인공지능을 구현하는 모든 이해관계자는 이를 고려하여 위험 요소에 대한 대응 방안을 마련하고, 위험이 제거 및 완화되었는지 확인하여야 한다.

• 인공지능 생명주기별로 발생 가능한 위험 요소에 대응하기 위한 기술적 방법을 면밀하게 분석하여, 인공지능 시스템의 신뢰성을 높이고 의료 사고의 확률을 줄일 수 있도록 설계하고 방안을 마련한다.

01-2a 위험 요소 제거 방안을 도출하고 파급효과가 감소하였는지 확인하였는가?

Yes ☐ No ☐ N/A ☐

• 위험 요소를 발생시킬 수 있는 구현 및 운영 방식, 소프트웨어 및 하드웨어 기능, 모델 학습 기법 및 전략 등의 기술적인 방법론을 도출하여야 한다. 이러한 방법론에 대한 분류와 개략적인 내용은 ISO/IEC 24028:2020에 제시되어 있다.

• 앞서 위험 요소를 분석하는 과정에서 위험 요소의 파급효과를 평가하였는데, 파급효과가 가장 큰 위험 요소를 우선순위로 대응 방안을 적용해야 하며, 위험의 파급효과가 클 때 인공지능 시스템의 판단 결과에 대한 사람의 개입을 고려하는 등의 위험 완화 방안을 적용해야 한다.

• 인공지능 생명주기 별로 발생 가능한 위험 요소에 대응하기 위한 기술적 방법을 면밀하게 분석하여, 인공지능 시스템의 신뢰성을 높이고 의료 사고의 확률을 줄일 수 있도록 설계하여야 한다. 대응 방안이 적용된 이후에는 파급효과를 재평가함으로써 위험 요소가 실제로 제거, 방지 혹은 이의 영향이 완화되었는지 확인하여야 한다. 다음은 도출한 위험 요소별 대응 방안의 예시이다.

  ✓ 진단 보조 인공지능의 학습 데이터 활용 시 발생 가능한 부작용 및 기타 의도하지 않은 효과
  - 정보 수집, 평가, 보고, 관리 등 일련의 위험 대응 체계를 마련하고 해당 대응 체계의 구체적인 절차와 방법을 명시

  ✓ 의료 윤리, 적법한 절차, 체계 대응
  - 연구윤리위원회나 기관감사위원회로부터의 승인 획득에 관한 설명을 사전에 제공
  - 각종 문제 발생 시에 데이터 활용 프로토콜을 수정하는 구체적 절차 마련

  ✓ 진단 보조 외, 의료진 대신 직접 처방을 내리거나 시술을 수행하는 인공지능 시스템의 위험
  - 문제 발생 시 환자에 대한 보상 제공 및 후속 치료 수행 등과 관련한 절차와 방법을 마련

  ✓ 보안 위협
  - 폐쇄 네트워크의 연결을 세분화하고 격리하여 관리
  - 기기 소프트웨어의 업데이트나 기타 자료 전송을 목적으로 네트워크에 연결할 때는 보안취약점을 반드시 확인하여 대응

참고 의료 인공지능 시스템의 생명주기 단계별 발생 가능 이슈와 대응 방안 예시

• 의료 인공지능 시스템 위험관리 프로세스가 구축되면 인공지능 생명주기 각 단계별로 발생할 수 있는 문제에 대하여 다양한 평가 및 측정 방법을 통해 적절한 실행과 함께 지속적인 테스트, 검토, 개선을 수행하여야 한다. 아래의 대응 방안 예시를 참고할 수 있으며, 더 자세한 내용은 ISO/IEC 24028:2020 또는 ISO14971:2019 등에 제시되어 있다.

프로세스	발생 이슈	대응 방안
1. 계획 및 설계	이해관계자의 시스템에 대한 이해도 결여	• 해당 의료 인공지능의 목적과 관련된 의료인이 참여하여 시스템에 대한 명확한 설명 제공(용어 표준화, 개념 정의 등)
	예측 가능한 위험 요소 존재	• 예측 가능한 오용, 사용 환경, 사용 중인 기술 등을 고려하여 위험을 허용 가능한 수준으로 감소 ※ 의도한 의학적 적응증, 환자 집단, 상호 작용하는 신체 부위나 조직의 형태, 사용자 프로파일[profile], 사용 환경 및 동작 원리 등의 정보 고려
2. 데이터 수집 및 처리	민감 데이터 오용 및 유출	• 의료 데이터를 안전하게 활용할 수 있도록 데이터심의위원회 DRB, Data Review Board를 구성하여 가명처리의 적정성, 제공 여부 및 방법 등을 마련
	데이터 스토리지[storage] 손상	• 의도적인 공격자의 데이터 스토리지 공격에 대한 대응책 마련
	데이터 편향(인지 편향, 사회적 편향 및 통계적 편향) 발생	• 인공지능 시스템 성능을 정의하고 데이터 편향 여부를 평가하여 편향이 발생할 수 있는 특정 그룹의 데이터 분포 측정
	데이터 소스 불투명성	• 전체 시스템 동작에 대한 외부 감사 등이 수행될 수 있도록 데이터 소스 공개
	인공지능이 비식별화 데이터를 자체 추론하여 데이터 재식별	• 비민감 데이터에서 민감 데이터를 추론하는 알고리즘을 적용하여 재식별화 가능한 데이터 선별
3. 인공지능 모델 개발	데이터 중독으로 인한 결과 조작	• 비정상적 학습 데이터 검출 필터링을 통해 데이터 중독으로 인한 손상 최소화
	인공지능 모델에 대한 적대적 공격	• 인공지능 모델이 잘못된 동작을 인지할 수 있도록 하여 적대적 공격에 대한 방어 전략 고려
	특성공학[feature engineering] 과정으로 인한 원본 데이터 손실	• 원시/원천 데이터 추적을 위해 데이터 변환 프로세스에 대한 문서화
	과대적합[overfitting] 및 과소적합[underfitting]	• 적절한 양의 예측 정보를 갖는 올바른 학습 데이터들을 선택하여 학습
	모델 업데이트로 인한 성능 저하	• 업데이트된 모델의 성능이 이전보다 저하되지 않고 특정 의료 목적에 적합한지 확인하는 검증 및 테스트 수행
	실제 조건에서의 불완전한 테스트	• 학습/테스트 데이터 외에 실제 데이터로 충분한 테스트 수행
4. 시스템 구현	시스템 소프트웨어 계층에 악성코드 존재	• 신뢰할 수 있는 실행 환경[TEE, Trusted Execution Environment]을 사용하여 수명주기 전반에 걸쳐 인공지능 모델 보호
	시스템 사양으로 인한 오류 장애 발생	• 의료 인공지능 시스템이 이기종* 환경에서 구현 가능하도록 정확한 시스템 사양을 정의(시스템을 테스트하여 검증 가능한 시스템 요구사항 제시) * 한 통신 인터페이스와 다른 통신매체 형식의 인터페이스로 통신 링크나 속도, 부호, 전송절차 등이 서로 다른 것

프로세스	발생 이슈	대응 방안
5. 운영 및 모니터링	인공지능 시스템 사용에 대한 책임 불분명	• 시스템 사용에 관한 법적·규제적·윤리적 의무를 이해하고 이행할 수 있도록 안내
	인공지능 시스템의 잠재적 위해성*으로 인한 피해 * 사람이 유해한 물질에 노출됐을 때 입는 피해 정도	• 의료 인공지능 시스템이 위해를 유발할 때, 관련 이해관계자에게 법적 책임을 부여하는 프로세스 구축 ※ EU과학신기술윤리위원회에서 자율 시스템 운영으로 인한 피해를 보상하는 프레임워크 확립 필요성 언급
	예측 불가능성 존재	• 의료 인공지능 시스템의 실패 가능 요인을 예측하거나, 예측할 수 없는 동작 등에 대한 안전장치fail-safe 등을 도입하여 환자의 안전 확보
	인공지능 생명주기 불투명	• 인공지능 생명주기 정보를 사용자 및 다른 외부 관계자에게 공개하여 상호 검증
	의료진 개입 프로세스 미비	• 최종 의사 결정자인 의료진이 의료 인공지능 시스템의 신뢰 수준을 재평가할 뿐만 아니라, 전반적인 시스템 운영을 개선할 수 있도록 피드백 기회 상시 제공

다양성 존중 | 책임성 | 안전성 | 투명성

요구사항 02 인공지능 거버넌스governance 체계 구성

대표행위자 | 시스템 기획자 협력 대상 | 비즈니스 결정권자 | 시스템 운영자 | 전문 의료진

- 의료 분야 인공지능 시스템은 확인되지 않은 오류로 피해나 사망 문제가 발생할 가능성을 잠재적으로 내포하고 있다. 이러한 인공지능 시스템의 영향과 결과를 예측하고 대비하기 위해 다양한 외부 전문가(예: 의사, 간호사, 데이터 과학자, 변호사, 임상 품질 및 학술 교수진 등)를 포함해 조직을 구성하는 것은 인공지능 신뢰성을 확보하는 데 중요한 요소이다. 따라서 인공지능 관련 법, 규제 및 정책, 관련 표준 및 지침을 정리하여 내부적으로 이행해야 할 윤리 원칙 및 규정을 수립하고, 이를 기준 삼아 감독하는 인공지능 거버넌스* 체계를 구성한다.

* 조직organization의 목적, 기회, 위험 및 이익을 파악하는 지속적인 프로세스

02-1 인공지능 거버넌스에 대한 지침 및 규정을 수립하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

의료 인공지능 시스템 및 모델 개발 시 발생 가능한 윤리 문제의 대응을 위해 본 항목을 고려하여 만족 여부를 판단하십시오.

- 의료 인공지능이 공중 보건에 긍정적인 영향을 미칠 수 있도록 의료 윤리와 안전을 비롯해 효과적인 배포에 대한 기본 방향을 제시하고 감시할 수 있는 거버넌스 체계를 구성하여 투명성을 확보하여야 한다. 즉, 건강 관리 윤리, 자비, 정의 및 비 악의 관련 기본 원칙을 담은 규정을 마련하고, 다양한 수준의 감사 시스템, 데이터, 지침 준수, 절차적 요건 충족 여부 등을 포함하여 감독하여야 한다.
- 또한, 인공지능 시스템 신뢰성을 확보하는 거버넌스 체계를 구성할 필요가 있다. 인공지능 시스템은 학습이나 추론 과정에서 윤리 및 지식 재산권IP, Intellectual Property 관련 문제, 보안 및 개인정보 이슈가 발생할 수 있기 때문이다. 이러한 위험 요소에 대비하기 위해 내부적으로 인공지능 거버넌스에 대한 지침 및 규정을 수립해야 한다.
- 내부적으로 수립해야 할 규정은 활용 측면에 따라 크게 두 가지로 구분하여 마련할 수 있다.
  - ✓ 첫째, 인공지능 관련 법, 규제, 정책, 표준 및 지침을 채택·정리하여 내부적으로 이행해야 할 지침 및 규정을 수립해야 한다.
  - ✓ 둘째, 인공지능 시스템 생명주기에 따른 조직의 역할과 책임을 명확하게 문서화해야 한다.

02-1a 내부적으로 준수해야 할 인공지능 거버넌스에 대한 지침 및 규정을 마련하였는가?

Yes ☐ No ☐ N/A ☐

- 윤리 원칙의 수립은 인공지능 거버넌스 체계에서 기본적으로 갖춰야 할 단계로(아래 참고), 인공지능과 관련된 법, 규제 및 정책을 이해한 후 내부적으로 윤리적 측면에서 이행해야 할 규정을 정의해야 한다. 즉, 인공지능과 관련된 위험을 인식하고 대비하기 위해 기업 성격에 맞는 핵심 가치를 선정하고 이와 관련된 표준 및 지침을 채택하여 내부 규정을 제공해야 한다.
- 의료 인공지능의 신뢰성 확보를 위한 윤리적 의무, 의료 데이터 보호 관련 법·제도, 의료 인공지능 알고리즘 개발 등에 관한 국제법과 국내법을 준수하여야 하며, 해당 여부를 자체적으로 감독할 수 있는 인공지능 거버넌스를 구축해야 한다.
- 인공지능 시스템의 신뢰성 확보를 위해서 인공지능 거버넌스 및 조직 전체의 업무, 역할, 의무 및 책임이 명확해야 한다. 이와 관련한 지침을 마련하고 조직 구성원에게 제공함으로써 자신의 역할과 책임을 인식할 수 있다.

참고 WHO, '의료 분야 인공지능 윤리와 거버넌스 지침서'

- 전 세계적으로 의료 인공지능에 대한 혁신이 가속화되면서 의료 인공지능 의사결정의 위험성, 알고리즘 편향 등에 대한 우려가 증가하고 있다. 이에 WHO는 의료 분야 인공지능 거버넌스가 준수해야 할 인공지능의 의료 윤리·안전에 대한 가이드라인을 제시하였다.
  ※ '의료 분야 인공지능 윤리와 거버넌스 지침서'는 의료 인공지능의 위험을 최소화하고 인공지능의 이점을 극대화하고자 WHO가 임명한 국제 전문가 위원회가 2년간 진행한 협의를 통해 만들어졌다.
- 의료진이 의료 인공지능과 기술적 의사결정에 대해 상호작용할 수 없으므로, 의료진이 의료 인공지능 시스템 의사 결정에 대하여 완전한 결정권을 행사해야 한다. 이를 위해서 WHO는 6가지 핵심 원칙을 제시한다.

6가지 핵심 원칙	설명
1. 자율성 보호	• 의료 인공지능 의사결정의 최종 결정권은 관련 의료진이며, 인공지능 의사결정 시 의료진의 의사결정은 보호되어야 한다.
2. 인간의 안전과 복지 증진	• 의료 인공지능 개발자는 시스템이 정상적으로 작동하는지 계속 모니터링해야 한다.
3. 투명성, 설명 가능성 보장	• 개발자는 의료 인공지능 시스템 설계에 대한 정보를 공유하여, 사용자가 이해할 수 있도록 투명성을 보장하여야 한다.
4. 책임 의식 함양	• 의료 인공지능 의사결정이 환자에게 위해를 가할 때, 기술적 문제 등에 따른 책임을 결정하는 메커니즘이 필요하다.
5. 포괄성 및 형평성 보장	• 의료 인공지능을 다국어로 사용할 수 있게 하여 인종적 편견 없이 다양한 데이터에 대해 학습하도록 한다.
6. 대응성 및 지속 가능한 AI 촉진	• 개발자는 의료 인공지능을 정기 업데이트해야 하며, 기관 및 기업은 의료 인공지능 시스템 문제 발생 시 해결 가능하여야 한다.

참고 위스콘신 보건 대학University of Wisconsin School of Medicine and Public Health의 거버넌스 지침 원칙[9]

- 내부 거버넌스에서 승인한 일련의 지침 원칙
  - ✓ 예측 모델(외부 공급업체 또는 내부 혁신 포함) 평가에는 UW Health 생산 데이터에 대한 성능 검증 및 적용하는 적절한 대상 레이블에 대한 임상의 검토가 포함됨
  - ✓ 모델 평가에는 통계적 측정(예: 민감도sensitivity, 특이도specificity, 양성 예측도positive predictive value) 및 관련 운영 및 건강 메트릭(예: 경보율, 걸음 수 대비 탐지율, 적절한 사용, 공정성, 비용 효율성 및 건강 결과에 대한 개입 효과)이 포함됨
  - ✓ 모델 출력은 임상 결정 지원의 5가지 권리(The right information, To the right person, In the right intervention format, Through the right channel, At the right time in the workflow)를 따름
  - ✓ 모델 모니터링(파일럿 또는 스케일 아웃)에는 통계적 측정, 운영 메트릭, 관련 결과 및 재평가 기준이 포함되며, 특히 절대 위험은 시간이 지남에 따라 변경될 수 있으므로 보정이 필요함
  - ✓ 건강 관리 윤리 자율성, 자비, 정의 및 비 악의의 기본 원칙은 모델 평가 및 검증의 모든 단계에 통합될 것임 인공지능 기반 도구로 해를 끼치지 않고, 생명 윤리 원칙이 거버넌스에 통합되도록 하는 것을 목표로 함

02-2 인공지능 거버넌스를 위한 조직을 구성하고 인력 구성에 대해 검토하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

02-1에 따라 인공지능 거버넌스에 대한 지침 및 규정을 마련한 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 02-1에서 언급했듯이, 의료 분야 인공지능 시스템은 확인되지 않은 오류로 피해 또는 사망 문제가 발생할 수 있다는 위험 요소가 존재한다. 따라서 다양한 위험 요소를 인식하고 관련 규정을 마련하여 이를 실행할 수 있도록 관리 및 감독하는 조직이 필요하다.
- 인공지능 거버넌스는 환자 안전과 책임뿐 아니라 임상의의 신뢰를 촉진하는 규정을 마련하고, 지침 준수 및 절차적 요건 충족 여부 등을 포함하여 감독하여야 한다. 또한, 이러한 조직은 각 담당자가 맡은 역할과 책임에 대해 충분히 인식하고 관련 역량을 갖춘 인력으로 구성할 필요가 있다.
- 단, 가능하다면 인공지능 거버넌스를 위한 조직은 외부 전문가(예: 의사, 간호사, 데이터 과학자, 임상 품질 및 학술 교수진 등)를 포함하여 구성할 필요가 있다. 외부 전문가들은 내부 조직에서 발생할 수 있는 편향된 시각을 보완하고, 집단 사고groupthink 등의 문제를 극복하는 데 도움을 주기 때문이다. 또한, 인공지능 시스템 운영에 필요한 방법론 및 전문지식을 외부 전문가를 통해 취득함으로써 의료 인공지능 사용 사례에 대해 민첩하고 혁신적인 접근 방식을 유지할 수 있다.

02-2a 인공지능 거버넌스를 위한 조직을 구성하였는가?

Yes ☐ No ☐ N/A ☐

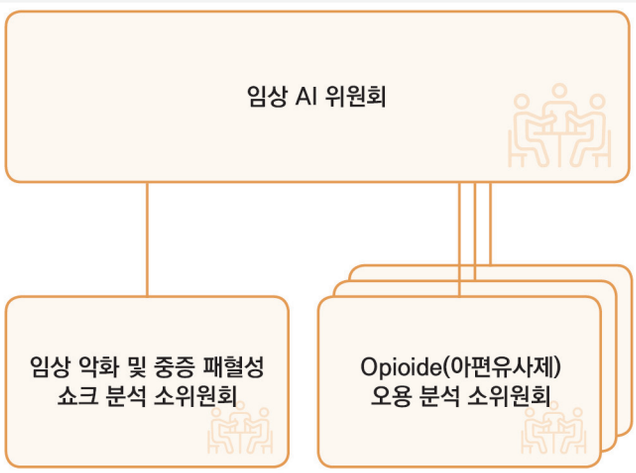
- 조직의 윤리 원칙 수립 후 이를 실행할 수 있도록 관리하는 것이 인공지능 거버넌스 체계의 목표이다. 내부 규정을 마련하고 이를 준수하는지 확인할 필요가 있다. 즉, 조직은 워크플로우 및 인공지능 시스템 구현을 포함하여 의료 시스템이 임상 치료에 영향을 미치는 인공지능 및 예측 솔루션 등을 감독한다. 여기에는 임상 치료(예: 환자 악화 또는 패혈증), 환자 접근 및 자원 할당(예: 입원 기간LOS, length of stay 예측, 입원 환자 용량 관리) 등을 목표로 하는 솔루션을 포함한다[9].
- 의료 인공지능 시스템의 거버넌스는 환자 안전과 책임뿐만 아니라, 채택을 개선하고 의미 있는 건강 결과를 촉진하도록 임상의의 신뢰를 촉진하는 것이 중요하다. 이에, 조직의 운영 원칙 또는 내부 규정을 마련할 때는 인공지능 모델의 유효성 및 사용자 수용 가능성 평가는 물론 효율성에 대한 지속적인 모니터링을 통한 안전한 배포에 이르기까지 인공지능 애플리케이션에 대한 감독을 수행할 수 있도록 고려한다[9].

참고 위스콘신 보건 대학University of Wisconsin School of Medicine and Public Health의 거버넌스 조직[9]

- 임상 인공지능 및 예측 분석 위원회의 분야별 구성원 및 역할 예시

- ✓ 의료 시스템을 구현하기 전에 솔루션의 평가 및 심사를 위한 게이트 역할을 제공함
- ✓ 의료 시스템 내 새로운 모델 개발 시 평가 및 심사를 위한 하위 작업 그룹을 위임하고 감독함
- ✓ 대학 및 의료 시스템의 기존 임상과 정보 구조 등을 비즈니스 파트너에게 보고함
- ✓ 하위 위원회 그룹(아래 그림)에 인공지능의 모든 임상 사용에 대한 가시성을 제공함

• 임상 인공지능 및 예측 분석 위원회 그룹 및 역할



사용 사례별로 승인된 소위원회

✓ 기관 차원 위원회(임상 AI 위원회)
- "예측 모델" 등 핵심 용어의 정의와 지침 원칙을 정의하고 설정함
- 인공지능 응용 프로그램에 대해 알고리즘 작업 그룹의 분석 등 세부 요청이 접수되면 특정 신청서의 범위를 가진 "알고리즘 소위원회"(예: 아편 유사제 오용 분석 소위원회)를 위임함

✓ 알고리즘 소위원회(임상 악화 및 중증 패혈성 쇼크 분석 소위원회 등)
- 기관 차원 위원회에서 확인한 지침 원칙, 내부 규정 등을 따름
- 특정 사용 사례에 대한 알고리즘을 평가 시, 위 지침 및 원칙을 적용하고 임상 AI 위원회 및 인공지능 예측 분석 위원회에 결과를 보고함

02-2b 인공지능 거버넌스를 위한 조직은 충분히 훈련된 인력으로 구성하였는가?

Yes ☐ No ☐ N/A ☐

• 인공지능 거버넌스 담당 조직은 자신이 맡은 역할과 책임에 대해 충분히 인식한 인력으로 구성해야 한다. 이들은 인공지능 생명주기에 걸친 모든 프로세스의 중심적인 역할로서, 이는 담당자가 이를 충분히 인식한 후 책임지고 관리해야 인공지능 시스템의 신뢰성을 확보할 수 있기 때문이다.

• 의료 분야 거버넌스 조직의 구성원은 거버넌스 기능(인공지능 시스템 사용 사례에 대해 민첩하고 혁신적인 접근 방식 유지)을 수행하는데 필요한 여러 분야의 전문가로 이루어져야 한다. 다음은 의료 분야 거버넌스 조직에 포함하여 구성할 수 있는 주요 분야와 대표되는 전문가 예시이다.

✓ 주요 분야 : 임상 주제 전문 지식, 데이터 과학, 정보학, 정보 기술, 인공지능 기술, 임상 운영, 생명윤리, 법률 자문 및 검토, 인적 요소 또는 디자인 사고 등

✓ 전문가 역할: 의사, 간호사, 데이터 과학자, 분석 전문가, 정보 서비스, 인공지능 시스템 개발자, 변호사, 임상 품질 및 학술 교수진 등

- 또한, 인공지능 거버넌스 담당 조직은 각기 다른 배경과 전문지식을 기반으로 충분히 숙련된 인력으로 구성해야 한다. 특히, 규정을 마련하는 역할을 맡은 담당자는 인공지능 윤리 및 신뢰성 분야의 원칙, 가이드라인, 표준 등에 대한 폭넓은 전문지식을 갖춰야 하며, 이를 적절히 해석하여 조직 업무에 적용하기 위한 기술력과 타 업무 담당자와의 의사소통 역량이 필요하다. 또한, 정의된 규정을 실행하고 관리하기 위해 각 담당자에게 관련 교육을 제공하여 충분히 훈련해야 한다.

02-3 인공지능 거버넌스 체계가 올바로 이행되고 있는지 감독하고 있는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

02-1에 따라 인공지능 거버넌스에 대한 지침 및 규정을 마련한 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 인공지능 거버넌스 체계를 운영하는 주체는 운영 결과에 대한 책임을 져야 하고, 이 책임은 위임할 수 없다. 따라서 인공지능 거버넌스 운영 담당자는 조직이 내부 지침 및 규정을 준수하는지에 대해 감독해야 한다.
- 또한, 의료 분야 인공지능 시스템의 예측 성능 등 위험관리를 위한 모니터링 프로세스 등을 적용하여 시스템 비 배포, 사용 중지 등 관리를 반복해 의료 인공지능 솔루션의 성능을 보장해야 한다.

02-3a 인공지능 거버넌스에 대한 내부 지침 및 규정 준수 여부를 감독하고 있는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 거버넌스 담당자는 인공지능 시스템의 생명주기에 따라 조직이 내부 규정을 준수함을 확인 및 감독해야 한다. 또한, 신뢰성 있는 인공지능 시스템을 목표로 적절히 관리 및 통제됨을 관련 이해관계자에게 입증해야 한다. 다음은 인공지능 거버넌스의 내부 지침 및 규정에 따른 감독 사례이다[9].
  - ✓ 감독 대상 인공지능 구축 목적: 인공지능 모델을 사용하여 병원의 응급환자 수가 많은 날을 예측
  - ✓ 활용: 응급환자 수 예측 결과에 따라 추가 의사 직원을 호출하기로 의사결정
  - ✓ 감독 결과: 초기에는 유용했으나, 평균 응급실 일일 방문자 수가 더 늘어나고 호출 교대가 가능한 의료진의 인력이 증가함에 따라 예측 모델의 필요성이 저하되어 폐기 결정
- 특히, 인공지능 시스템 위험관리와 관련된 내부 규정을 이행하는지 감독함으로써 인공지능 시스템의 잠재적 위험에서 조직 및 이해관계자를 보호하고 조직의 역량을 향상할 수 있다.
- 따라서 인공지능 거버넌스 체계에서 감독을 담당하는 조직은 인공지능 시스템에 대한 이해를 바탕으로 역할에 대한 책임 및 권한을 명확히 인지하여 인공지능 시스템 생명주기에 걸쳐 모든 규정이 이행되는지 감독해야 한다.

02-4 인공지능 거버넌스 조직이 신규 및 기존 시스템의 차이점을 분석하였는가?

Yes No N/A
☐ ☐ ☐

해당여부 판단

인공지능 도입을 위해 기존 의료 인공지능 시스템 및 서비스를 분석한 경우 본 검증항목을 고려하여 만족 여부를 판단하십시오.

- 도입 및 개발을 계획 중인 인공지능 시스템이 기존에 공개된 의료 시스템과 활용 대상 및 역할 측면에서 유사한지 고려하고, 사전에 알려진 해당 의료기기의 위험성 항목을 확인한 결과를 기반으로 시스템을 계획 및 설계해야 한다.
- 또한, 기존에 공개된 의료기기가 리콜recall된 사례를 비교 분석하여, 발생 가능한 잠재적 위험성을 최대한 분석 및 해결해 신뢰할 수 있는 의료기기를 개발할 수 있도록 한다.

참고 기존 의료기기 리콜을 통해 알려진 위험성 분석 사례

- Medtronic사의 MiniMed Paradigm, MiniMed 508 인슐린 펌프와 함께 사용되는 원격 컨트롤러의 잠재적인 사이버 보안 위험으로 인한 리콜 사례[10]
  - ✓ 리콜 등급
    - Class I 리콜
    - 가장 심각한 유형의 리콜
  - ✓ 리콜 사유
    - 잠재적인 사이버 보안 위험으로 인해 권한이 없는 사람(환자, 환자 간병인 또는 의료 제공자가 아닌 사람)이 무선 통신을 기록하고 재생할 수 있음
    - 승인되지 않은 사람이 특수 장비를 사용하여 환자에게 인슐린을 과다 전달하여 저혈당(저혈당증)을 유발하거나, 인슐린 전달을 중단하여 고혈당 및 당뇨병성 케톤산증, 심지어 사망까지 초래하도록 지시할 수 있음

02-4a 신규 시스템과 기존 동일 목적의 의료 시스템을 비교하여 안전성과 효율성 확보가 가능한지 분석하였는가?

Yes No N/A
☐ ☐ ☐

- 신규 의료 인공지능 시스템을 시판하기 전 기존에 출시된 기기와 비교하여 장치에서 감지되는 위험 수준과 인지된 위험 수준에 따라 필요한 데이터 요구사항 및 제어사항을 분석할 필요가 있다.
- 비교 분석을 통해 신규 의료 인공지능 장치가 기존의 동일 목적으로 시장에 출시된 장치와 동등한 안정성과 효과성을 입증해야 하며, 객관적인 기준, 근거, 검증을 기반으로 의료 인공지능 서비스의 안전성을 보장하는 방향으로 비교 분석을 추진하여야 한다.

참고 FDA 510(k) 승인 프로세스[11]

- FDA 510(k) 승인 프로세스는 의료기기 시판 전에 이미 시장에 출시된 기기와 실질적 동등성을 기반으로 안전하고 효과적임을 입증하는 데 사용되는 절차이다.
- 동등성을 결정하려면 ❶장치가 시판된 것과 동일한 용도로 사용됨을 입증하고, ❷기기와 선행 기기 간의 기술적 차이가 기기의 유효성과 안전성에 부정적인 영향을 미치지 않는다는 것을 증명해야 한다.

안전성 투명성

요구사항 03 인공지능 시스템의 신뢰성 테스트 계획 수립

대표행위자 | 품질 관리자 협력 대상 | 시스템 기획자 시스템 엔지니어 비즈니스 결정권자 전문 의료진

- 전통적인 소프트웨어와 달리, 인공지능은 출력 결과에 대한 불확실성uncertainty을 내포한다. 이러한 인공지능의 불확실성을 줄이는 것은 안전성과 같은 신뢰성 확보에 중요한 요소이다. 따라서 소프트웨어의 품질을 확인하는 테스트 외에도 인공지능 의료기기 시스템의 신뢰성 확인을 위한 테스트가 추가 요구된다. 테스트를 위해서는 인공지능 의료기기 시스템의 운영 환경을 고려한 계획 수립이 필요하며, 계획에 따라 생명주기 전 단계에서 정기적·지속적 테스트를 수행한다.

  * 인공지능에 해당하는 속성뿐만 아니라 기존 소프트웨어 시스템에 적용되는 전통적 속성도 적용되었는지 확인이 필요하다. 따라서, 본 요구사항에 기술된 내용 외에도 시스템 성능, 보안 등 품질 관점의 검증 절차도 반드시 병행되어야 할 것이다.

03-1 인공지능 시스템의 특성을 고려한 테스트 환경을 설계하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

인공지능 시스템의 위험 분석 결과에 따라 사고 발생 가능성 및 오동작으로 인한 파급력이 클 때 본 항목을 고려하여 만족 여부를 판단하십시오.

- 의료 인공지능 시스템은 진단 및 임상 관리 개선, 의약품 개발, 질병 추적 및 대응, 보건의료 시스템 고도화 등 활용도가 매우 높지만, 개인의 생명 문제와 직결되어 인공지능 의사결정에 따른 신뢰성 확보 문제를 항상 내포한다. 따라서 의료 인공지능 시스템의 안전성과 투명성을 확보하는 테스트와 검증 계획 수립이 필요하다.
- 유네스코의 인공지능 윤리 권고에서는 인권에 대한 잠재적 위협 가능성이 있다고 식별된 인공지능 시스템은 출시 전 이해관계자에 의해 윤리 영향 평가의 하나로 광범위한 테스트를 거쳐야 하며, 필요하다면 실제 상황과 동일 조건에서 테스트를 진행하여야 한다고 권고한다.
- 정확한 테스트를 위해서는 실제 환경 테스트를 수행하는 것이 적절하지만, 테스트는 합리적인 시간 및 비용 범위 내에서 수행되어야 하므로 운영 조건이 매우 복잡한 시스템이라면 실제 환경 테스트가 적절하지 않을 수 있다. 또한, 실제 환자와 물리적으로 상호작용하는 인공지능에 실제 환경 테스트를 적용한다면 위험한 상황이 발생할 우려가 있는데, 이 때 가상테스트를 고려할 수 있다.
- 따라서, 의료 인공지능 시스템의 특징을 고려하여 적절한 테스트 환경을 결정한 후 테스트 환경을 설계하는 것이 필요하다. 테스트 환경 설계 시 고려해야 할 사항의 예시는 아래와 같다.

✓ 인공지능 시스템의 운영 환경이 복잡하고 다양한 이해관계자가 참여하는가?
✓ 인권에 대한 잠재적 위협 가능성이 우려되는 시스템인가?
✓ 테스트는 합리적인 시간 및 비용 범위 내에서 수행 가능한가?
✓ 임상시험 전 검증 시 구체적인 테스트가 필요한 항목은 무엇인가?

03-1a 테스트 환경 결정 시 인공지능 시스템의 운영환경을 고려하였는가?

Yes ☐ No ☐ N/A ☐

- 테스트 환경 설정 시 실제 임상시험 절차를 참고하고, 가능한 참고한 대로 신뢰성 확보를 위한 테스트 계획을 수립하는 것이 바람직하다. 임상시험 시에 사용하는 테스트용 데이터셋을 직접 확보할 때는 의료기관과 긴밀한 협의가 필요하다. 때에 따라 테스트용 데이터셋 확보나 피험자 동의 등을 위해 임상시험 절차 수준에 따라 임상생명심사위원회IRB의 승인이 필요할 수 있다.
- 의료 인공지능 시스템의 초기 테스트 이후에도 실제 임상 환경에 출시되기 전에 결함을 식별하도록 임상시험이 수반되어야 한다. 구체적인 임상시험방법 설계 시에는 식품의약품안전처의 인공지능 의료기기 임상시험방법 설계 가이드라인을 참고할 수 있다. 의료 분야에서는 민감도, 특이도, 양성 예측도, 음성 예측도negative predictive value, ROCreceiver operating characteristic Curve, AUCarea under the curve 등 성능 유효성 검증항목을 참조 표준reference standard과 비교함으로써 임상시험을 설계할 수 있다. 예를 들어, 진단 보조 성능을 검증하고자 다수의 병원에서 다양한 환자를 대상으로 한 가지 이상의 의료기기로 촬영한 영상 데이터를 확보 후 시스템을 테스트하여 정상 동작 유무를 확인하도록 설계할 수 있다.

참고 테스트 환경 결정 시 반영할 수 있는 식약처의 의료기기 임상 유효성 평가 항목(일부 발췌)

[표본 데이터 선정]
- '표본 데이터' 선정은 매우 중요한 과정으로서, 임상시험 설계 시에는 선정기준 및 제외기준을 명확히 하고, 수립된 기준에 따라 데이터를 선정하여야 한다.
- 선정기준 및 제외기준은 의료기기의 적응증에 따라 질환군, 질환의 빈도, 성별 등 목표 집단을 반영하여야 한다.
- 또한, '표본 데이터'는 의료기기의 개발과정 동안 사용된 학습 데이터와 독립성이 유지되어야 하고, 편향이 발생하지 않도록 모집된 데이터 집단에서 무작위로 배정해 추출할 것을 권고한다.
- '표본 데이터'의 수는 대상 질병, 임상시험의 목적, 임상시험 평가변수, 검정력 등을 고려하여 후향적 임상시험에 적합한 통계학적 방법에 따라 산출할 수 있다.
- 이러한 '표본 데이터'의 수는 1차 유효성 평가변수의 종류, 임상 결과의 기대치 정도, 비교 대상 및 방법 등에 따라 달라질 수 있으며, 적절한 통계가설을 설정한 후 피험자 데이터 수 산출 공식을 적용하여 산정하도록 한다.

참고 의료기기 임상시험 테스트 환경 고려 사항의 예[12]

[AI 기반 의료기기의 일반화 도전 과제]

- 대부분의 AI 시스템은 대부분의 의료 데이터 유형에 대해 임상 적용 가능성이 작다. 특히, 일반화는 현장 간의 기술적 차이(장비, 코딩 정의 전자건강기록EHR, Electronic Health Records 시스템, 실험실 장비 및 분석의 차이 포함)와 현지 임상 및 관리 관행의 차이로 인해 어려울 수 있다.

[테스트 환경 수립 시 필요한 고려사항]

- 새로운 인구 및 환경에 대한 일반화 성능을 확보하도록 어느 정도 적용 대상 위치(병원 등)별 교육이 필요할 수 있다.
- 의료 영상 분류를 포함한 간단한 작업은 규모가 크고 수집한 데이터와 이질적인 다중 센터 데이터셋의 큐레이션curation으로 극복할 수 있다.
- 또한, 특정 목적으로 학습된 인공지능 모델에서 학습 데이터와 다른 인구 모집단의 데이터셋을 사용하였을 때 성능 변화가 있을 수 있음을 인지해야 한다.
- 따라서, 일반화 성능 확보가 필요한 인공지능 모델은 모델 학습을 위해 데이터를 제공한 기관뿐만 아니라 다른 기관에서 수집한 적절한 크기의 데이터셋을 사용한 외부 검증이 필요하다.

참고 인공지능 기반 의료기기의 임상적 유효성 확인 방법[5]

- 기계학습 가능 의료기기MLMD에 적용 가능한 임상적 테스트 환경을 조성하는 유효성 확인 방법으로는 크게 ❶전향적 연구, ❷후향적 연구 그리고 전향적 연구와 후향적 연구를 병행하는 ❸전향적·후향적 연구가 있다.
- 의료진의 유효성 확인 방법 설계 완료 후, 의료 인공지능 시스템의 특징을 고려하여 임상적 신뢰성을 확보하는 테스트 계획을 수립한다.

구 분	설 명
전향적 연구	• 위험 요소를 미리 설정한 후, 일정 기간에 변화를 추적하는 연구법으로 위험 요소가 일으키는 변화를 관찰하는 연구
후향적 연구	• 피험자의 모집 대신 이전의 진료 또는 임상시험을 통해 획득된 피험자의 의료 데이터를 이용하여 의료기기의 안전성과 유효성 검증을 실시하는 임상 연구 ※ 후향적 연구는 피험자의 진료 기록, 의료 영상, 생체 신호, 병리 검사, 유전 정보 등의 의료 데이터를 사용할 수 있음

03-1b 가상테스트 환경이 필요한 인공지능 시스템의 경우, 시뮬레이터를 확보하였는가?

Yes No N/A
☐ ☐ ☐

- 일반 분야와 달리 의료 분야의 특성상 실제 환경 테스트 또는 임상시험의 승인 절차가 까다로울 수 있다. 본격적인 테스트 절차 전에 가상테스트 환경에서 시뮬레이터를 통한 테스트를 수행하면 시간과 비용을 절약할 수 있다. 특히, 테스트 시 환자에게 직접적으로 손상을 줄 위험이 있는 시나리오가 포함된 인공지능 시스템은 가상테스트 환경을 구축해 위험을 완화할 수 있다. 다음은 현재 의료 분야에서 개발되고 있는 가상테스트 시뮬레이터 및 활용 예시이다.

  ✓ 가상 환자 시뮬레이션을 통한 치료 예측 및 모의 수술: 메디컬 트윈medical twin

  ✓ 〈메디컬 트윈 활용 예시〉

  - 의료 영상 정합 및 멀티피직스multiphysics 모델링 기반 고정밀 심뇌혈관 트윈 기술
  - 인공지능을 활용한 트윈 제작 가속화 및 질환 진단·예후 예측 성능 향상
  - 경동맥 3D 트윈 기반 협착 진행 및 플라크 파열 예측 기술
  - 관상동맥 허혈 진단 및 발달 예측 기술
  - 시뮬레이션 기반 심뇌혈관 중재 시술 내비게이션 자동화 기술
  - 3차원 혈관 정보 실시간 정합 기반 햅틱 시술 가이드 개발
  - 뇌동맥류 메시mesh 기반 뇌동맥류 성장 및 파열예측 기술

03-2 인공지능 시스템의 테스트 설계에 필요한 협의 체계를 구성하였는가?

Yes No N/A
☐ ☐ ☐

해당여부 판단

인공지능 시스템 사용 대상이 다양할 때 본 항목을 고려하여 만족 여부를 판단하십시오.

- 인공지능 시스템의 출력에 대한 설명이 필요한 시스템이라면, 시스템 출력을 확인하는 대상 사용자(예: 의료진, 환자 등)에 따라 출력에 대한 도출 방법을 이해하는 정도인 설명가능성*에 대한 평가 기준이 달라질 수 있다. 인공지능의 작동 방식을 이해하는 정도인 해석가능성의 평가 기준 역시 대상 사용자에 의존한다.

  * ISO/IEC TR 29119-11:2020에서는 설명가능성을 '인공지능 시스템이 주어진 결과를 어떻게 도출했는지 이해하는 정도'라고 정의하며, 해석가능성을 '인공지능 기술이 작동하는 방식에 대한 이해 정도'로 정의한다.

- 따라서 의료진, 환자마다 서로 다른 기준으로 해석할 수 있는 인공지능 시스템을 테스트하려면 기대 출력에 관한 결정이나, 시스템 출력에 대한 설명가능성 및 해석가능성의 평가 기준을 마련하는 협의체를 구성하고, 구성원 간 합의를 도출해 테스트 방법론을 설계하는 방식이 적절하다.

03-2a 인공지능 시스템의 기대 출력을 결정하기 위한 협의 체계를 구성하였는가?

Yes No N/A
☐ ☐ ☐

- 의료진을 위한 진단 보조를 수행하는 인공지능 시스템이라면, 시스템의 기대 출력을 결정하기 위해 해당 도메인의 내·외부 전문가로 구성된 협의체를 구성하여야 한다. 이때 기대 출력을 결정하기 위해 다수의 의료진이 동의하는 데 시간이 걸릴 수 있음을 인지하여야 한다.
- 의료진이 아직 안전성이 검증되지 않은 의료 인공지능 시스템에 의존하여 진단하거나 치료 방법을 결정할 때 부정확한 결과로 환자의 건강에 위해를 끼칠 수 있다. 식품의약품안전처의 인공지능 의료기기의 허가·심사 가이드라인에 따르면 의료 인공지능 시스템의 의사결정에 대해 의료인의 개입 절차가 없거나 판단하기 어려울 때 의료 인공지능 시스템 의사결정의 근거에 대한 검토가 필요하다. 따라서, 의료 인공지능의 기대 출력과 구체적인 의사결정에 대한 타당성 검토를 수행하기 위한 의료 분야 전문가 3인 이상으로 구성된 협의체를 구성하도록 권고한다.
- 질병에 대한 판정 기준이나 확진 데이터 확인이 필요한 경우에는 인공지능 시스템의 임상적 유효성을 예측하는 참조 표준을 확보(아래 참고)할 수 있다.

참고 임상의 전문가 그룹에 의한 참조 표준 구축 및 판독 방안[8]

- 임상의 전문가 그룹에 의해 확진된 데이터를 이용하여 참조 표준을 구축하는 경우, '질병에 대한 명시적인 판정 기준'이나 '참조 표준 의료기기를 이용한 검사 결과 또는 질병의 표준검사 방법에 따른 확진 데이터'가 없을 때 가능하다.

구분	설명
임상의 전문가 그룹	– 의료기기의 적응증과 임상시험 목적 등에 적합한 전공과 경력을 갖춘 복수의 임상의로 구성 – 편향 방지를 위해 참조 표준 구축에 참여한 임상의는 임상시험 평가자로 참여하지 않을 것을 권고
질병 판독 방법	– 모든 임상의는 동일 진료 지침을 이용하여 질병 판독
임상의 합의 판정 방법	– 판독 의견 불일치 시, 적절한 합의 과정을 통해 해결 ※ 필요시, 해당 질환과 관련한 임상 학회나 단체 등의 견해 반영 가능 ※ 합의 방법은 임상시험에 참여하는 판독자의 수를 고려하여 자율적으로 설정 가능 예) 3인의 판독자 참여 시 2인의 일치된 판독 결과를 채택 가능 예) 2인의 판독자 참여 시 관련 분야의 경력이 우수한 임상의가 주도하여 합의 가능 – 무작위 배정과 눈가림을 통해 편향 최소화
임상시험 디자인	– 임상시험의 목적에 맞게 설계하고, 제품의 특성, 유효성 입증 방법(우월성, 동등성, 비열등성 평가 등), 유효성 평가 기준 등을 고려 – 후향적 임상시험도 목적에 따라 평행설계나 교차설계 적용 가능 – 임상시험 설계 시 시험군 또는 대조군 평가에 임상의가 참여한다면 임상의는 임상시험 피험자 데이터의 판독 결과를 알 수 없도록 눈가림 방법 필요

03-2b 설명가능성 및 해석가능성 확인을 위한 사용자 평가단을 구성하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템의 설명가능성과 해석가능성을 테스트하려면 대상 사용자가 시스템의 출력과 작동 방식을 얼마나 쉽게 이해하는지 확인하여야 한다.
- 진단 보조 인공지능 의료기기 사용자는 의학적 진단을 수행하는 의료진과 그러한 진단 결과를 받는 환자 등 크게 두 부류로 구분할 수 있다. 사용자가 의료진이면, 의료진의 의학적 진단 및 판단을 효과적으로 보조하는 설명가능성 및 해석가능성을 어느 정도로 설정해야 하는지 논의할 필요가 있으며, 이는 개발에 참여하는 의료진의 시스템 평가 및 합의를 통해 달성될 수 있다. 사용자가 환자이면, 환자의 데이터 수집 또는 진단 정보 제공 시 불편했던 점을 조사하는 사용자 평가를 수행함으로써 환자를 위한 설명가능성 및 해석가능성을 개선할 수 있다.
- 따라서 의료진과 환자 각각으로 구성된 사용자 평가단을 구성하여 설명을 어느 수준으로 제공할지 결정하고, 이를 모델이나 시스템 구현 과정에 반영해야 한다. 이를 위해, 계획 및 설계 단계에서 임상 분야별로 대상 사용자를 명확히 정의한 후 사용자 평가단을 구성해야 한다.
- 사용자 평가단의 평가 결과에 따라 테스트의 통과 또는 실패 여부를 결정할 기준을 마련하는 것이 필요하다. 예를 들어, 평균 점수가 일정 점수 이상일 때 통과를 결정하는 등의 정량적 기준 마련이나, 평균 점수 계산 시 절사평균*의 활용 여부 등 산출 기준 마련 등이 있다. 또한, 정량적 산출 기준과 함께 인공지능 시스템에 의해 출력된 정보에 대한 임상적 근거를 파악할 수 있는 학습 데이터셋의 출처, 학습 데이터셋과 추론 결과 사이의 상관관계 등에 대한 충분한 설명 제공 여부를 평가 기준으로 고려하는 것이 가능하다.

* 편차가 큰(극단치가 존재) 자료는 산술 평균이 적합하지 않으므로, 자료의 총 개수에서 일정 비율만큼 가장 큰 부분과 작은 부분을 제거한 후 평균 산출

02 데이터 수집 및 처리

책임성 투명성

요구사항 04 데이터의 활용을 위한 상세 정보 제공

대표행위자 | 데이터 과학자 협력 대상 | 데이터 공급자 전문 의료진 인공지능 모델 개발자

- 의료 인공지능 데이터는 수집 과정에서 장비의 종류나 병원의 방침에 따라 상당한 차이를 보일 수 있다. 병원마다 각각 최적화된 장비를 운용하기에 개발 과정에서 병원이 운용하는 의료기기에 적합한 데이터를 수집하여 학습해야 하기 때문이다. 따라서 데이터로 인해 발생 가능한 문제에 능동적으로 대응할 수 있도록 개발 환경을 고려하여 데이터 활용을 위한 상세 정보를 제공한다.

04-1 데이터의 명확한 이해와 활용을 지원하는 상세한 정보를 제공하는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

의료 분야 인공지능 알고리즘 또는 모델 개발 시 데이터셋을 직접 구축하거나, 향후 추가 데이터 수집 가능성이 있을 때 본 항목을 고려하여 만족 여부를 판단하십시오.

- 의료 데이터는 개인 정보 비식별화 처리로 데이터의 출처 확인이 어렵다. 따라서 이후 데이터를 재활용하는 상황이나 동일한 형식의 추가 데이터 수집 필요 시 원시 데이터[raw data]의 특징을 파악하기 위한 메타데이터[metadata]를 제공해야 한다.
- 또한, 개발자뿐만 아니라 인공지능 시스템과 관련된 이해관계자들이 수집 데이터를 이해하고 발생 가능한 편향이나 오류를 방지할 수 있도록 학습 데이터 및 메타데이터, 라벨링 작업 가이드 등의 데이터에 대한 정보가 확보되어야 한다.
- 이해관계자들에게 전달되어야 할 정보의 예로는 수집 데이터의 출처와 형식, 데이터 수집·정제·가공 방법, 데이터 라이선스, 편향 유발 가능성이 있는 보호변수[protective attribute] 등이 있다.

04-1a 정제 전과 후의 데이터 특성을 설명하였는가?

Yes ☐ No ☐ N/A ☐

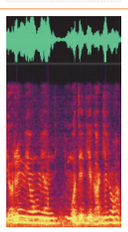
- 의료 분야는 데이터 정제 과정에서 의학적 지식을 갖춘 의료진의 참여가 필수적이며, 도메인 지식 기반의 데이터 표준화 작업과 정제작업이 필요하다.
- 이 과정에서 조직적 데이터 큐레이션, 데이터 거버넌스 확보 등의 절차가 수행되어야 하며, 이해관계자들의 적절한 데이터 활용을 돕도록 정제 전과 후의 데이터 특성에 대한 정보를 명시하여야 한다.
- 데이터 품질 보장, 데이터 관리 최적화, 데이터 구축 목적 수립, 데이터 종류 분석 등 정제 기준 정보 및 정제 도구 정보 제시가 필요하다. 다음은 데이터 종류별 데이터 정제 기준의 예시이다.

✓ 이미지 데이터: 이미지 크기, 비율, 화질, 촬영 장비, 개인 정보, 지식저작권 등
✓ 텍스트 데이터: 텍스트 분량, 텍스트 문법 정확성, 텍스트 내용 적절성, 주제와의 연관성 등
✓ 음성 데이터: 음량, 발음 정확성, 소음 및 잡음, 안 들림(허용 범위 기준), 개인 정보, 저작권 등
✓ 비디오: 화질, 영상 손실 여부, 개인 정보, 정치적 견해, 특정 인물 비하 등
✓ 3D: 포인트 클라우드 획득, 메시 데이터 최적화, 표준 모델 생성 등
✓ 센서: 단위, 결측값, 센서의 기록 시간 등

- 또한, 데이터셋 구축 과정에서 데이터 품질을 높이기 위해 일부 데이터가 다시 정제되며, 다음은 정제 후 학습 데이터 특징으로 설명할 수 있는 항목의 예시이다.

✓ 데이터 선별 및 처리 설명 항목: 중복성 방지, 이상 데이터 제거, 샘플링 등
✓ 통계적 설명 항목: 클래스별 학습 데이터 수, 피실험자 수 등

AI Hub 내 헬스케어 분야 주요 데이터셋 유형별 데이터 수집/정제 예시(2022년 10월 기준)

데이터명	데이터 예시	데이터 종류	데이터 유형	데이터 포맷	데이터 정제 방법
소아 흉부 이미지 데이터		X선 이미지	이미지	DICOM	- 데이터 선별(화질이 선명하지 않거나 움직임이 있을 때, 폐 영역 일부가 잘린 때, 나이대별 질환 종류에 부합하지 않을 때, 측면 영상일 때 제외) - 개인 정보 복구 불가능한 비식별화 처리
측면두부 규격방사선 사진 데이터		X선 이미지	이미지	DICOM	- 데이터 선별(부정교합을 진단하는 19개 계측점 모두 포함하지 않을 때 제외) - 개인 정보 비식별화 처리
후두질환 판독을 위한 후두 내시경 데이터		내시경 이미지	이미지,	JPG, PNG	- 데이터 선별(후두 주변부가 보이지 않을 때, 화질이 선명하지 않으면 제외) - 개인 정보 비식별화 처리
심장질환 진단을 위한 심전도 데이터		심전도	텍스트	XML	- 데이터 선별(식별자를 통해 데이터 확인이 안 될 때, 전문학회 진료 지침 또는 주요 임상 연구 정의 기준에 맞지 않을 때 제외) - 개인 정보 비식별화 처리
			이미지	JPG	
음성질환 판별을 위한 음성 데이터		음성	오디오	WAV	- 데이터 변환을 통한 개인 정보 비식별화 처리 ※ 보건의료 데이터 활용 가이드라인 준수 - 표준화(소리 크기 등)
			텍스트	CSV	

04-1b 학습 데이터와 메타데이터metadata를 구분하고 각 명세자료를 확보하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 학습 데이터셋을 활용하기 위해서는 데이터셋에 대한 정보를 파악해야 하는데, 이러한 정보를 메타데이터라고 한다. 메타데이터는 JSON, XML 등의 형식으로 제공할 수 있다.
- 메타데이터와 학습 데이터는 구분되어야 하며, 각각에 대한 명세자료를 작성하여 개발자 관점에서 인공지능 모델 학습 등에 활용이 용이하도록 해야 한다.
- AI Hub에서 제공하는 헬스케어 데이터셋은 데이터의 유형(이미지, 비디오, 텍스트, 오디오, 3D, 센서 등)에 따라 데이터 영역, 형식, 유형, 출처, 라벨링 유형 및 형식, 데이터 활용 서비스, 데이터 구축 연도와 구축량에 대한 정보를 명세서로 제공한다.
- 의료 분야는 메타데이터상에도 환자 번호와 성명 등 개인의 민감한 정보를 포함할 때가 많다. 따라서 개인정보보호법에 따라 보건의료 데이터 활용 가이드라인을 준수하여 가명 처리나 비식별화를 수행하여야 한다.
- TTA 정보통신단체표준 TTAK.OT-10.0245에서는 PACSPicture Archiving and Communication System 표준 Log 메타데이터 구성 요소를 표준으로 제정하였다. 해당 표준에서는 의료 영상 데이터를 대상으로 5개 범주(환자 정보, 검사 정보, 영상 정보, 장비 정보, 시리즈 정보)의 메타데이터 구성 요소를 제시하고 있으며, 해당 메타데이터에서도 환자 정보는 개인 정보로 구분될 수 있으니 주의를 기울여야 한다.

의료 영상 데이터 PACS 표준 메타데이터 구성 요소

범주	요소	내용	필수 여부
Patient	Patient	환자 정보	필수
	Clinical Trial Subject	임상실험 분야	선택
Study	General Study	일반검사 내용	필수
	Patient Study	개별환자 검사내용	선택
	Clinical Trial Study	임상실험 검사	선택
Series	General Series	일반적 시리즈 내용	필수
	Clinical Trial Series	임상실험 시리즈 내용	선택
Equipment	General Equipment	의료 장비에 대한 일반적 내용	필수
Image	General Image	영상의 일반적 속성	필수
	Image Pixel	이미지 픽셀	필수
	SOP Common	관련된 SOP instance를 확인하고자 요구되는 속성 명세	필수

※ 요소별 하위요소는 TTA 표준 TTAK.OT-10.0245 참조

Use Case 학습 데이터와 메타데이터 명세서

• AI Hub 내 헬스케어 분야 '소아 흉부 이미지 데이터' 학습 데이터 명세서 예시(2022년 10월 기준)

데이터 영역	헬스케어	데이터 유형	이미지
데이터 형식	DICOM	데이터 출처	고려대구로병원 외 8개 병원 (참여기관)
라벨링 유형	세그멘테이션 이미지 및 임상정보 Text	라벨링 형식	PNG / JSON
데이터 활용 서비스	소아 흉부질환 진단 서비스	데이터 구축년도 / 데이터 구축량	2021년 / 소아 흉부 이미지 데이터 21,217건

• AI Hub 내 헬스케어 분야 '소아 흉부 이미지 데이터' 데이터 구성 및 어노테이션 포맷 예시(2022년 10월 기준)

No	속성명	항목설명	필수여부	타입	비고
annotations(어노테이션)					
1	project_code	프로젝트코드	Y	string	D74
2	identifier	고유번호	Y	string	병원코드+일련번호+순번 (예) H01_00001_01
dicom_info(DICOM정보)					
3	modality	모달리티 구분	Y	string	(예) CR
4	study_id	스터디 ID	Y	string	
5	series_no	시리즈 번호	Y	integer	
6	instance_no	인스턴스 번호	Y	integer	
patient(환자정보)					
7	age_group	연령대	Y	string	[표] 참조
8	diagnosis	진단명	Y	string	[표] 참조
9	pneumonia_type	폐렴유형		string	1:바이러스성, 2:세균성, 3:알수없음, 4:바이러스성+세균성
10	report	진단내역		string	
11	sex	성별	Y	string	M:남, F:여
12	age	나이		string	1세미만 : 0~11개월 1세이상 : 1세~15세
13	height	키		number	cm
14	weight	몸무게		number	kg
mask_image(어노테이션 이미지)					
15	org_dicom_file	원본파일 경로	Y	string	
16	body_part_mask	장기부위 마스킹 파일경로	Y	string	
17	lesion_part_mask	병변부위 마스킹 파일경로		string	

〈 Annotation 예시 〉


```
{
  "project_code": "D74",
  "identifier": "H07_00690_01",
  "dicom_info": {
    "modality": "CR",
    "study_id": "00000",
    "series_no": 1,
    "instance_no": 2
  },
  "patient": {
    "age_group": "A1",
    "diagnosis": "05",
    "pneumonia_type": "3",
    "report": "Slightly accentuated peribronchovascular markings, especially right lung.",
    "sex": "M",
    "age": "0개월",
    "height": 39.5,
    "weight": 1.3
  },
  "mask_image": {
    "org_dicom_file": "D74/05.폐렴.A1.0-2개월/H07_00690_01/org/H07_00690_01.dcm",
    "body_part_mask": "D74/05.폐렴.A1.0-2개월/H07_00690_01/mask/H07_00690_01.png",
    "lesion_part_mask": "D74/05.폐렴.A1.0-2개월/H07_00690_01/mask/H07_00690_02.png"
  }
}
```

04-1c 보호변수protective attribute의 선정 이유 및 반영 여부를 설명하였는가?

Yes ☐ No ☐ N/A ☐

- 인종, 성별, 연령, 직업, 가족력, 장애 여부 등 일반 분야에서 사회적 문제가 될 수 있는 보호변수가 의료 분야에서는 인공지능 모델 설계 목적에 대한 중요한 특성으로 사용된다.
- 그러나 인종 다양성을 띠는 해외에서는 예기치 못한 윤리적 편향 문제 사례가 보고된다. 따라서 모델의 설계 목적 외에 발생 가능한 편향에 대비해 학습 데이터 수집 및 라벨링 단계에서 의료진의 임상적 판단에 근거하여 보호변수의 선정 및 반영이 필요하다.
- 또한, 수집·구축된 데이터의 향후 사용자를 고려하여 개발하는 인공지능 시스템의 목적과 데이터셋의 보호변수 선정 이유, 과정 및 반영 여부에 대한 설명이 제공되어야 한다.

참고 보호변수 조사 사례[13]

- MIT 연구진은 인공지능 모델이 의료 이미지만으로 환자의 인종을 정확히 예측할 수 있음을 발견하였다. 인공지능 모델이 어떻게 인종(백인, 흑인, 아시아인)을 식별하는지 조사하고자 흉부 X-Ray, 사지 X-Ray, 흉부 CT 스캔 및 유방조영술 이미지 데이터를 사용하여 해부학적 차이, 골밀도, 이미지 해상도 등과 같은 변수의 영향력을 조사하였다.

04-1d 라벨링 작업자를 위해 교육을 시행하고 작업 가이드 문서를 마련하였는가?

Yes ☐ No ☐ N/A ☐

- 데이터 라벨링 작업은 인공지능 모델을 학습하기 위한 원시 데이터의 주석(정답) 작업에 해당하며, 특정 질환을 진단하는 과정을 전문적이고 세밀하게 가공해야 하므로 이 역시 전문 의료진의 참여가 필수적이다.
- 단, 세부 전문성과 경력 등의 차이, 사회적 편향으로 인해 의료진의 진단 기준이 다를 수 있어, 다수의 전문의를 선정하여 합의하는 과정을 통해 데이터셋의 품질을 확보해야 하며, 작업자의 교육 및 상세한 작업 가이드 문서를 마련하는 것이 중요하다.
- 라벨링 작업은 데이터 종류에 따라 작업 대상, 범위, 상세 절차 및 라벨링 도구 등이 달라질 수 있다. 일반적인 라벨링 작업 절차는 아래와 같으며, 작업 절차에 따라 작업자 대상 교육과 가이드 문서가 확보되어야 한다.
  - ✓ 데이터 획득 및 정제: 원시 데이터 획득 및 데이터 정제 작업을 진행한다.
  - ✓ 라벨링 작업 대상과 범위 정리: 원시 데이터 내의 어떤 항목들을 라벨링 하는지 대상과 범위를 정의한다. 특히, 데이터 종류에 따라 세부적인 기준을 마련해야 한다(데이터 일부 라벨링, 개인정보 비식별화, 클래스 정의 및 관리 등).
  - ✓ 라벨링 방법 및 절차 수립: 라벨링 할 정보에 따라 자동·반자동·수동 등의 작업 방식을 결정하고, 작업의 배분 및 데이터별 라벨링 기준 등 상세한 작업 기준을 마련한다.

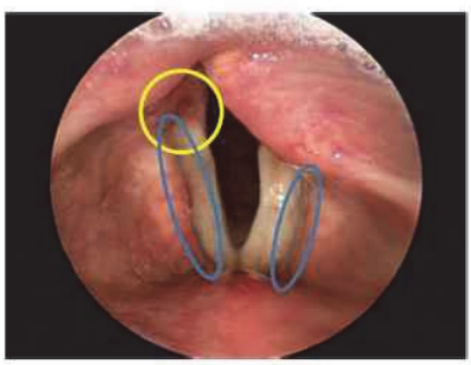
✓ 라벨링 작업 진행: 상세 작업 기준으로 작업자 교육 후, 데이터 라벨링 작업을 시행한다(앞서 결정한 작업 방식에 따라, 자동·반자동일 경우, 적절한 라벨링 도구 선정 및 교육 진행).

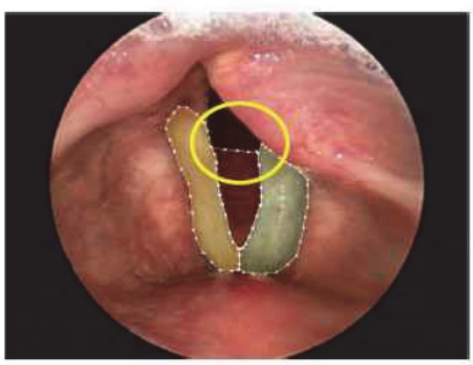
참고 라벨링 작업자를 위한 작업 가이드 예시(출처: AI Hub)

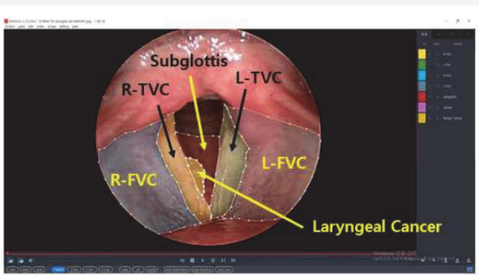
• '후두질환 판독을 위한 후두 내시경 데이터'의 어노테이션/라벨링 작업 기준 예시

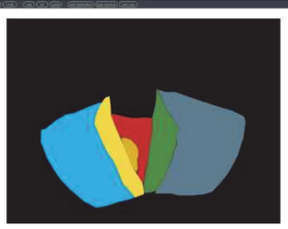
■ 라벨링 기준

- TVC: 위쪽 부분은 색이 바뀌는 부분을 기준으로 삼음
- TVC와 FVC: TVC와 FVC 사이의 경계를 기준으로 삼음
- Subglottis: 하부성대가 구분이 안되는 경우, 온전히 다 보이는 진성대(TVC)의 윗부분을 기준으로 대략 1cm 밑을 기준으로 삼음









Classification			비고	
해부학적 구조물	TVC (진성대)	R-TVC	Right True Vocal Cord	
		L-TVC	Left True Vocal Cord	
	FVC (가성대)	R-FVC	Right False Vocal Cord	
		L-FVC	Left False Vocal Cord	
	Subglottis (하부성대)			
	Laryngeal Cancer		악성종양	
병변	Benign Tumor	Laryngeal nodule	출혈성 용종	양성 종양
		Intracordal cyst	성대 내낭	
		Leukoplakia	성대백반증	

해부학적 구조물 및 병변 클래스 리스트

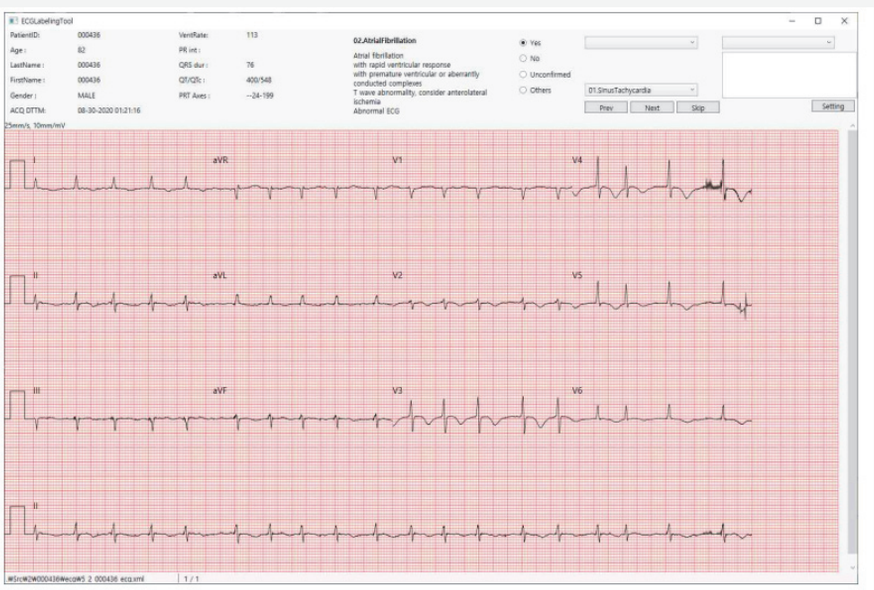
• ‘심장질환 진단을 위한 심전도 데이터’의 어노테이션/라벨링 도구 설명 예시

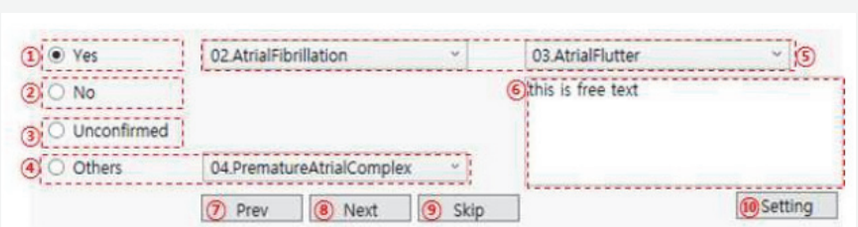
■ 어노테이션/라벨링 도구

– INFINITT ECG Viewer

※ XML 형태의 원시데이터를 읽어 심전도 그래프를 보여주는 프로그램

※ 심전도 부정맥질환의 라벨링을 위해 제작되었으며, 추후 XML 원시데이터 뷰어로 제공





초기설정	Src 폴더에 XML 파일과 JSON 파일을 넣고 프로그램을 실행한다.
1	Yes를 선택하면 XML 파일이 Src 폴더에서 Confirm 폴더로 이동한다.
2	No를 선택하면 XML 파일이 Src 폴더에서 Discard 폴더로 이동한다.
3	Unconfirmed를 선택하면 XML 파일이 Src 폴더에서 Unconfirmed 폴더로 이동한다.
4	Other를 선택하고 오른쪽의 진단명을 선택하면 해당 폴더로 XML 파일이 이동한다. 주진단명을 변경하기 위해 사용한다.
5	진단명을 선택하면 XML 파일 하단에 UserDiagnosis1, UserDiagnosis2 항목으로 추가된다. 부진단명을 입력하기 위해 사용한다.
6	텍스트를 입력하면 XML 파일 하단에 Comment 항목으로 추가된다.
7	이전 ECG를 다시 불러온다. 키보드 방향키 [←]로도 동작한다.
8	현재 진단이 저장되고 다음 ECG를 불러온다. 키보드 방향키 [→]로도 동작한다.
9	현재 진단을 보류한다. 키보드 [Page Down]키로도 동작한다.
10	설정 창이 뜬다. Source, Confirm, Discard, Unconfirmed 폴더 경로를 설정한다.

04-2 데이터의 출처는 기록 및 관리되고 있는가?

Yes No N/A
☐ ☐ ☐

해당여부 판단

인공지능 알고리즘 또는 모델 개발에 활용하는 데이터셋을 직접 구축하거나 오픈소스 데이터셋을 활용 시 본 항목을 고려하여 만족 여부를 판단하십시오.

- 의료 분야는 일반 분야와는 달리 제품의 목적에 특정 질환이나 지역, 의료 기관 협업 등 제한적 상황으로 인해 데이터의 신뢰도를 확보하고자 제품 목적에 적합한 데이터를 자체 수집할 때가 많으며, 이에 따라 오픈소스 데이터셋의 활용도는 낮은 편이다.
- 자체 수집 시 의료 기관이 수집한 환자의 의료 데이터를 사용하기 위한 연구 계획서를 심의위원회에 제출하여 사용에 대한 승인을 받을 수 있다. 이러한 일련의 승인 절차를 통해 데이터의 신뢰성을 확보할 수 있다.
- 때에 따라서는 오픈소스 데이터셋을 활용할 수도 있는데, 오픈소스 데이터셋은 다수의 사용자가 데이터를 활용하는 과정에서 발견한 오류가 추후 나타날 수 있으며, 이로 인한 데이터셋 수정, 재구축으로 데이터의 버전이 변경될 수 있다.
- 이러한 데이터셋 자체 원인으로 발생할 수 있는 인공지능 모델의 문제 대응을 위해서는 학습에 사용한 데이터셋의 명확한 출처, 구축 시점, 오픈소스 데이터셋 버전 등의 정보를 관리하여야 한다.

04-2a 신뢰할 수 있는 출처로부터 제공되는 데이터셋을 사용하였는가?

Yes No N/A
☐ ☐ ☐

- 의료 기관에서 수집한 데이터를 활용하려면 기관생명윤리위원회의 승인을 획득해야 하며, 승인 절차는 보건복지부의 보건의료 데이터 활용 가이드라인에 명시되어 있다.
- 대표적인 의료 오픈소스 데이터셋 플랫폼은 '보건의료 빅데이터 플랫폼(보건복지부)', '바이오헬스 빅데이터 플랫폼(산업통상자원부)', '마이데이터(과학기술정보통신부)'등이 있다. WEF[World Economic Forum]는 오픈소스 데이터셋 활용 이전에 신뢰할 수 있는 데이터인지 미리 확인할 것을 권고한다.

참고 지도학습을 위한 데이터 품질 관리 요구사항 – 출처의 신뢰성 확보

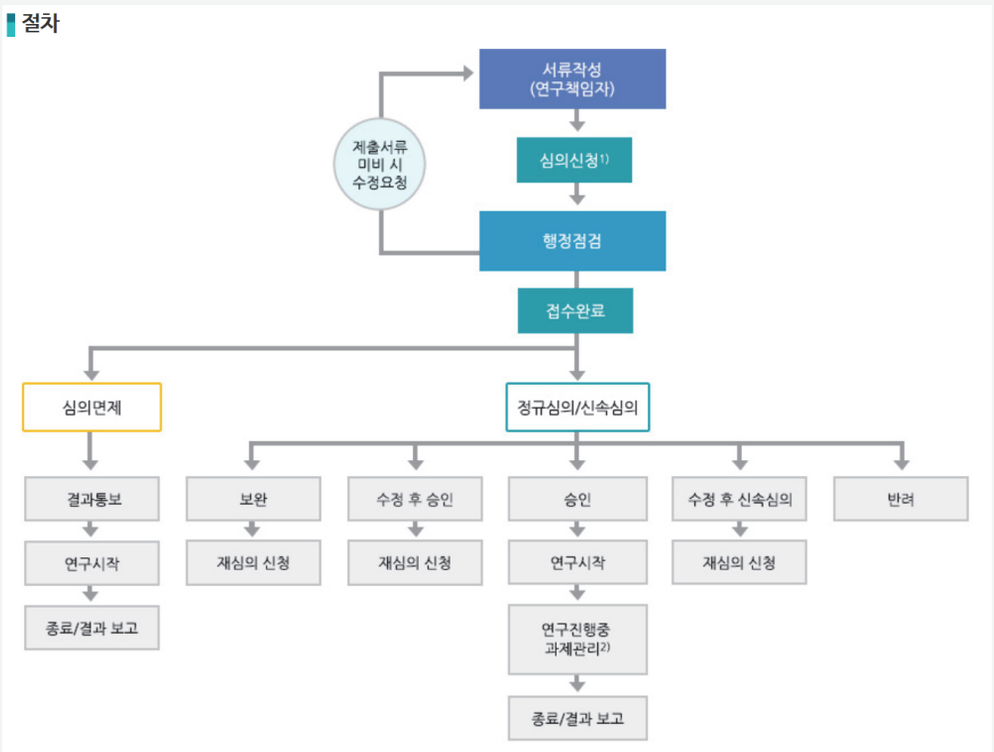
TTA 정보통신단체표준 TTAK.KO-10.1339에서는 지도학습 계열의 인공지능 기술에 활용되는 데이터 획득 시 출처의 신뢰성 확보 측면에서 고려해야 할 내용을 정리하였다.

- 데이터 획득 시 직접 생산 혹은 제삼자에 의해 생산된 데이터 중계의 2가지 방법으로 데이터를 획득할 수 있는데, 제삼자에 의해 생산된 데이터를 중계하여 획득할 때 데이터의 출처에 대하여 신뢰성을 확보하여야 하며 다음과 같은 요소를 고려할 수 있다.
  - 제삼자가 데이터 획득 시 개인정보보호, 지식재산권, 사전 승인/허가 등과 관련하여 정식 절차를 거쳐 문제없이 획득하였는지 여부
  - 제공하는 데이터셋의 규모가 충분히 커 데이터 사용자가 원하는 학습용 데이터를 제공하는 데 문제가 없는지에 대한 여부
    ※ 규모가 충분하지 않은 경우, 데이터 획득을 재차 시도하고자 할 때 수급에 문제가 있을 수도 있음
  - 해당 데이터의 지속적인 업데이트 및 추가 제공 등의 여부
  - 데이터와 함께 설계서 내용이 명확하게 제공되는지 여부
  - 해당 데이터 활용건수 및 인용건수가 많아 범용성이 높은지 여부
- 반면, 데이터를 직접 생산(이미지/동영상 촬영, 발화 녹음, 텍스트 작성 등)하는 경우, 위의 내용 중 첫 번째 사항을 고려하여야 한다.

참고 기관생명윤리위원회 정보포털에 안내된 IRB 심의 절차

- 기관생명윤리정보위원회는 인간 대상 연구에 대한 심의, 조사 감독, 교육 등을 제공하여 연구대상자의 권리와 안전 및 복지를 보호하는 기관으로써, 사람을 대상으로 하는 의료 분야 인공지능 시스템 개발 역시 환자의 의료 정보 활용 이전에 승인을 얻어야 한다.
- 연구대상자 등 공공에 미치는 위험이 미미하면 심의받지 않을 수 있으며, 심의면제 대상은 아래와 같다.

1. 연구대상자를 직접 조작하거나 그 환경을 조작하는 연구 중 다음 각 목의 어느 하나에 해당하는 연구
   가. 약물 투여나 혈액 채취 등 침습적 행위를 하지 않는 연구
   나. 신체적 변화가 따르지 않는 단순 접촉 측정 장비 또는 관찰 장비만을 사용하는 연구
   다.「식품위생법」시행규칙 제3조에 따라 판매 등이 허용되는 식품 또는 식품첨가물을 이용하여 맛 또는 질을 평가하는 연구
   라.「화장품법」제8조에 따른 안전기준에 맞는 화장품을 이용하여 사용감 또는 만족도 등을 조사하는 연구
2. 연구대상자 등을 직접 대면하더라도 연구대상자 등이 불특정하며,「개인정보보호법」제23조에 따른 민감 정보를 수집하거나 기록하지 않는 연구
3. 연구대상자 등에 대한 기존의 자료나 문서를 이용하는 연구



(출처: 기관생명윤리위원회 정보 포털)

04-2b 오픈소스 데이터셋을 활용하는 경우, 출처를 명시하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 모델 학습에 오픈소스 데이터셋을 사용한 경우, 학습 시점에는 발견되지 않았던 오류나 편향된 결과가 나올 수 있다. 편향된 결과는 사회 인식 변화에 따른 윤리적 문제와도 결부될 수 있어 오픈소스 데이터셋 구축 당시 인식하지 못한 데이터 편향의 발생 가능성이 있다.
- 따라서 오픈소스 데이터셋을 활용하여 학습 기반 인공지능 모델을 구축할 경우, 과거·현재·미래 시점에 발생할 수 있는 데이터 편향의 원인을 파악하고자 확보된 데이터의 명확한 출처 및 관련 정보를 명시하여 관리해야 한다.

안전성

요구사항 05 데이터 강건성 확보를 위한 이상abnormal 데이터 점검

대표행위자 | 데이터 과학자 협력 대상 | 데이터 공급자 인공지능 모델 개발자 전문 의료진

- 의료 인공지능 모델 및 시스템의 성능은 품질이 저하된 데이터나 입력 데이터의 누락(예를 들어, 심전도상의 과도한 운동, 인공물) 등으로 인하여 손상될 수 있다. 따라서 의료 인공지능 개발 시 품질이 좋지 않거나 사용할 수 없는 입력 데이터를 식별하고, 해당 문제를 처리하는 방법을 지정한다.
- 통계적 방법 및 기법을 활용하여 이상값outlier을 식별하고 처리할 시에는 의료 전문가의 교차 검토를 통해 데이터의 제외 또는 반영 여부를 판단·결정하고 문서화한다.

05-1 이상 데이터의 식별 및 정상 여부를 점검하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

인공지능 모델 개발 시 학습 데이터를 직접 구축하거나 사전 학습된 인공지능 모델에 사용된 학습 데이터에 대한 정상·오류 여부가 명확하게 확인되지 않은 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 이상 데이터란 학습용 데이터를 구성하는 데이터셋의 수집 및 가공 과정에서 발생할 수 있는 다양한 오류error 와 일반적인 데이터의 범위에서 크게 벗어난 데이터 이상값을 포괄한다. 학습 데이터의 수집 및 가공 과정에서 발생하는 이상 데이터는 데이터상의 노이즈, 학습 데이터 내의 편향, 잘못된 라벨링, 라벨링 누락 등 다양하게 존재할 수 있으며, 이를 점검하여 대처하지 않으면 인공지능 모델의 성능 및 강건성을 충분히 확보할 수 없다.
- 단, 낙상이나 사고를 감지하는 이상 탐지anomaly detection 시스템에 활용되는 인공지능 모델은 이상 데이터가 제거해야 할 데이터가 아닌 학습 데이터가 될 수 있으므로 유의하여야 한다.
- 비정형 데이터unstructured data 를 학습에 활용하는 경우, 데이터 전처리 과정에서 이상 데이터의 식별을 위한 별도 기법을 마련하여야 한다.

05-1a 전체 학습용 데이터 분포를 시각화하여 발생 가능한 오류들을 확인하였는가?

Yes ☐ No ☐ N/A ☐

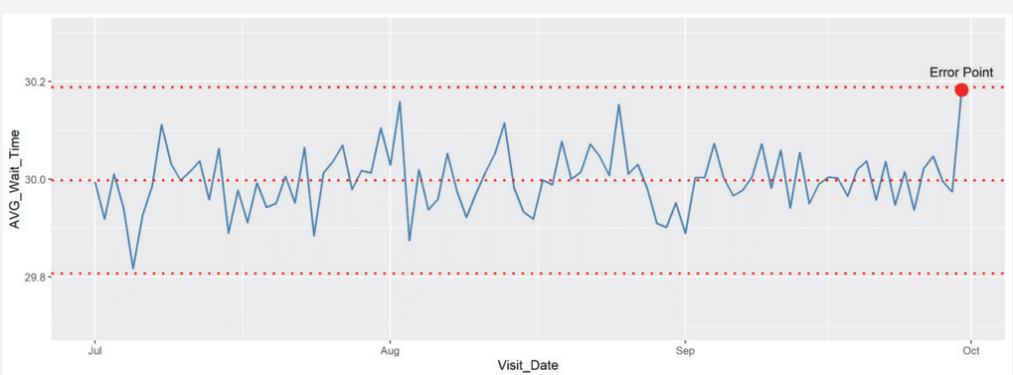
- 데이터 전처리 과정 중 하나인 데이터 정제 단계 이후, 데이터 전체 분포를 시각화하여 추가적인 입력 오류를 확인할 수 있다. 환자 진료 기록 데이터 등 간호사, 전문의 등이 수기로 입력하는 데이터는 NULL 또는 N/A와 같은 데이터 누락이 발생하지 않았으나, 오기입 등 인적 오류로 이상값을 발생시킬 수 있어 데이터 분포의 시각화를 통해 이러한 이상치를 확인할 수 있다.
- 적은 데이터에서는 손쉽게 이상값, 오류 등을 발견·관리할 수 있으나, 백만 건 이상의 환자 진료 기록 데이터 등 대량의 데이터를 활용·관리할 때는 시각화를 통해 사람의 실수로 발생 가능한 오류들을 더욱 용이하게 확인할 수 있다. 또한, 이러한 데이터 분포 시각화는 인공지능 모델 학습을 위한 데이터 탐구 및 이해에 많은 도움을 준다.
- 데이터 분포 시각화 방법은 데이터의 특성에 따라 다양한 기법이 존재한다. 전체 데이터의 평균, 분산, 편차 등을 활용하여 데이터 분포를 시각화하는 분포 도표, 범주형 데이터를 시각화하는 범주형 도표, 2차원 행렬 데이터를 시각화하는 행렬 도표 등이 있다.

참고 정상으로 보이는 입력 데이터의 오류를 확인하는 시각화 분석 사례[14]

- 환자의 등록, 진료 시간을 입력하고, 진료까지의 대기 시간을 계산하여 평균 대기 시간을 분석 예측하는 데이터의 오류 발생 사례
  - 하루 평균 6,000명의 방문 환자에 대한 대기 시간을 분석하고 예측하는 학습용 데이터 중 환자 등록 시간 오기입으로 대기 시간 자동 계산 결과가 781분으로 계산된 이상값이 생성됨

Patient ID	Registration	Seen by Professional	Calculated Wait Time
001	2019-09-30 08:00:00	2019-09-30 08:25:00	25 Minutes
002	2019-09-30 12:05:00	2019-09-30 12:38:00	33 Minutes
003	2019-09-30 00:00:00	2019-09-30 13:01:00	781 Minutes

  - 일일 평균 대기 시간을 시각화하여 분석 시, 이상값이 생성된 날짜에서 평균 대기 시간이 정상 범주보다 높은 평균 대기 시간을 가지는 이상값이 발생한 것을 확인할 수 있음



05-1b 학습 데이터 이상값 식별 기법을 적용하였는가?

Yes ☐ No ☐ N/A ☐

- 이상값의 탐색은 일관성 있는 분석 결과를 산출하기 위해서 우선 수행되어야 하며, 이상값이 포함된 자료 분석은 모형의 오류, 편향된 결과를 도출할 수 있다.
- 이상값은 비합리적인 이상값과 합리적인 이상값으로 구분할 수 있다. 비합리적인 이상값은, 입력 오류 등 자료의 오염으로 인해 발생한 이상값을 의미하고, 합리적인 이상값은 정확하게 측정되었으나, 다른 자료들과 전혀 다른 경향이나 특성을 보이는 이상값을 의미한다.
- 이상값 탐색 시에는 가면효과masking effect와 수렁효과swamping effect를 주의해야 한다. 가면효과는 일부 극단값에 의해 이상값으로 분류되어야 할 측정값들이 정상 범주의 값으로 나타나는 현상이며, 수렁효과는 정상 범주의 측정값이 이상값과 근접하여 같은 이상값으로 나타나는 현상이다. 가면효과와 수렁효과를 해결하려면 강건한 중심값centroid의 측정과 이상값에 영향을 덜 받는 공분산 행렬을 사용해야 한다.
- 의료 분야에서의 이상값을 식별할 때는 이상값이 의미 있는 의료 정보를 내포/반영할 수도 있음에 유의하여 이상값을 결정할 수 있도록 한다. 다수의 병원에서 진료하는 질병의 종류 및 진료비 데이터, 병원별 총 진료비, 입원/퇴원 날짜에 따른 입원 기간, 매일 평균 입원 비용 등 다차원 의료 데이터를 활용할 경우 전통적인 통계 방법을 적용하여 식별된 이상값이 실제로는 유의미한 값일 수 있으므로, 통계 기반 이상값 알고리즘을 보완(이상값 식별 방법의 조합, 개선 등) 적용하는 것을 고려할 수 있다.

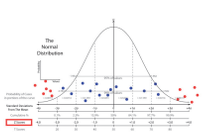
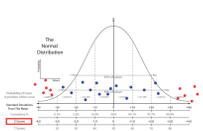
접근 방법에 따른 이상값 탐색 방법의 분류

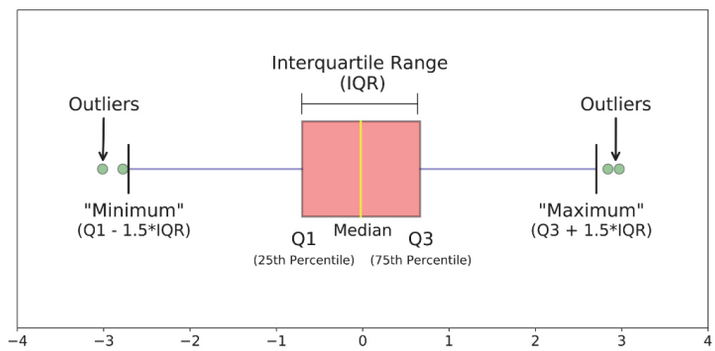
접근 방법	이상값 탐색 방법 분류
자료의 크기	소표본, 대표본
자료의 차원	일차원, 이차원, 다차원
변수의 개수	일변량, 이변량, 다변량
목표 변수의 유무	지도 방법, 비지도 방법
통계적 방법	모수적 방법, 비모수적 방법, 준모수적 방법

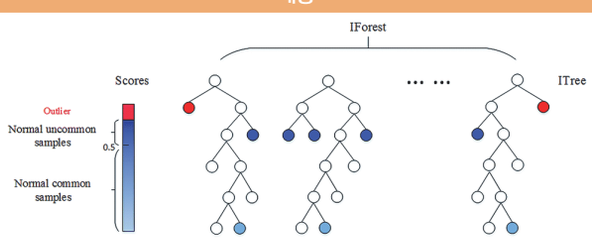
자료의 구조에 따른 이상값 탐색 방법의 분류

자료의 구조	이상값 탐색 방법
단변량 자료	• 표준화 점수 • 수정된 표준화 점수 • 통계적 가설검정 • 사분위수범위 • 수정된 사분위수범위 • 준사분위수범위
다변량 자료	• 회귀진단에서 이상값 탐색 • 마할라노비스 거리 • LOF(Local Outlier Factor) • iForest(isolation Forest)
시계열 자료	• Shewhart • 누적합(CUSUM) • 지수가중이동평균 • Hidiroglou-Berthelot

자료의 구조에 따른 이상값 탐색 방법의 분류

검토항목			
단변량 자료	– 시계열 자료가 아니고, 변수의 개수가 1개인가?		
	알고리즘		내용
	표준화 점수(Z-score)를 활용한 이상값 탐색		표준화 점수는 평균이 μ이고, 표준편차가 σ인 정규분포를 따르는 관측치들이 자료의 중심(평균)에서 얼마나 떨어져 있는지를 나타낸다. n개의 각 관측치에 대한 표준화 점수는 다음과 같다. $$Z_i = \frac{x_i - \mu}{\sigma}, i = 1, 2, 3, \cdots, n$$ 일반적으로 표준화 점수의 절댓값이 3보다 크면 이상값으로 정의하지만 절대적인 기준은 없으며, 경험에 근거하여 판단기준을 제시하는 것이 합리적이다.
	수정된 표준화 점수(Modified Z-score)를 활용한 이상값 탐색		표준화 점수는 평균과 표준편차에 의존하므로 산출 과정에 이상치의 영향을 받는 문제점이 있다. 수정된 표준화 점수는 표준화 점수의 문제점을 보완하고자 중앙값(*median,* $\tilde{x}$)과 중앙값의 절대편차MAD, median absolute deviation를 이용하여 산출한다. $$MAD = median(\left\| x_i - \tilde{x} \right\|), \tilde{x}\text{는 중앙값}$$ n개의 관측치에 대한 수정된 표준화 점수는 다음과 같다. $$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD}, i = 1, 2, 3. \cdots, n$$ 수정된 표준화 점수를 활용한 이상값 탐색 방법은 관측치의 수가 적을 때 적합한 방법으로 알려져 있다.

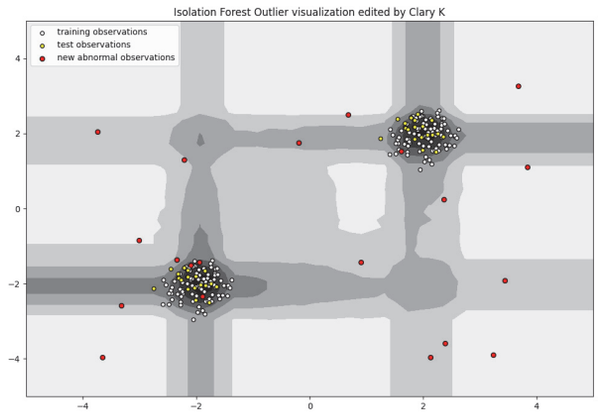
<table>
<tr><td rowspan="3">다변량 자료</td><td colspan="2">- 시계열 자료가 아니고, 변수의 개수가 2개 이상인가?</td></tr>
<tr><td>알고리즘</td><td>내용</td></tr>
<tr><td>사분위수 범위를 활용한 이상값 탐색</td><td> 

상자그림boxplot은 최솟값, 최댓값, 제1사분위수(Q_1), 제2사분위수(Q_2), 제3사분위수(Q_3)를 활용하여 데이터를 시각적으로 요약한 그래프이다.
상자그림에서 표현되는 최솟값과 최댓값은 이상값을 제외한 데이터의 최댓값과 최솟값을 의미하여, 이상값은 사분위수범위를 활용하여 정의한다.
사분위수범위는 제1사분위수(Q_1)와 제3사분위수(Q_3)의 차이로 정의되며, 사분위수 범위를 활용한 이상값 정의 수식은 아래와 같다.
$$(Q_1 - c \times IQR, Q_3 + c \times IQR), IQR = Q_3 - Q_1, c\text{는 상수}$$
일반적으로 상수 c는 1.5나 3을 적용하며, 사분위수범위의 1.5배를 초과하는 관측값은 약한 이상값, 3배를 초과하는 관측값은 강한 이상값으로 정의한다.</td></tr>
<tr><td>수정된 사분위수 범위를 활용한 이상값 탐색</td><td>수정된 사분위수범위를 활용한 이상값 정의 방법은 위의 상자그림을 활용한 방법을 일반화한 것으로, 비대칭 분포에서도 이상값을 합리적으로 정의할 수 있다.
수정된 사분위수범위는 데이터의 치우침 정도에 대한 척도인 왜도skewness를 활용한다.
왜도에 대한 강건한 통계량인 medcouple(MC)을 적용하여 이상값을 정의하며 수식은 아래와 같다.
$$MC = median\{h(x_{(i)}, x_{(j)})\}, x_{(i)} \le \tilde{x} \le x_{(j)}$$
- $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$, $\tilde{x}$는 중앙값
- $h(x_{(i)}, x_{(j)}) = \frac{(x_{(j)} - \tilde{x}) - (\tilde{x} - x_{(i)})}{x_{(j)} - x_{(i)}}$
수정된 사분위수범위를 활용한 이상값 정의 수식은 아래와 같다.
$$(Q_1 - 1.5e^{-4MC} IQR, Q_3 + 1.5e^{3MC} IQR), \text{if } MC \ge 0$$
$$(Q_1 - 1.5e^{-3MC} IQR, Q_3 + 1.5e^{4MC} IQR), \text{if } MC < 0$$</td></tr>
<tr><td>iForest (Isolation Forest)</td><td> 

iForest 기법은 관측치 사이의 거리 또는 밀도에 의존하지 않고, 데이터마이닝 기법인 의사결정나무decision tree를 이용하여 이상값을 탐지하는 방법이다.
의사결정나무 기법으로 분류모형을 생성하여 모든 관측값을 고립시켜 나가면서 분할 횟수로 이상값을 탐색한다. 즉, 데이터의 평균적인 관측값과 멀리 떨어진 관측값일수록 적은 횟수의 공간 분할로 고립시킬 수 있다.</td></tr>
</table>

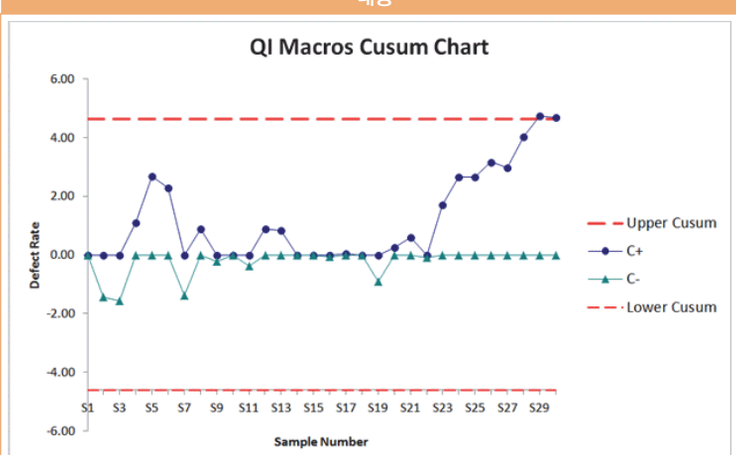
의사결정나무 모형에서 적은 횟수로 leaf 노드에 도달하는 관측값일수록 이상값일 가능성이 크다.

```
# 파이썬 코드 예시

import numpy as np
import matplotlib.pyplot as plt
from sklearn.ensemble import IsolationForest

# 랜덤 스테이트 수치 설정
state_value = np.random.RandomState(20)

# 학습 데이터 생성
X = 0.3 * state_value.randn(100, 2)
X_train = np.r_[X + 2, X - 2]

# 정규값과 비슷한 데이터 추가
X = 0.3 * state_value.randn(20, 2)
X_test = np.r_[X + 2, X - 2]

# 정규적이지 않은 데이터 추가(인위적인 이상값 구현을 위함)
X_outliers = state_value.uniform(low=-4, high=4, size=(20, 2))

# IsolationForest 모델 만들기
isol = IsolationForest(max_samples=100, random_state=state_value)

# 만들어 둔 데이터셋에 알고리즘 피팅
isol.fit(X_train)
y_pred_train = isol.predict(X_train)
y_pred_test = isol.predict(X_test)
y_pred_outliers = isol.predict(X_outliers)

# 샘플 데이터 및 가까운 벡터를 평면에 플로팅
xx, yy = np.meshgrid(np.linspace(-5, 5, 50), np.linspace(-5, 5, 50))
Z = isol.decision_function(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)

# 캔버스 사이즈 정의 및 타이틀 작성
plt.figure(figsize=(12, 8))
plt.title("Isolation Forest Outlier visualization edited by Clary K")

# 학습 데이터, 테스트 데이터, 결측값 데이터 각각 스캐터플로팅
plt.contourf(xx, yy, Z, cmap=plt.cm.binary)
ob1 = plt.scatter(X_train[:, 0], X_train[:, 1], c="white", s=20, edgecolor="k")
ob2 = plt.scatter(X_test[:, 0], X_test[:, 1], c="yellow", s=20, edgecolor="k")
ab = plt.scatter(X_outliers[:, 0], X_outliers[:, 1], c="red", s=30, edgecolor="k")

# X, y축 범위 지정 및 범례 지정
plt.axis("tight")
plt.xlim((-5, 5))
plt.ylim((-5, 5))
plt.legend(
    [ob1, ob2, ab],
    ["training observations", "test observations", "new abnormal observations"],
    loc="upper left",
)
plt.show()
```



– 시계열 자료인가?

알고리즘	내용
시계열 자료	누적합 (CUSUM, Cumulative sum) 관리도



누적합 관리도는 처음부터 현재까지 통계량의 누적합을 사용하는 방법으로, 작은 변화에 대한 효과가 누적된 통계량을 통해 추세의 작은 변화를 감지하는 데 유용하다.

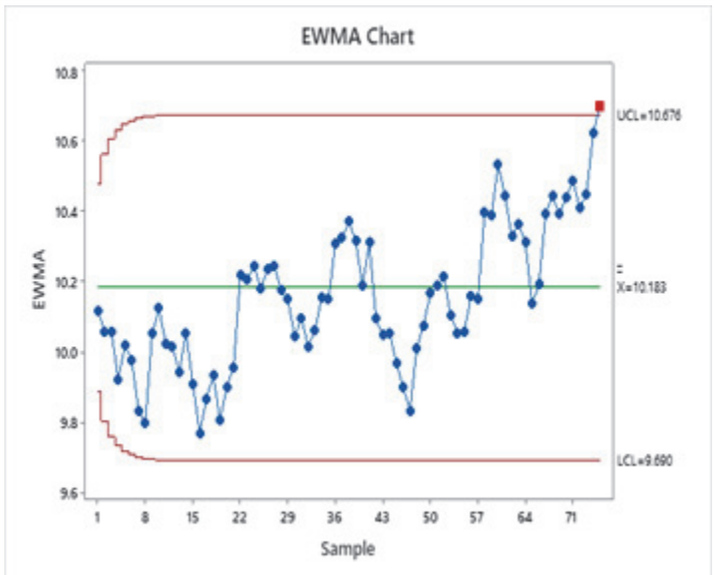
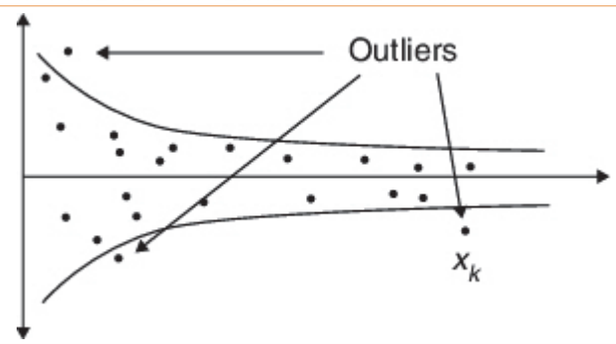
누적합 관리도는 통계량 C_i^+, C_i^- 이 의사결정 구간 H를 벗어나는 시점을 이상값으로 정의하며, 누적합 통계량에 대한 수식은 아래와 같다.

$$C_i^+ = \max\left[0, x_i - (\mu_0 + K) + C_{i-1}^+\right]$$

$$C_i^- = \max\left[0, (\mu_0 - K) - x_i + C_{i-1}^-\right]$$

의사결정구간 $H = h\sigma$, 허용값 $K = k\sigma$로 정의된다. σ는 표준편차를 의미하여 h는 4 또는 5, k는 0.5를 적용한다. μ_0는 목표값으로 관측값들이 수렴해야 할 이상적인 수치를 의미한다.

누적합 방법은 작은 변화에 대한 이상값을 탐지하므로, 변이가 큰 자료보다 안정적인 자료에 적용하기에 적합하다.

지수가중 이동평균(Exponetially Weighted Moving Average, EWMA) 방법	 지수가중이동평균 방법은 최근 관측값에 큰 가중치를 주어 최근 변화를 반영하여 이상값을 탐지하는 방법이다. 지수가중이동평균($EWMA_t$)은 아래와 같이 산출할 수 있다. $$EWMA_t = \lambda Y_t + (1-\lambda)EWMA_{t-1}, t = 1,2,\cdots,n, 0 < \lambda < 1$$ $EWMA_0$는 목푯값, Y_t는 t시점의 관측값, λ는 가중치이다. Hunter는 0.2와 0.3 사이의 값으로 가중치 λ를 정의하는 것을 제안하였다. 지수가중이동평균에 대한 관리 한계는 아래와 같이 정의된다. $$LCL = EWMA_0 - k\sqrt{\frac{\lambda}{2-\lambda}s^2}$$ $$UCL = EWMA_0 + k\sqrt{\frac{\lambda}{2-\lambda}s^2}$$ s는 표준편차, k는 상수이며, 일반적으로 상수 $k=3$을 사용한다. 하지만, 사용자의 경험적 판단에 의해 결정하는 것이 가장 합리적이다.
Hidiroglou-Berthelot 방법	 Hidiroglou-Berthelot(H-B) 방법은 다른 시계열 자료의 이상값 탐색 방법과 다르게, 이전 시점과 현재 시점의 비율로 이상값을 탐지하는 방법이다. 보통, 정기적으로 수행되는 경기동향조사에서 활용되며, 단위의 크기size of unit를 고려하여 이상값에 대한 허용범위를 정의하는 방법이다. Hidiroglou와 Berthelot이 제안한 이상값 정의 방법은 아래와 같은 절차로 수행한다. ① 이전 시점($t-1$) 관측값에 대한 현지 시점(t) 관측값의 비(r_i) 산출

$$r_i = x_i(t)/x_i(t-1)$$

② 0을 중심으로 대칭 분포하는 s_i로 관측값의 비(r_i) 변환

$$s_i = \begin{cases} 1-\frac{r_m}{r_i}, 0 < r_i < r_m \\ \frac{r_i}{r_m}-1, r_i \geq r_m \end{cases}, r_m \text{은 } r_i \text{의 중앙값}$$

③ 앞서 산출한 s_i와 관측값에 대해 E_i 산출

$$E_i = s_i \times Max\left[x_i(t), x_i(t-1)\right]^{\mu}, 0 \leq \mu \leq 1$$

μ는 단위의 크기가 E_i에 미치는 영향력을 조절한다. μ가 0이면, 단위의 크기와 상관없이 E_i는 동일한 값을 가지며, μ가 1이면, 단위의 크기가 E_i에 큰 영향을 준다.

④ E_i에 대해 다음과 같이 d_{Q_1}, d_{Q_3}를 산출한다.

$$D_{Q_1} = Max(E_M - E_{Q_1} | AE_M|)$$
$$D_{Q_3} = Max(E_{Q_3} - E_M | AE_M|)$$

E_M은 E_i의 중앙값, E_{Q_1}은 E_i의 제1사분위수, E_{Q_3}은 제3사분위수를 의미하고, A는 상숫값이며 일반적으로 0.05를 적용한다.

⑤ 이상값 정의 기준은 다음과 같다.

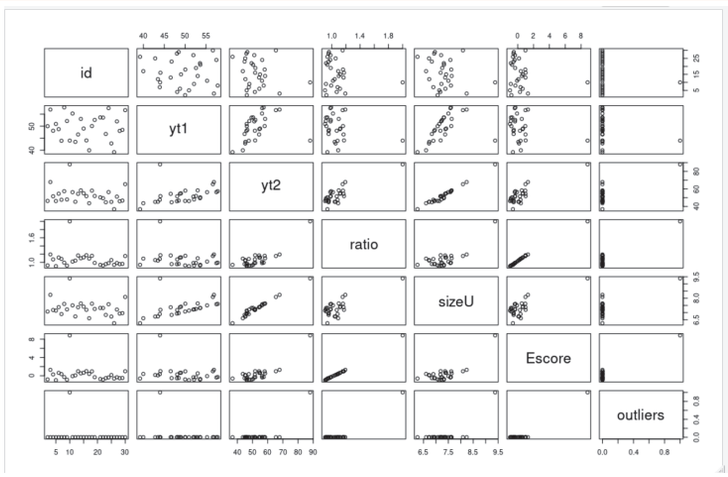
$$(E_M - Cl_{Q_1}, E_M + Cl_{Q_3})$$

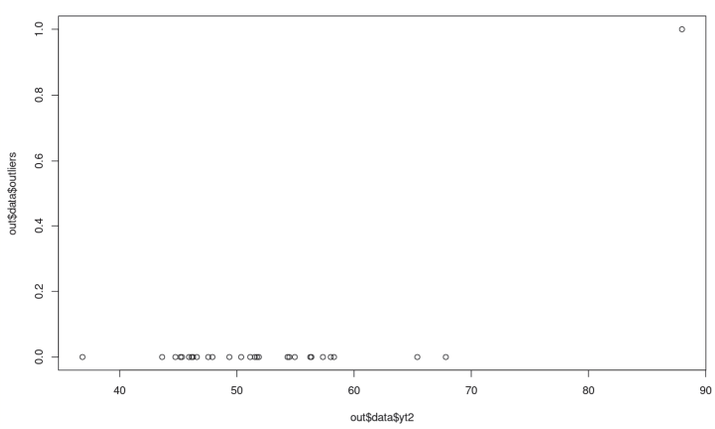
C는 정상 범위를 결정하는 상숫값이며, 일반적으로 6을 적용한다.

Hidiroglou-Berthelot 방법은 비율에 대해 사분위 범위를 적용하여 제안한 방법으로 통계적 가정이 필요하지 않으나, 결정해야 할 상수가 많고, 산출 과정이 복잡하다.

```
# R H-B 방법 코드 예시

install.packages("univOutl")
library("univOutl")

# 평균이 50이고, 표준편차가 5인 정규분포 난수값 30개 x0 벡터 생성
set.seed(222)
x0 <- rnorm(30, 50, 5)
# x0 벡터 1번에 NA 할당
x0[1] <- NA
# 난수 시드 값 재설정
set.seed(333)
# 최소 0.9, 최대 1.2인 유니폼 분포 난수값 30개 rr 벡터 생성
rr <- runif(30, 0.9, 1.2)
# rr 벡터 10번에 2 값 할당 (강제 이상값 생성 위함)
rr[10] <- 2
# x0 벡터에 rr 벡터를 곱하여 x1 벡터 생성
x1 <- x0 * rr
# x1 벡터 20번에 0 값 할당 (강제 이상값 생성 위함)
x1[20] <- 0

# 두 개의 다른 시간에 주기적 데이터의 이상값 감지
out <- HBmethod(yt1 = x0, yt2 = x1)
# 제외된 식별자 표시 [NA or 0]
out$excluded
# 비율의 중앙값 [ratio -> yt1/yt2]
out$median.r
# E-점수를 사용하여 이상값을 찾기 위해 식별된 상한 및 하한.
out$bounds.E

# E-점수가 한도를 벗어나면 이상값으로 간주
0out$outliers
```

```
cbind(x0[out$outliers], x1[out$outliers])
out <- HBmethod(yt1 = x0, yt2 = x1,
                return.dataframe = TRUE)
out$excluded
head(out$data)
```

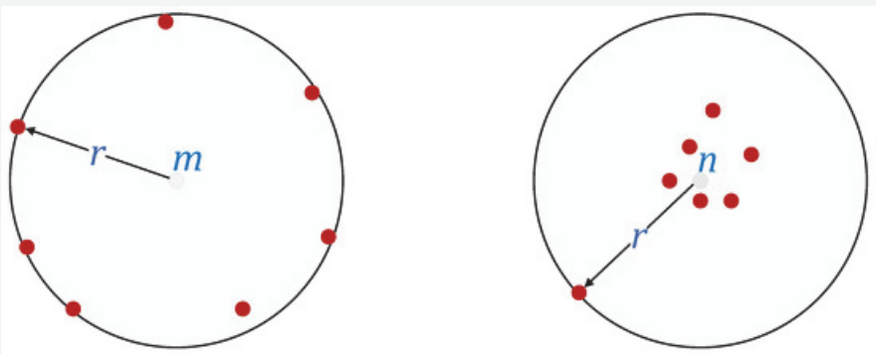




참고 각 병원의 기계학습 기반 의료 품질 분석을 위한 이상값 분석 사례[15]

- 전통적인 통계 기반 평가 방법 및 기계학습 기반 기술을 통합한 사우디아라비아 지역의 의료 데이터 기반 의료 품질평가 모델의 이상값 분석 방법 사례
  - 사우디아라비아의 의료 빅데이터를 바탕으로 관리 결함, 안전 위험, 보험 사기 및 과도한 치료 등 불법 행위가 빈번한 병원을 판별하는 의료 품질의 평가 솔루션 개발이 목적임
  - 이때, 전체 병원을 대상으로 전통적인 통계 기반의 이상값 탐지 기법을 적용할 때, 너무 많은 이상값이 발생한 것으로 잘못 해석될 수 있는 문제 발생
  - 또한, 주변과 유사한 의료 서비스를 제공하는데도 낮은 품질의 서비스를 제공하는 것으로 저평가될 수 있어, 이를 보완한 평가 방법이 필요하게 됨
- 문제 해결을 위한 개선된 KNN 기반 이상값 탐지 방법의 연구 개발 및 적용

- 기존의 통계 기반 평가 방법을 벗어나는 것은 비현실적이므로 통계를 기반으로 이상값 감지 알고리즘을 개선함
- 최근 몇 년 동안 이상값 감지 분야에서 많은 관심을 끌고 있는 KNN[K-Nearest Neighbor]방법을 도입, 주로 인근 이웃의 이상치 정도와 비교하여 평가함

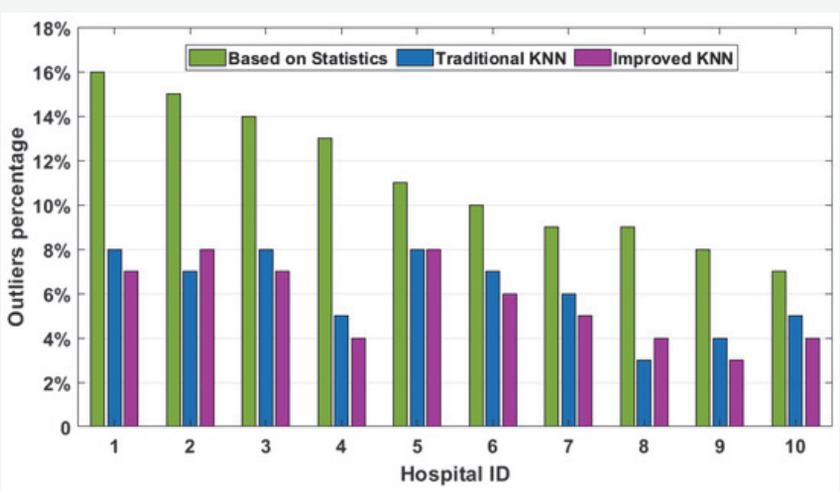


KNN 방법 기반 이상값 탐지

(k=7일 때, m, n 그룹은 동일한 포인트 값을 가지지만, 각 그룹 내 이상값 정도는 완전히 다른 의미를 가짐)

- 통계 기반 이상값 탐지 알고리즘을 사용하여, 통계와 KNN을 기반으로 한 비교 조건이 동일한 차원인지를 확인하는 등 향상된 KNN 기반 이상값 감지 알고리즘을 제안 및 적용함

• 개선된 KNN 기반 이상값 탐지 결과 비교



- 전통적인 통계 기반 이상값 탐지 기반 검출 결과(녹색) 상위 10개 병원을 대상으로 KNN(파란색) 또는 개선된 KNN 기반 이상값 탐지 기법(자주색)을 적용하여 비교 분석
- 전통적인 통계 기반 방법에서 검출률이 높게 나왔더라도, KNN 기반의 분석 방법은 상대적으로 다른 병원과 평균적인 서비스를 제공하는 것으로 이해할 수 있음
- 특히, KNN 기반의 이상값 분석 방법을 토대로 의료 품질을 분석한 결과 사우디아라비아의 국민 건강 정보 센터의 의료 품질 순위의 분석 결과와 동일한 결과를 얻음

05-2 데이터 공격에 대한 방어 수단을 강구하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

학습 데이터의 수집, 구축 및 보관 과정에서 위변조 등을 통한 데이터 중독, 회피 공격이 예상되는 경우 본 항목을 고려하여 해당 여부를 판단하십시오.

- 인공지능 서비스 개발 또는 운영 과정에서 의도적으로 학습 데이터를 변질시키거나 입력 데이터에 최소한의 변조를 가해 예상과는 다른 결과를 출력하는 공격에 노출될 수 있으므로, 이에 대처할 방안을 검토 및 적용하는 것이 바람직하다.
- 의료기기 및 소프트웨어에 대한 적대적 데이터 공격의 실제 사례는 아직 많지 않지만, 이론적 가능성은 지속해 제기되고 있다. 데이터 공격 시도와 그에 대한 방어 기법이 꾸준히 연구되고 있으므로, 최신 연구 동향을 파악하여 데이터 공격에 대한 방어 수단을 세워야 한다.
- 현재까지 다양한 연구를 통해 흉부 X선 영상, 안저 영상, 피부 영상 등에 대한 적대적 공격이 의료보험 부정수급이나 임상시험을 단축하고자 악용될 가능성이 이론적으로 검토되고 있다. 또한, 의료 데이터에 대한 다양한 중독과 회피 공격 중 HopSkipJump, FGSM[Fast Gradient Sign Method], Carlini & Wagner, Crafting Decision Tree, Zeroth Order Optimization 등이 이론적으로 가능하므로 주의하여야 한다 [16].

데이터 공격 및 방어 기법 예시

공격 기법 분류	공격 기법 내용	대표 방어 기법
데이터 중독 공격	• 인공지능 서비스는 일반적으로 입력 데이터 분포의 변화에 적응하기 위해 모델 배치 후 수집된 새로운 데이터를 사용해 재교육된다 (예: 침입 감지 시스템). 이때, 공격자는 세심하게 조작된[perturbed] 데이터를 주입하여 서비스의 정상적인 기능을 손상하는 방식으로 학습 데이터를 오염시킬 수 있다. • 종류: FGSM, HopSkipJumpAttack	• 적대적 학습 adversarial training • Gradient Masking (Distillation) • Feature Squeezing
회피 공격	• 공격자는 학습 모델이 입력을 올바르게 식별할 수 없도록 기존의 입력 데이터에 대해 미묘한 차이의 노이즈를 추가하여 조작된 입력 데이터를 생성한다. 이러한 변화는 사람의 눈에 잘 띄지 않지만, 심층학습 모델의 출력에 큰 영향을 미친다. • 종류: Crafting Decision Tree, Carlini & Wagner, Zeroth Order Optimization	

05-2a 데이터 중독poisoning, 회피evasion 등 공격에 대한 방어 대책을 마련하였는가?

Yes ☐ No ☐ N/A ☐

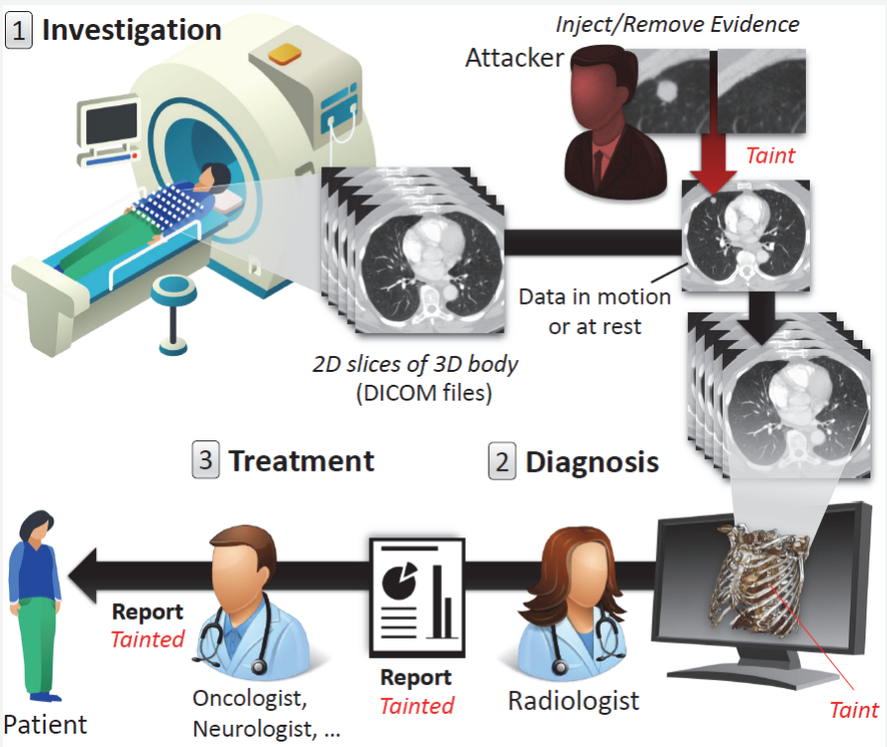
- 의료 분야에서는 외부 공격자가 의료 영상 이미지 데이터에 접근하여 콘텐츠를 변경하는 등 데이터 공격으로 인해 오진이 유발될 가능성이 있다[17]. 병원 내 의료 데이터가 가진 취약점의 예는 다음과 같다.
  - ✓ 외부 인터넷에 노출된 방사선 의료 기기 네트워크
  - ✓ 의료(헬스케어) 산업에서 보유한 빈약한 보안 추적 기록
  - ✓ 인터넷 네트워크 보안의 부재(오래된 소프트웨어/운영체제, 없거나 적합하지 않은 암호화, 외부로 노출된 인프라 등)
- 따라서, 병원을 통해 수집하는 인공지능 학습 데이터 뿐 아니라, 내부에서 수집 및 구축한 학습용 데이터를 대상으로도 공격자의 악의적인 적대적 공격이 발생할 수 있으므로, 이에 대한 대책의 마련이 필요하다.
- 적대적 공격을 방어하고 인공지능 서비스의 강건성을 높이기 위한 다양한 방어 기법이 존재한다. 특히 데이터 설계 및 모델 학습 단계에서의 회피 공격과 중독 공격 방어를 위한 대표적 기법으로는 적대적 학습, Gradient Masking, Feature Squeezing 등이 있으며, 의료 데이터 분야에도 역시 적용될 수 있다.

적대적 공격에 대한 방어 기법 예시

방어 기법 분류	방어 기법 내용
적대적 학습	• 가장 직관적으로 쉽게 떠올려볼 수 있는 알고리즘이다. 모델을 학습시킬 때, 적대적 사례로 활용할 수 있는 모든 경우의 수를 미리 고려하여 학습 데이터셋에 포함하는 것이다. 그러나 충분한 수와 다양성이 보장된 적대적 데이터를 생성하는 과정 없이는 적대적 학습은 그 성능을 보장하기 어렵다.
Gradient Masking /Distillation	• 대부분의 적대적 공격은 모델 추론 과정에서의 경사gradient를 보고 공격이 이루어진다. 그러므로 학습 모델의 경사가 그대로 노출되는 것을 방지하거나gradient masking, 정규화 방법 등을 통해 경사가 두드러지지 않게 하여 적대적 공격에 방어할 수 있는 방법distillation들이 제안되었다.
Feature Squeezing	• 적대적 공격을 막기 위한 방법으로는 본래의 학습 모델과 별도로, 주어진 입력이 적대적 사례인지 아닌지를 판단하는 학습 모델을 추가하는 방법이 있다. 그 외에 다수의 학습 모델을 조합하여 시스템을 구성하면 특정 모델에 대한 화이트박스 공격을 피할 수 있으며, 특정 모델에 적용되는 적대적 공격이 불가능해지므로 좋은 방어법이 된다는 연구 결과가 제시되었다.

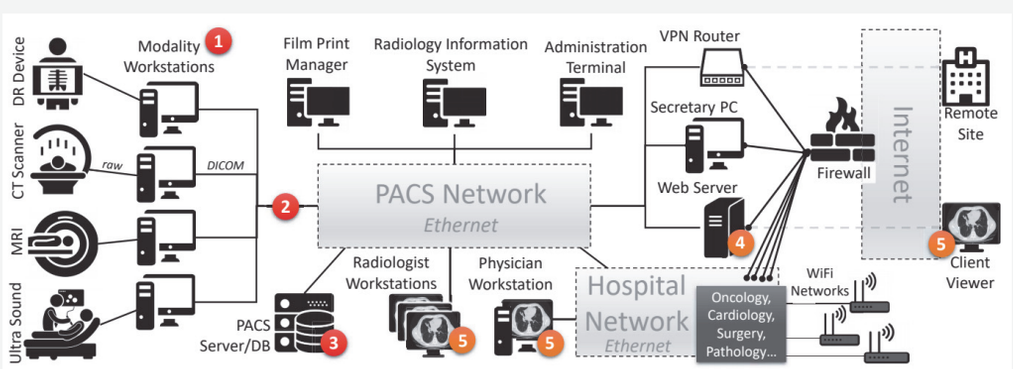
참고 CT 이미지를 대상으로 공격자가 데이터를 중독시키는 시나리오 연구 사례[17]

- 의료 CT 촬영 데이터를 대상으로 하는 데이터 위변조 공격 시나리오



데이터 중독 공격 시점

(조사(1)와 진단(2) 사이에서 의료 이미지를 변조하여 방사선 전문의와 의사 모두 공격자가 설정한 오류를 믿음)

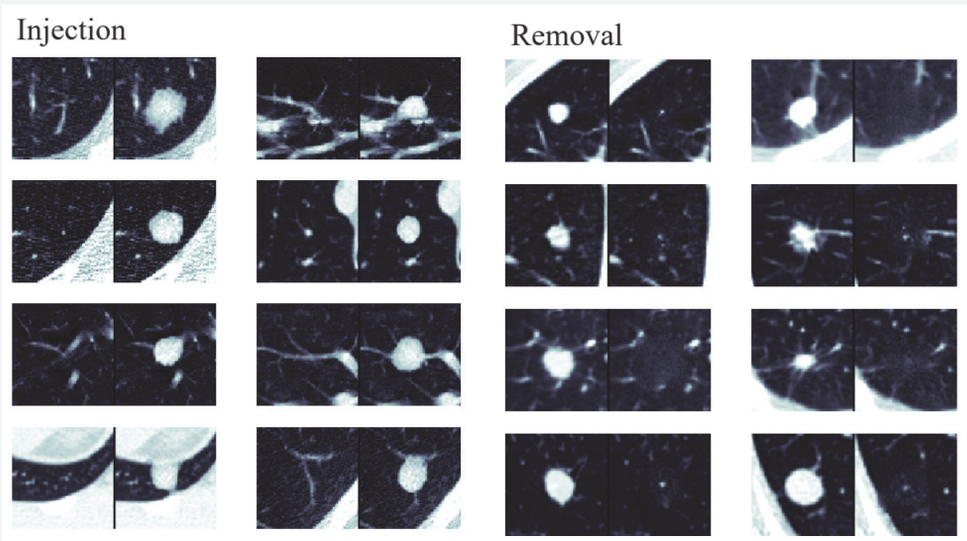


데이터 중독 공격 위치

(병원이 PACS 네트워크 사용 시, 공격자는 1~3 위치에서 모든 스캔을 변조할 수 있음. 공격자는 4~5 지점에서 스캔의 하위 집합을 조작할 수 있음)

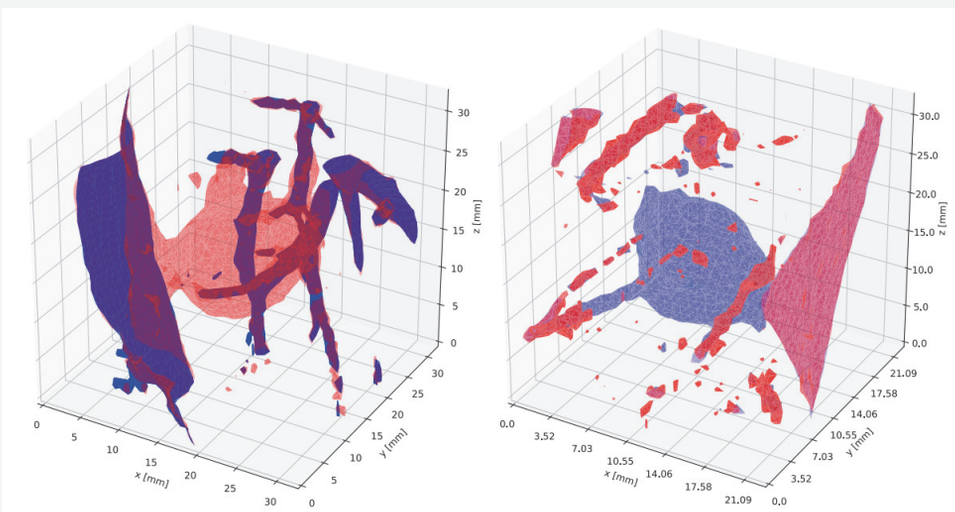
- 의료 CT 촬영 데이터 공격 시나리오
  - PACS 네트워크를 사용하는 병원을 가정하여 공격자는 방사선 전문가가 진단을 수행하기 전에 저장 데이터 또는 이동 데이터를 대상으로 CT 스캔 내용을 변경할 수 있음
  - 이동 데이터는 PACS 서버 또는 방사선 전문의의 개인 컴퓨터에 저장된 DICOM[Digital Imaging and Communications in Medicine] 파일을 나타냄
  - 때에 따라 DICOM 파일을 DVD에 저장한 다음 환자 또는 외부 의사가 병원으로 전송함

• CT-GAN 기법을 활용하여 의료 CT 촬영 데이터 공격



데이터 중독(변조) 공격 결과

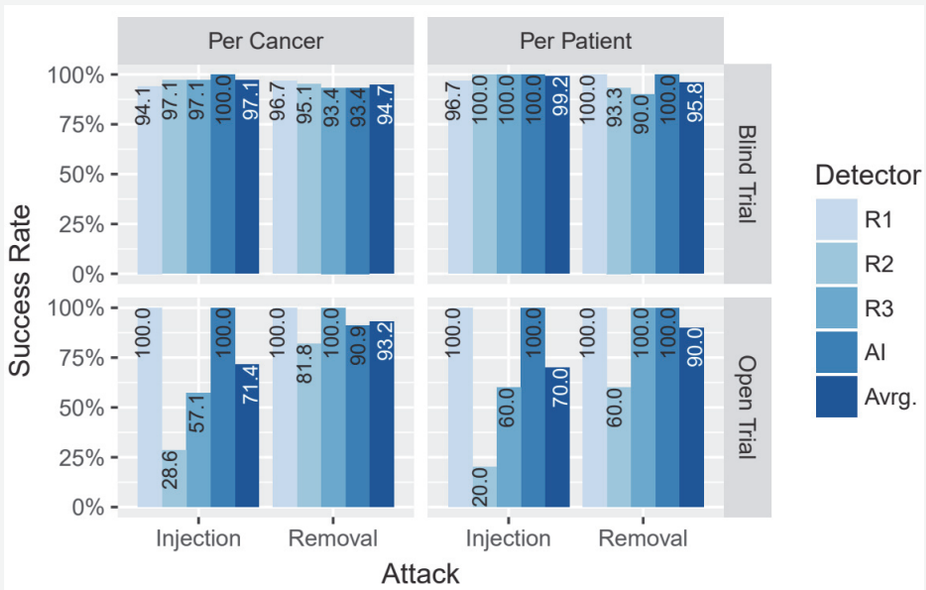
(샘플 주입(왼쪽) 및 제거(오른쪽), 각 이미지에 대해 왼쪽이 변조 전, 오른쪽이 변조 후 이미지임)



데이터 중독(변조) 공격 결과

(CT 스캔한 3D 샘플에 대해서 주입(왼쪽), 제거(오른쪽), 변조 이전(파란색), 변조 이후(붉은색) 결과 비교)

- 1) 2년, 5년, 7년의 경험이 있는 3명의 방사선 전문의를 모집하고, 80개의 완전한 폐 CT 스캔을 진단하도록 한 뒤, 공격(변조) 여부는 모른 채 스캔 결과를 분석하도록 요청
- 2) 공개적으로 진행한 결과에서는 공격(변조)에 대한 정보를 제공하고, CT 스캔 이미지에서 가짜, 실제 및 제거된 결절을 식별하도록 요청



데이터 중독(변조) 공격 성공 결과

((위) 1)의 블라인드 테스트 결과 공격 성공률, 암 주입은 평균 99.2%, 암 제거는 평균 95.8% 평균 성공률을 보였으며, 해당 데이터에서 인공지능 모델은 중독된 데이터를 구분하지 못함, (아래) 공격 정보를 제공했을 때 공격 성공률)

다양성 존중 | 책임성 | 투명성

요구사항 06 수집 및 가공된 학습 데이터의 편향 제거

대표행위자 | 데이터 공급자 협력 대상 | 데이터 과학자 | 전문 의료진 | 인공지능 모델 개발자

- 학습에 필요한 데이터 수집 및 가공 시 발생할 수 있는 편향을 인식하고 이를 제거하는 방안을 적용한다. 주로, 데이터 수집 시 발생할 수 있는 편향을 확인해야 하며, 학습을 위한 특성을 선택하거나 데이터 라벨링 및 샘플링 시에도 편향이 발생할 수 있으므로 제거 방안을 마련한다. 단, 이미 편향성 검토가 완료된 데이터를 활용하거나, 의료 분야 특성상 현실적으로 모든 환자군의 데이터를 검증하기 어려울 때는 샘플링 기법 등을 통해 데이터를 검증한다.

06-1 데이터 수집 시, 인적·물리적 요인으로 인한 편향 완화 방안을 마련하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

인공지능 학습용 데이터셋을 직접 수집, 구축하는 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 인적 요인으로 인한 편향은 사람이 의식적 혹은 무의식적으로 특정 정보에 대해 편향되는 점에서 기인한다.
  - ✓ 인적 편향: 자동화 편향, 그룹 귀인 편향group attribution bias, 암묵적 편향implicit bias, 그룹 내 편향in-group bias 등이 포함됨
- 인적 편향을 방지하도록 데이터 수집 시 명확한 수집 및 검수 기준을 수립하여 수집하는 작업자별로 데이터 특성이 편향되지 않도록 방지하거나, 다양하고 충분한 수의 검수자를 확보함으로써 검수 시에 편향을 바로잡아야 한다.
- 물리적 편향은 의료 데이터를 수집하는 장비상의 문제로 인해 발생할 수 있다. 예를 들어, X-Ray 촬영 장비마다 설정값이 달라 각 환자에 대한 특정 색상, 밝기, 해상도 등 물리적으로 한정된 이미지 데이터가 수집될 수 있다. 따라서, 특정 장비로 수집된 데이터셋을 사용하여 알고리즘을 학습시켰더라도 추후 다른 장비로 수집된 데이터셋을 적용할 시에는 오류가 발생할 수 있다.
- 이에 따라 특정 인종이나 특정 장비에 편향된 학습이 이루어질 수 있으므로, 가능한 여러 지역과 인종을 대상으로 다양한 의료기기를 활용하여 각종 인적·물리적 편향 요인을 제거하고 다양성을 보완하는 것이 바람직하다.

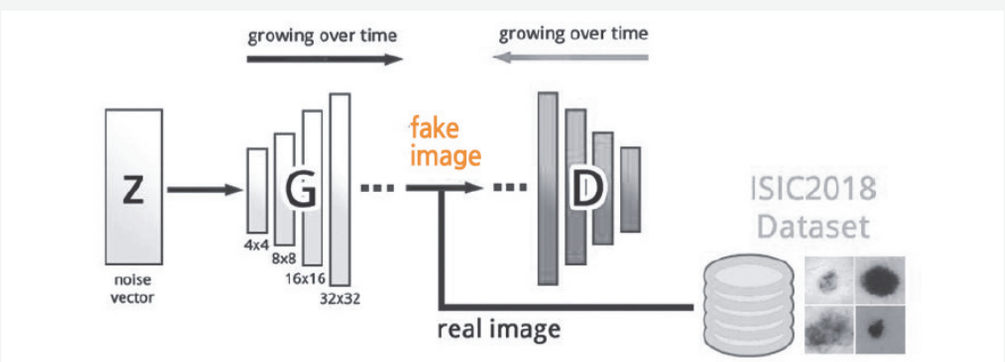
06-1a 인적 편향을 제거하기 위한 절차적, 기술적 수단을 적용하였는가?

Yes No N/A
☐ ☐ ☐

- 의료 기관에서 보유한 한정된 데이터로는 실제 제품화하는 데 필요한 데이터를 취득하기 어려운 실정이며, 수집된 데이터를 활용하여 목적에 맞는 데이터셋을 구성해야 한다. 이때, 질병의 특성으로 인해 데이터 자체가 희귀하거나 특정 그룹에 편향된 데이터 불균형 등 인적 편향 문제에 직면할 수 있으므로, 이를 완화하는 수단을 고려해야 한다.
- 편향된 실제 데이터를 기반으로 합성 데이터를 생성하여 데이터의 다양성을 확보하고 편향을 완화할 수 있다. 그러나, 작업자에 의해 왜곡된 합성 데이터는 실제 데이터와는 완전히 다른 결과를 도출할 수 있으며, 작업자 간 개인별 편차로 인적 편향이 발생할 수 있다. 이를 줄이기 위해 데이터 수집 작업 가이드라인을 마련하고, 다양한 작업자를 모집하여 특정 배경과 성향을 배제하고, 합성데이터의 품질을 확보하여야 한다.

참고 합성 데이터 생성 기법(GAN) 활용 사례[18]

- 피부 이미지 데이터는 피부 유형 간 데이터양의 차이가 존재하므로 이를 해결하고자 양성 및 악성 피부병변의 이미지 10,000개의 데이터셋을 대상으로 실제 데이터와 유사한 고해상도의 합성 데이터를 생성했다.
- 생성된 이미지를 검증하고자 피부과 전문의 및 심층학습 전문가로 구성된 팀이 생성된 이미지와 실제 이미지를 혼합한 데이터셋을 재분류하는 Visual Turing Test를 진행하였으며, 검증 결과 생성 데이터가 실제 데이터와 구분이 어렵다는 결과를 도출했다.



참고 합성 데이터의 품질 확인 사례[19]

- 오픈소스로 공개된 합성 데이터 생성기 Synthea에서 생성한 120만 명의 매사추세츠 환자 코호트를 대상으로 대장암 검진, 만성 폐쇄성 폐질환(COPD) 30일 사망률, 고관절/무릎 치환술 후 합병증 발생률, 고혈압 관리 등 4가지 지표를 대상으로 데이터의 품질을 확인했다.
- 합성된 데이터의 통계와 매사추세츠 환자, 미국 전역 환자의 실제 데이터를 기반으로 공개된 값을 비교한 결과 대장암 검진 지표에 대해서는 합성 데이터와 실제 데이터 간 유사한 비율을 나타냈으며, 데이터의 신뢰성이 높은 것으로 확인했다.
- 이 외 지표에 대해서는 합성 데이터와 실제 데이터 간 차이가 존재했다. 이를 해결하려면 합성 데이터가 실제 데이터의 특성을 반영할 수 있도록 지속해서 개선해야 하며, 활용하기 이전에 합성 데이터의 품질을 검증하는 절차가 필요하다.

Measure	SyntheticMass Rate	Massachusetts Rate	National Rate
Colorectal Cancer Screening	68.7% (68.5, 68.9%) (215,919/314,355)	77.3%	69.8%[a]
COPD 30-Day Mortality: Strict	0.7% (0, 1.7%) (3/409)	7.0% (1233/17636)	8.0%
COPD 30-Day Mortality: Expanded	4.7% (4.6, 4.8%) (8612/181,458)	7.0% (1233/17636)	8.0%
Complications of Hip/Knee Replacement	0% (0/207)	2.9% (700/23949)	2.8%
Controlling High Blood Pressure	0% (0/241,311)	74.52%	69.7[a]%

06-1b 데이터의 다양성 확보를 위해 이기종 수집 장치를 활용하였는가?

Yes ☐ No ☐ N/A ☐

- 의료 데이터는 X-Ray, CT, MRI 등의 기기를 이용해 영상 데이터를 확보하지만, 데이터의 다양성이나 가외성* 측면을 보강하고자 같은 종류의 데이터를 2대 이상의 장비로 촬영할 때는 드물다.

  * 같은 기능이 여러 기관에서 개별적으로 이루어지는 중복성을 아울러 이르는 말

- 그런데도 학습 단계 시 사용한 장비와는 다른 장비에서 얻은 데이터를 통해 진단을 수행할 때 오류가 급격히 증가하였다는 사례도 있으므로, 하드웨어로 인한 데이터 편향이 발생하지 않도록 주의하여야 한다.
- 또한, 디지털 헬스케어 산업이 성장하며 의료 마이데이터 정책이 추진되고 있으며, 여러 의료 기관에 흩어진 데이터를 관리하는 빅데이터 사업이 진행되고 있다. 이와 같은 통합 데이터를 활용하여 인공지능 시스템의 정확도를 개선한 사례도 나타나고 있다.

06-1c 하드웨어로 인해 발생할 수 있는 데이터의 편향을 점검하였는가?

Yes ☐ No ☐ N/A ☐

- 진단에 필요한 온도계, 산소 농도계 등 의료 데이터 수집 장치에 대한 인종적 편향 문제는 지속해서 제기되어 왔으며, 이는 하드웨어 장비 설계 시 인종적 격차를 고려하지 않은 사실을 나타낸다.

참고 의료 데이터 수집 장치에 대한 인종적 편향 사례[20,21]

- 에모리 대학교의 최근 연구에서 연구원들은 흑인 환자의 체온을 측정하기 위해 100건 이상의 임상 연구에서 테스트된 이마 온도계를 사용했다. 연구자들은 구강 체온계를 사용했을 때보다 이마 체온계를 사용하는 흑인 환자에게서 열을 감지할 확률이 26% 낮음을 발견했다.
- 미시간 대학병원의 연구원들은 맥박 산소 측정기가 피부가 어두운 환자의 혈중 산소 수치를 과대평가하여 필요한 치료를 지연시키고 환자를 위험에 빠뜨릴 수 있음을 발견했다. 연구에 따르면 맥박 산소 측정기는 흑인 환자의 혈액에서 비정상적으로 낮은 산소 농도를 감지할 가능성이 3배 낮다.
- Seyyed-Kalantari와 동료들은 모델이 흉부 X-Ray 자체에만 접근할 수 있을 때도 인공지능 모델이 인종 및 기타 인구통계 그룹에 걸쳐 자동화된 흉부 X-Ray 진단의 정확도에 상당한 차이가 있음을 보여주었다. 이러한 모델을 사용하면 백인 남성 환자보다 흑인 여성 환자가 건강한 것으로 잘못 식별되는 환자가 더 많아질 수 있다. 더욱이, 인종적 격차는 단순히 학습 데이터에서 이러한 환자 그룹의 과소 대표로 인한 것이 아니며, 그룹 구성원과 인종적 격차 사이에 통계적으로 유의미한 상관관계가 존재하지 않았다.

- 장비의 기술적인 한계를 극복하는 절차를 마련할 수도 있지만, 과거 비윤리적인 피해를 일으킨 원인이 되었으며, 기기적인 편향을 해결하는 과정에서는 의료기기와 환자의 상호작용을 고려하여 윤리적인 피해가 발생하지 않도록 주의해야 한다.

참고 의료 데이터 수집 장치의 인종적 편향 해결 과정에서의 윤리적 피해 사례[21]

- 피부색이 어두운(검은색 또는 갈색) 환자의 경우 더 정확한 방사선 사진을 얻고자 방사선 조정이 권장되었다. 인종적인 차별뿐만 아니라 매우 비만하거나 근육질인 사람들과 특정 질환(경화증, 골수염, 파제트병)을 앓는 환자에게 적용되었다.
- 흉부 X-Ray 촬영 과정에서 사람이 받는 방사선의 양은 10일간의 자연 노출과 비슷하다. X-Ray 방사선량이 40~60% 증가하여도 사람의 생명에는 거의 영향을 미치지 않는다. 하지만 누적 효과는 건강 상태에 영향을 줄 수 있으며, 이러한 조정은 연구에 근거하여 위해 정도를 분명하게 밝혀야 한다.

- 대표적으로 인종적 차별이 발생하는 데이터는 흉부 X-Ray, 사지 X-Ray, 흉부 CT 및 유방조영술 등의 이미지 데이터로 데이터 정제 과정에서도 하드웨어 편향을 완화하는 계획을 수립해야 하며, 편향을 완화하기 위한 시도가 이루어지고 있다.

참고 X-Ray 장비로 인한 편향을 완화하는 분석 사례[13,20,22]

- 일부 연구자는 일반적으로 흑인이 골밀도가 더 높아서 색상 차이가 인공지능 모델이 인종을 감지하는 요소가 된다고 가정했고, 이를 차단하고자 필터를 적용하여 모델이 색상 차이를 확인할 수 없도록 처리하였다.
- 인종을 예측할 수 있는 고주파수 이미지 특징(즉, 텍스처) 및 저주파 이미지 특징(즉, 구조적)의 주파수 영역 차이를 평가하였고, 이미지 품질의 차이가 의료 이미지에서 인종 인식에 어떻게 영향을 미칠 수 있는지 분석하였다. 마지막으로 특정 이미지 영역(예: 오른쪽 상단 모서리의 방사선 마커 등 이미지의 특정 패치 또는 지역적 변형)이 인종 정체성 인식에 이바지했는지를 분석하였다.

06-2 학습에 사용되는 특성feature을 분석하고 선정 기준을 마련하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단 인공지능 알고리즘 또는 모델을 직접 개발하거나 민감한 특정 변수를 추가로 활용하는 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 의료 데이터의 수집 과정은 큰 비용과 복잡한 절차를 수반하므로, 상대적으로 고소득인 국가나 인종에 대한 데이터가 우선 구축되었다. 이처럼 수집된 데이터 자체에 편향이 존재하면 그 데이터를 통해 학습된 인공지능 시스템의 일반화 성능 및 신뢰성이 낮아진다. 따라서 의료 데이터를 수집 및 가공할 시에는 학습에 사용되는 특성을 특히 더욱 면밀하게 인식하고, 이에 따라 발생할 수 있는 편향을 제거하는 방안을 효과적으로 적용하여야 한다.
- 의료 데이터는 수집 당시의 환경이나 수집 방식에 따라 편향이 발생할 수 있다. 데이터를 수집할 때 모델의 목적에 부합하도록 알고리즘의 결론을 지지할 수 있는 데이터를 선택적으로 취하면 선택 편향selection bias이 발생할 수 있고, 면밀한 검토 없이 알고리즘 또는 인공지능을 이용하여 자동으로 데이터를 수집하면 자동화 편향이 발생할 수 있다. 편향 완화를 위해 데이터에 포함된 차별적인 요소를 사전에 가려내는 것이 중요하며, 이를 위해 학습을 위한 특성을 분석하고 선정 기준을 수립하는 것이 바람직하다.
- 차별적인 요소란 학습 결과가 사회적 물의 및 차별을 일으킬 수 있는 특성으로, 국제기구나 글로벌 기업이 언급하는 민감한 특성의 예시는 아래와 같다. 이와 같은 요소는 데이터 학습 시 반영되지 않아야 하는 특성으로 선정하고, 이에 따라 발생할 수 있는 편향을 방지해야 한다.

사회적 물의를 일으킬 수 있는 민감한 특성들

기관명	특성
UNESCO	나이, 성별, 인종, 민족적·사회적 기원, 혈통, 언어, 종교, 정치적 사상, 국적, 출생 시 사회적·경제적 상황, 장애
ALTAI	나이, 성별, 인종, 민족적·사회적 기원, 혈통, 언어, 종교, 정치적 사상, 소수 민족 구성원, 재산, 출생, 성적 지향
ISO/IEC 24027:2021	나이, 성별, 인종, 수입, 가족관계, 교육 수준, 키·체중, 장애 여부
IBM Watson OpenScale	나이, 성별, 인종, 결혼 여부, 주소
Google	인종, 성별, 장애 여부, 종교

06-2a 보호변수 선정 시 충분한 분석을 수행하였는가?

Yes ☐ No ☐ N/A ☐

- 보호변수 선정 시 충분한 분석을 진행하지 않을 경우, 모델의 성능이 저하될 수 있다. 따라서 모델 추론 결과에 영향을 미치는 특성을 식별한 경우 주어진 데이터셋으로부터 데이터 일부분을 변경하면서 모델의 결과가 어떻게 변하는지 관찰하고 분석하여야 한다.
- 기계학습 기반 회귀 및 분류 모델의 경우, 데이터 변화에 따른 추론 결과의 추이를 시각화하여 보여 주는 도구(예: Google What If Tool)를 사용하여 설정한 보호변수가 불공평한 결과에 얼마나 영향을 미치는지, 성능이 어떻게 변하는지 알 수 있다.
- 의료 인공지능은 현재 지속해 발전하는 중이므로, 알고리즘 성능 극대화가 우선시되어 편향 및 보호변수 문제는 아직 주요 고려 대상이 아닐 수 있다. 의료 분야에서는 '작은 데이터'나 '특정 질환' 등의 데이터 소수결 문제가 중요하며, 오히려 병변이 발현된 희귀 데이터와 그 외 양질의 데이터를 최대한 많이 수집하는 것에 주안점을 두기 때문이다. 그런데도 일반적인 인공지능과 마찬가지로 편향의 가능성은 항상 존재하므로, 이론적인 편향 발생 가능성을 검토하고, 보호변수를 설정해 편향 제거 방안을 마련하여야 한다. 다음은 흔히 논의되는 의료 데이터 편향의 예시들이다.

참고 의료 데이터 편향의 예시

- 급성 신부전을 진단하는 모델이 주로 남성 환자들에게서 수집된 데이터를 통해 학습되면 데이터셋 자체의 성비 불균형으로 인해 여성 환자들에 대한 임상적 유효성이 낮을 수 있다[23].
- 흉부 X-Ray 데이터를 통한 학습 시에 남성 데이터가 여성 데이터보다 현저히 많으면 여성 환자에 대한 진단 성능이 떨어질 수 있다[24].
- 유색 인종에 대해 저산소 혈증이라고 잘못 진단할 확률이 높은 등, 맥박 산소 포화도 측정 시 인종이나 피부색에 따라 진단 정확도에 차이가 발생할 수 있다[25].
- 의료비는 인종, 국가, 소득 수준 등에 따라 불가피한 차이를 보이므로, 단순히 과거 의료비 자료만으로 미래 의료 기술에 대한 수요 예측을 시도하면 알고리즘상 인종 편향이 발생한다[26].

06-2b 편향을 발생시킬 수 있는 특성의 영향력을 완화하였는가?

Yes ☐ No ☐ N/A ☐

- 의료 데이터에 포함된 차별적 요소 등의 특성으로 인해 편향이 발생할 수 있다. 대표적으로는 데이터셋의 구성 등의 특성이 편향을 유발할 가능성이 없는지 검토하여야 한다.
- 해외 의료 데이터셋을 사용할 때는 해당 데이터셋의 특성이 지나치게 특정 성별이나 백인 및 선진국 국민 위주로 편향되지는 않은지 검토하여야 한다. 해당 의료 데이터로 학습된 결과가 성별, 인종, 생활수준 등에 따라 다르게 나타날 수 있기 때문이다. 예를 들어 서양인 위주의 학습 데이터로 학습된 결과를 아시아인 위주의 의료기관에서 사용하면 알고리즘의 성능이 크게 저하될 것이다. 따라서, 인종 등 편향을 유발할 수 있는 데이터셋의 구성적 특성이 모델에 미치는 영향력을 완화할 수 있도록 데이터를 전처리하거나, 개발하는 모델과 시스템의 주 수요층을 차지할 것으로 예상되는 인종을 대상으로 학습 데이터의 특성을 선정하거나, 가능하면 배경이 다양한 환자군으로 이루어진 데이터셋을 사용하여야 한다.
- 편향 완화를 위한 간단한 접근법으로는 편향을 발생시키는 특성을 배제하는 특성 선택feature selection 기법을 고려해볼 수 있다. 필터filter 방법, 래퍼wrapper 방법, 임베디드embedded 방법 등이 있다. 이러한 방법들은 데이터 내 특성들의 통계적 상관관계를 분석하여 높은 상관계수를 갖는 특성을 사용하거나, 특성 일부에 대해 좋은 성능을 갖는 부분 집합subset을 활용하는 것이다.
- 편향과 관련된 특성을 제거하는 경우, 다른 편향을 발생시키거나 강화할 수 있어 모든 경우에 효과적인 방법은 아닐 수 있다. 예를 들어, 의료 데이터상의 인종 편향을 완화하고자 인종 변수를 제거하면 오히려 모델의 성능이나 정확도가 떨어져 가중치 재조정 등을 통한 방법을 마련하는 것이 더 바람직하다는 연구가 있다[27]. 따라서 편향을 완화하기 위한 다양한 기법(예: 가중치 재지정, 라벨링 재지정, 변수 블라인딩, 샘플링)을 고려해야 한다.
- 단, 시스템 사용 목적에 따라 의도된 편향이거나 학습 과정에서 편향 완화가 가능한 경우에는 예외로 할 수 있다.

06-2c 데이터 전처리 시 특성이 과도하게 제거되었는지 검토하였는가?

Yes No N/A
☐ ☐ ☐

- 특성 선택 기법을 통해서 잠재된 편향을 완화하고 모델 성능을 향상시킬 수 있으나, 지나칠 경우 과적합overfitting 문제 혹은 오히려 편향의 원인을 제공하기도 한다.
- 특히, 모든 데이터에서 특성 선택을 시행할 경우, 교차 검증에서 동일한 특성을 사용하게 되므로 편향을 야기할 수도 있다. 따라서 과도한 특성 선택 및 배제를 방지하기 위한 점검이 필요하다.
- 또한, 의료 분야는 인공지능의 임무 목적에 따라 특정 질환을 진단하는 데이터셋의 클래스로서 의도적으로 편향된 인종, 성별, 연령 등의 특성을 사용할 수 있다. 이러한 상황에서 집중 또는 편향된 데이터의 제거가 필요하면, 의학적 지식이 있는 의료진이 검토하여 임무 목적과 잠재된 편향을 구분해야 한다. 특히 과도한 전처리 시에는 모델의 성능이 저하되거나 모델의 목적에 부합하지 않은 결과가 도출될 수 있으므로, 특성을 과도하게 제거하지 않고 특성마다 가중치를 달리 부여하는 등의 주의가 필요하다.

과도한 특성 선택 및 배제를 방지하기 위한 점검표

점검 항목	조치사항
도메인 지식을 가지고 있는가?	만약 가지고 있다면, 도메인 지식을 바탕으로 임시 특성들을 구성하는 것이 좋다.
특성들이 서로 연관 있는가?	만약 그렇지 않다면, 스케일을 맞추기 위해 정규화하는 것이 좋다.
특성들 사이에 상호 의존성이 있는가?	만약 그렇다면, 관련 있는 특성을 결합하여 특성 셋을 확장하는 것이 좋다.
입력 변수들을 비용·속도 등의 이유로 제거해야 할 필요가 있는가?	만약 그렇지 않다면, 특성들을 분리하거나, 특성의 가중치 합을 구성하는 것이 좋다.
모델에 대한 특성의 이해 혹은 필터링을 위해 특성들을 개별적으로 평가해야 하는가?	만약 그렇다면, variable ranking 방법을 사용하는 것이 좋다.
Predictor가 필요한가?	만약 그렇지 않다면, 특성 선택을 할 필요가 없다.
데이터가 지저분한가?	만약 그렇다면, top ranking variable을 이용해 이상값을 제거하는 것이 좋다.
무엇을 먼저 해야 할지 아는가?	만약 모른다면, linear predictor를 사용하고, 전진 선택forward selection 기법이나 0-norm 임베디드 기법을 사용해보는 것이 좋다.
새로운 아이디어와 시간, 컴퓨팅 자원, 데이터가 충분한가?	만약 그렇다면, 다양한 방법을 시도하는 것이 좋다.
안정적인 솔루션을 원하는가?	만약 그렇다면, 여러 번 해보고 bootstrap을 쓰는 것이 좋다.

06-3 데이터 라벨링 시, 발생 가능한 편향을 확인하고 방지하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

의료 분야의 인공지능 알고리즘 또는 모델을 개발하고자 데이터 셋을 직접 수집 및 라벨링하는 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 의료 인공지능 모델을 개발하고자 수집된 데이터를 라벨링 시, 의료 분야에 대한 전문 지식과 라벨링 작업을 위한 사전 지식이 요구된다. 이와 같은 이유로 의료 분야에서 라벨링 작업자는 전문 도메인 지식을 갖춘 전문 의료진, 전문 도메인 지식은 부족하지만 빠르게 라벨링 작업을 처리할 수 있는 크라우드워커로 구분할 수 있다.
- 라벨링 작업자의 특정 의도, 실수로 인한 특성 누락, 무의식적인 판단으로 인한 편향이 발생할 수 있다. 다음은 라벨링 작업자 구분(전문 의료진, 크라우드워커)에 따른 편향 발생 원인에 대한 예시이다.
  - ✓ 전문 의료진이 라벨링을 수행하는 경우, 라벨링 작업 과정에 대한 이해도 부족, 라벨링 작업 및 도구 사용 미숙 등의 이유로 편향이 발생할 수 있다.
  - ✓ 크라우드워커가 라벨링을 수행하는 경우, 의료 분야 전문성 부족, 작업 및 판단 기준의 일관성 결여 등의 이유로 편향이 발생할 수 있다.
- 따라서, 라벨링 작업자가 발생시킬 수 있는 편향의 잠재적인 원인을 사전에 파악하고, 라벨링 결과의 평가 및 작업 기준의 교육 등을 통해 편향 발생을 방지해야 한다. 또한 다양한 라벨링 작업자를 섭외하여 작업자별로 나타날 수 있는 편향을 최소화하거나, 검수자를 충분히 확보하여 편향 방지 작업을 수행하는 것이 바람직하다.

06-3a 데이터 라벨링 기준을 명확히 수립하고 작업자에게 제공하였는가?

Yes ☐ No ☐ N/A ☐

- 의료 인공지능의 잠재적 편향은 사람이 데이터에 주입한 편견 및 사람에 의해 진행되는 라벨링의 과정으로 인해 발생할 수 있다.
  - ✓ 암 등의 질병 발병률이 남성보다는 여성에게서 높게 나타난다고 생각하는 편견, 즉 성별에 대한 고정 관념으로 인해 여성에게 과도하게 진단할 수 있다.
  - ✓ 여러 연구[28, 29]를 통해 통증을 진단할 때, 백인보다 흑인의 통증 원인을 과소 진단하는 인종적 편견이 존재함이 확인되었으며, 이러한 진단 데이터는 라벨링 작업자에게 그대로 전달되어 편향을 발생시키는 원인이 된다.
- 이러한 잠재적 편향은 라벨링 작업을 위한 가이드라인이 명확하지 않아 개인의 판단에 의존함으로써 발생한다. 이를 방지하려면 다양한 의료진과 긴밀하게 협업해 라벨링 기준 수립에 세심한 주의를 기울여야 할 뿐만 아니라, 라벨링 가이드라인을 상세히 마련하여 표준화된 작업 표준을 구축하여야 한다. 다음은 라벨링 작업자에게 전달하기 위한 가이드, 교육 방안 수립 등을 위한 절차 예시이다.

✓ 의사결정 프로세스 매핑: 임상 의료진과 긴밀하게 협업해 데이터 라벨링 기준 및 지침 수립

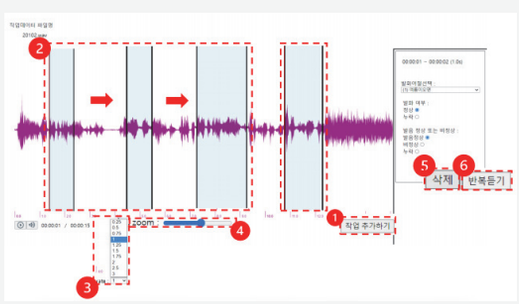
✓ 적합한 라벨링 도구 활용: CT 및 MRI 촬영 이미지, 음성, 심전도 데이터 등 다양하고 많은 데이터를 효과적으로 관리하고, 모든 작업자에게 동일한 작업 환경을 제공하여 라벨링 시 주관성 배제

✓ 작업 가이드 수립 및 교육 제공: 크라우드워커에게는 의료 분야의 특징, 질병, 라벨링 대상에 대한 교육을 제공하고, 전문의, 간호사 등 전문 의료진에게는 도구에 대한 명확한 설명, 절차 등 작업 가이드 및 교육 제공

참고 라벨링 작업자를 위한 작업 가이드 예시(출처: AI Hub)

'소아 흉부 이미지 데이터' 라벨링 도구 가이드 예시

1. 자동으로 생성된 폐부위 세그멘테이션을 확인하여 수정이 필요하면 부분수정 또는 초기화를 하고 다시 그립니다.
2. 폐부위 전체 삭제 시 초기화 버튼으로 클릭하고 수동 생성(Freedraw 방식) 합니다.
3. 폐부위 부분수정 삭제 시 "Eraser"을 선택하고, "Eraser size"의 사이즈(Eraser 붓펜의 두께)를 조정하고 지웁니다.
   - 덧그리기, Undo, Ctrl 버튼과 마우스 휠 위(확대)/아래(축소)/드래그(이동) 가능합니다.
4. "레이블링 저장" 버튼을 클릭하여야 최종 저장됩니다.
   - 질환인 경우 병변부위를 세그멘테이션하여야 레이블링 완료됩니다.
5. 쓰기 버튼 클릭 시 환자 등록정보 확인 가능합니다. (단,수정 불가하며, 수정 필요시 등록현황 상세 화면에서 수정 가능)



'음성질환 판별을 위한 음성 데이터' 라벨링 도구 가이드 예시

※ 각 항목에 대한 세부 내용은 '음성질환 판별을 위한 음성 데이터' 저작도구 설명서 참조

1. 원하는 구간에 작업 추가
2. 추가한 작업 구간은 크기를 조절할 수 있고, 다른 구간으로 이동할 수 있음
3. 재생속도 조절(0.25 ~ 3)
4. 파형 크기 확대 및 축소
5. 잘못 작업한 구간 삭제
6. start 타임라인과 end 타임라인 사이의 구간 반복 듣기

06-3b 다양한 라벨링 작업자를 섭외하기 위해 노력하였는가?

Yes ☐ No ☐ N/A ☐

- 데이터 라벨링 단계에서 인적 편향을 줄이려면 다수의 데이터 라벨링 작업자 확보가 우선적으로 요구된다. 또한, 라벨링 작업자들을 인구 통계학적 특성 및 배경지식 등이 다양하고 고르게 분포되도록 구성하는 것이 바람직하며, 주요 분포 고려 요소는 다음과 같다.
  - ✓ 인종, 종교, 성별, 민족, 장애 여부, 언어, 국적, 경제적 상황 등
- 작업자의 다양성을 검증하기 위해서는 크게 2가지를 확인해야 한다. 첫째, 크라우드소싱crowdsourcing등의 방법을 도입하였는지 점검한다. 둘째, 데이터 라벨링 작업자의 인구 통계적 특성, 배경지식 등을 조사하고 분석함으로써 실제로 라벨링 작업자가 다양하고 고르게 분포하는지를 확인한다.
  - ✓ 크라우드소싱: 데이터 라벨링 과정에 라벨링 관련 교육을 받은 일반인이 참여토록 외부 발주하는 것을 의미하며, 이를 통해 기존 라벨링 작업자 집단보다 더욱 다양한 작업자를 확보할 수 있음
- 데이터 라벨링 과정에서 전문 의학지식이 필요할 때는 의료 분야 전문가를 라벨링 작업자로 섭외하여야 한다. 다음은 라벨링 작업자로 고려할 수 있는 다양한 의료 분야 전문가이다.
  - ✓ 의료 전문가: 의사(내과, 정형외과 등), 치과 의사 등
  - ✓ 약학 및 의학 물리학 전문가: 약사, 의학 물리학자 등
  - ✓ 의료 보조원: 간호 보조원, 보육 보조원, 구급차 운전사, 치과 보조원, 간호사, 운동 치료사, 척추 지압사, 정신 운동 치료사, 언어 치료사, 방사선 조작자, 의료 실험실 기술자, 청력 보철사, 안경사, 보조기 전문의 등

참고 의학 전문 지식을 갖춘 라벨링 작업자 확보 방안 예시(출처: AI Hub)

- ‘소아 흉부 이미지 데이터’ 라벨링 작업자 확보 및 활용 방안
  - ■ 어노테이션/라벨링 조직
    - 어노테이션 수행자들은 영상의학과 전문의들로 구성
    - 어노테이션 수행자들을 대상으로 어노테이션 내용 숙지 후, 어노테이션 시행
- ‘음성질환 판별을 위한 음성 데이터’ 라벨링 작업자 확보 및 활용 방안
  - ‘데이터 라벨러 육성 과정’을 수료한 작업자만 프로젝트 참여 신청이 가능하도록 설정
    ※ 개인 정보 보호를 위해 보안 서약서 및 동의서를 제출한 작업자만 프로젝트 진입이 가능한 플랫폼 활용
  - 어노테이션 작업 전 가이드라인 숙지 필수
  - 크라우드소싱으로 모집된 작업자 중에서 검수자 자격시험을 통과한 작업자만 선발

06-3c 다양한 라벨링 검수자를 확보하기 위해 노력하였는가?

Yes ☐ No ☐ N/A ☐

- 다양한 데이터 라벨링 작업자를 확보했음에도 불구, 인적 편향이 발생할 수 있다. 따라서, 데이터 라벨링 검수자를 확보하고, 라벨링 결과가 데이터 수집 목적 및 데이터 스펙과 다른 부분은 없는지 등을 확인하며, 수정을 요청하는 등의 작업을 실시해야 한다.
- 데이터 라벨링 검수자 역시 데이터 라벨링 작업자와 마찬가지로 다양하고 고르게 분포할 수 있도록 구성하는 것이 바람직하다. 그러므로 크라우드소싱 등의 방법을 도입하였는지 그리고 검수자에 대한 조사와 분석을 통해 그 분포가 다양하고 고르게 형성되는지 점검한다.
- 의료 분야 전문가가 아닌 일반인이 라벨링을 수행한 경우, 검수 과정에서 의료진의 참여가 권장되며, 오류 및 편향 발생이 최소화되도록 검수자가 사전에 합의된 라벨링 가이드라인을 준수할 수 있도록 하는 것이 중요하다.
- 의료 분야 전문가가 라벨링을 수행할 때도 의료진의 라벨링 검수가 필요할 수 있다. X-Ray 이미지에서 암세포의 위치를 식별하고 종양의 윤곽을 그리는 작업 등 의료 전문가마다 다른 결과를 도출할 수 있기 때문이다. 이처럼 의료 분야 전문가의 라벨링 과정에서도 인적 편향이 발생할 수 있으므로, 적어도 두세 명의 전문 의료진 검수자를 확보하여 검수하는 것을 고려해야 한다.

참고 의학 전문지식을 갖춘 라벨링 검수자 확보 및 활용 방안 예시(출처: AI Hub)

- '소아 흉부 이미지 데이터' 라벨링 검수자 확보 및 활용 방안
  - ■ 3차에 걸친 품질 검수 수행
    - (1차) 소아과 전문의 검수
    - (2차) 각 병원 소아과 책임교수
    - (3차) 영상의학과 전문의 및 소아과 전문의로 구성된 자문단

    ※ 각 검수 단계에서 검수 기준에 부합하는지 점검하여 통과 또는 오류 결정하며, 오류 발생 시 어노테이션 재수행
- '음성 질환 판별을 위한 음성 데이터' 라벨링 검수자 확보 및 활용 방안
  - ■ 의료진 교차 검증을 통한 검수
    - 개인 정보 보호를 위해 의료 기관 내에서만 시행되며, 2인 의료진이 교차 검증을 수행
    - 검수를 통해 문제가 있는 음절을 확인하고, 해당 내용을 데이터 가공 기관에 전달하여 수정 요청
    - 수정된 자료는 2인 의료진이 교차 검증하여, 최종 pass 될 때까지 위 과정을 반복
  - ■ 검수 기준
    - 청각적 검수(정성적) 과정으로, 2인의 의료진(2인 중 최소 1인은 후두 음성 질환 진료를 보는 이비인후과 전문의)이 모두 정상적 발화라고 인정할 때만 정상적인 어노테이션으로 인정

06-4 데이터 편향 방지를 위한 샘플링을 수행하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

의료 데이터 활용 시 클래스 불균형 문제가 발생할 소지가 있거나, 클래스 불균형 발생을 확인한 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 연구에 따르면 의료 데이터에서 대부분 다수 클래스(음성 또는 건강한 환자)와 소수 클래스(양성 또는 아픈 환자) 간 불균형 문제가 발생한다[30]. 이로 인한 오분류(위음성 및 위양성)는 의료 진단 결과에 직접적인 영향을 주므로 데이터 불균형 여부에 주의를 기울여야 한다.
- 클래스 간 데이터 분포 균형을 맞추고자 샘플링 기술을 사용할 수 있다. 샘플링은 모집단에서 일정한 기준으로 데이터를 추출하여 표본을 만드는 기법이다. 일정한 기준으로 추출된 표본은 모집단의 분포를 대표하는 동시에 모집단의 클래스 간 불균형으로 인한 편향을 방지하여야 한다.
- 대표적인 기법으로 언더 샘플링under sampling과 오버 샘플링over sampling을 예로 들 수 있다. 일반적으로 의료 분야에서는 데이터 수집 자체가 어려운 소수 범주의 데이터를 다수 범주의 데이터 수에 맞게 늘리는 오버 샘플링 기법을 사용하여 편향을 방지할 수 있다.

06-4a 편향 방지를 위한 샘플링 기법을 적용하였는가?

Yes ☐ No ☐ N/A ☐

- 의료 데이터는 일반적으로 정상 범주와 이상 범주의 데이터 관측 수가 현저히 차이 나는 불균형 데이터이다. 질병의 발병률을 기반으로 하는 데이터의 경우, 데이터 내에 존재하는 음성(건강한 환자) 클래스의 샘플 수는 양성(아픈 환자) 클래스의 샘플 수보다 많으며, 클래스 간 데이터 분포의 차이는 크다. 이때 데이터 불균형을 해소하기 위해 오버 샘플링 기법을 적용해야 한다.

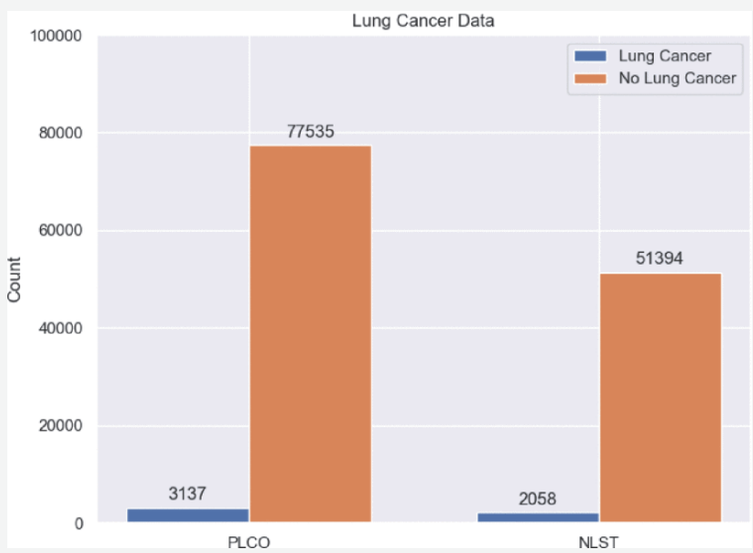
참고 오버 샘플링 기법의 예시

- 오버 샘플링이 필요한 이유: 클래스 불균형이 심할 때, 학습된 모델은 소수 클래스를 제대로 학습하지 못하고 다수 클래스에 과적합 된다. 이러한 모델은 어떤 데이터가 들어오더라도 다수 클래스로 분류하는 문제가 발생한다.
- 랜덤 오버 샘플링ROS, Random Over Sampling
  - ✓ 기존에 존재하는 소스의 클래스를 단순 복제하여 균형을 맞추는 방식
  - ✓ 단순 복제하여 분포는 변화하지 않지만, 숫자가 늘어나 더 많은 가중치를 받는 원리
  - ✓ 똑같은 데이터가 증식되므로 과적합의 위험이 존재
- SMOTESythetic Minority Over-sampling Technique[31]
  - ✓ 소수 클래스의 데이터와 유사한 새로운 합성 데이터를 생성하여 균형을 맞추는 방식
  - ✓ 임의의 소수 클래스에 해당하는 관측치 X를 잡고, X에서 가장 가까운 K개의 이웃 X(nn: nearest neighbors)를 찾은 후, K개의 X(nn)와 X 사이에 임의의 새로운 데이터 X'를 생성

- 보더라인 SMOTE[Borderline-SMOTE][32]
  - ✓ SMOTE에서 조금 변형을 준 알고리즘
  - ✓ 다수 클래스와 소수 클래스가 서로 인접한 경계선, 즉 보더라인의 분포가 매우 중요
  - ✓ 경계선에 있는 소수 클래스의 데이터에 대해서 SMOTE를 적용
- ADASYN[Adaptive Synthetic Sampling][33]
  - ✓ 보더라인 SMOTE에서 조금 더 변형을 준 알고리즘
  - ✓ 보더라인 근처에서 [Danger, Safe, Noise]의 3가지 경우로 판단해 SMOTE로 진행했던 부분을 가중치를 통해 SMOTE를 적용

참고 폐암 데이터에 대한 클래스 불균형 방법 적용 연구[34]

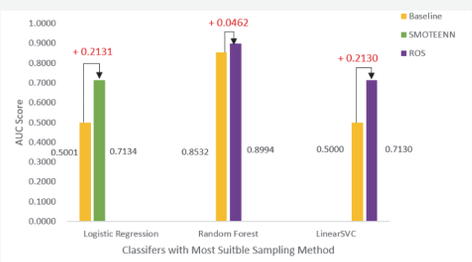
- 대상 데이터
  - ✓ PLCO[Prostate, Lung, Colorectal, and Ovarian], NLST[National Lung Screening Trial] 두 가지 폐암 데이터셋 사용

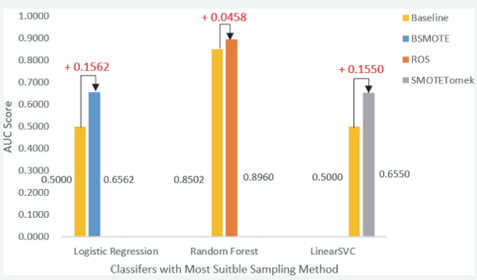


PLCO 및 NLST의 폐암 데이터 클래스 수 비교 – 건강한 사람의 데이터가 훨씬 더 많음

- ✓ PLCO 데이터
  - – 전체 80,672명을 대상으로 조사한 결과 3,137명(약 3.89%)이 폐암임을 확인
  - – 연령, 체질량 지수[BMI], X-ray 병력, 교육, 흡연 상태, 흡연 연수, 연간 흡연 팩 수, 금연 후 연수, 폐암 가족력, 기관지염, 폐기종 병력 및 폐암 확인
- ✓ NLST 데이터
  - – 전체 53,452명을 대상으로 조사한 결과 2,058명(약 3.85%)이 폐암임을 확인
  - – PLCO와 유사하게 연령, 체중, 키, X-ray 이력, 교육, 흡연 상태, 흡연 연수, 연간 흡연 팩 수, 금연 시 연령, 가족(형제, 자녀, 부모)의 폐암 병력, 기관지염, 폐기종 병력 및 폐암 확인

• 샘플링 기법 적용 후 3가지 클래스 분류기의 성능 변화 분석





✓ PLCO 데이터(왼쪽) 및 NLST 데이터(오른쪽) 대상 최고 성능의 샘플링 방법 비교

✓ PLCO 데이터는 ROS와 하이브리드(언더+오버) 샘플링 방법인 SMOTEENN이 가장 높은 성능을 보임

✓ NLST 데이터는 ROS, B.SMOTE 및 하이브리드 샘플링 방법인 SMTOE-Tomek이 가장 높은 성능을 보임

• 클래스 불균형을 보완하여 학습 시, 모델의 분류 능력을 향상시킬 수 있음

✓ 각 샘플링 기법을 비교했을 때 언더 샘플링 기법은 표준편차가 가장 높고, 오버 샘플링 기법이 표준편차가 가장 낮음

✓ ROS를 사용한 랜덤 포레스트 분류 방법은 최고의 예측 성능을 달성하였고, 연구 대상의 데이터에 가장 적합한 것으로 분석함

03 인공지능 모델 개발

책임성 안전성

요구사항 07 오픈소스 라이브러리의 보안성 및 호환성 확보

대표행위자 | 인공지능 모델 개발자 협력 대상 | 시스템 엔지니어

- 의료 인공지능 모델 설계 및 개발 단계에서 개발 기간을 단축하고 최신 기술 동향을 빠르고 유연하게 적용하기 위해 다양한 오픈소스 라이브러리 활용 여부를 고려한다.
- 오픈소스 활용을 결정하였다면 사용할 라이브러리가 안정적으로 업데이트 중인지, 주의해야 할 라이선스 기준은 무엇인지 등을 확인한다. 오픈소스를 사용 중인 경우, 사용하던 오픈소스가 어느 날 라이선스 정책이 바뀌거나 취약점이 새롭게 발견될 수도 있다. 따라서 사용 중인 오픈소스의 목록 및 버전을 지속해서 확인하여 운영 및 보안상의 위험 요소를 점검한다.

07-1 오픈소스 라이브러리의 안정성을 확인하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

의료 인공지능 모델 또는 시스템을 개발하는 과정에서 오픈소스 라이브러리를 사용하는 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 오픈소스 라이브러리는 의료 분야의 연구, 임상 협력 및 적용을 개선하고자 활용할 수 있다.
- 오픈소스 라이브러리는 특정 단체가 관리하기도 하거나, 개인 혹은 기업이 관리한다. 오픈소스를 운영하는 방식은 다양하므로 사전에 꼼꼼히 체크해야 향후 발생할 수 있는 위험risk을 최소화할 수 있다.
- 인공지능 모델 개발에 오픈소스 라이브러리를 사용한다면, 안정성 확인을 위해 해당 오픈소스 라이브러리가 얼마나 많은 사용자를 보유하는지, 업데이트는 자주 이루어지는지, 이슈가 발생했을 때 대응은 신속하게 이루어지는지 등을 따져봐야 한다.

07-1a 활성화된 오픈소스 라이브러리를 사용하였는가?

Yes ☐ No ☐ N/A ☐

- 오픈소스 라이브러리는 공급자가 라이브러리 개선 및 업데이트를 중지하는 경우, 개발 중인 의료 인공지능에 위험을 초래할 수 있다[35]. 의료 영상, 전문 의료 분야 등에서 연구 및 임상 협업을 가속화하고, 이러한 위험 요소를 방지하고자 더욱 활성화된 오픈소스 라이브러리를 사용하는 것을 목표로 해야 한다. 특히 가능하다면, 건강, 의료 전문가, 의료 전문 그룹, 의료 전문 센터 등에서 개발하거나 기여한 라이브러리의 도입을 고려해 볼 수 있다.

- 오픈소스 라이브러리의 안정성은 많은 개발자가 적극적인 참여가 있을 때 가능하다는 의견이 있다. 따라서, 사용하려는 오픈소스 라이브러리의 개발 과정을 주의 깊게 살펴볼 필요가 있다.
- '기업 공개소프트웨어 거버넌스 가이드-정보통신산업진흥원'에 따르면, 오픈소스 프로젝트의 활성화 정도를 확인하는 것도 안정성을 확인하는 한 가지 방법일 수 있다. 해당 오픈소스가 활발한 커뮤니티에서 논의되는지, 그 커뮤니티 내 구성원들이 적극적으로 협력하고 있는지는 아주 중요한 선택의 표지석일 수 있다.
  - ✓ 오픈소스 라이브러리를 GitHub에서 관리 중이라면, 오픈된 이슈 개수나 Pull Request 수, 마지막 커밋 일시 등을 통해 오픈소스 개발이 얼마나 활발하게 이루어지고 지속해서 발전할 가능성이 어느 정도인지 파악할 수 있다.
  - ✓ 그 밖에도 해당 오픈소스와 관련된 StackOverflow 질문 수, 오픈소스 다운로드 수, Google 질의 query 결과 수 등 간단한 측정을 통해서 해당 라이브러리의 활성화 정도를 확인할 수 있다.
  - ✓ 논문 및 연구 결과를 비교 분석하는 경우, Papers with Code에서 최근 라이브러리, 데이터셋, 프레임워크 등의 활성화 정도를 사용자의 등급, 라이브러리에 대한 기여도 등을 확인하여 파악할 수 있다. Papers with Code에서는 건강 응용 프로그램 및 프로젝트 카테고리 등 더욱 상위 개념으로 그룹화하여, 의료 분야 섹션만을 확인할 때는 주의가 필요하다.
  - ✓ Redhat의 경우, 오픈소스 기반의 수익화 모델(호환성, 보안 강화, 기술지원 등 제공)을 개발하고 있으며, 오픈소스 라이브러리 업데이트 시 커뮤니티 내 구성원들이 제안한 개선 사항도 적용한다. 이처럼 수익화 모델 기반의 오픈소스 라이브러리 역시 개인 및 기업의 참여가 활성화된 프로젝트로 판단할 수 있다.

Use Case 신뢰할 수 있는 오픈소스 라이브러리 출처 분석 예시

〈GitHub 출처 분석 예시(2023.01.11. 기준)〉

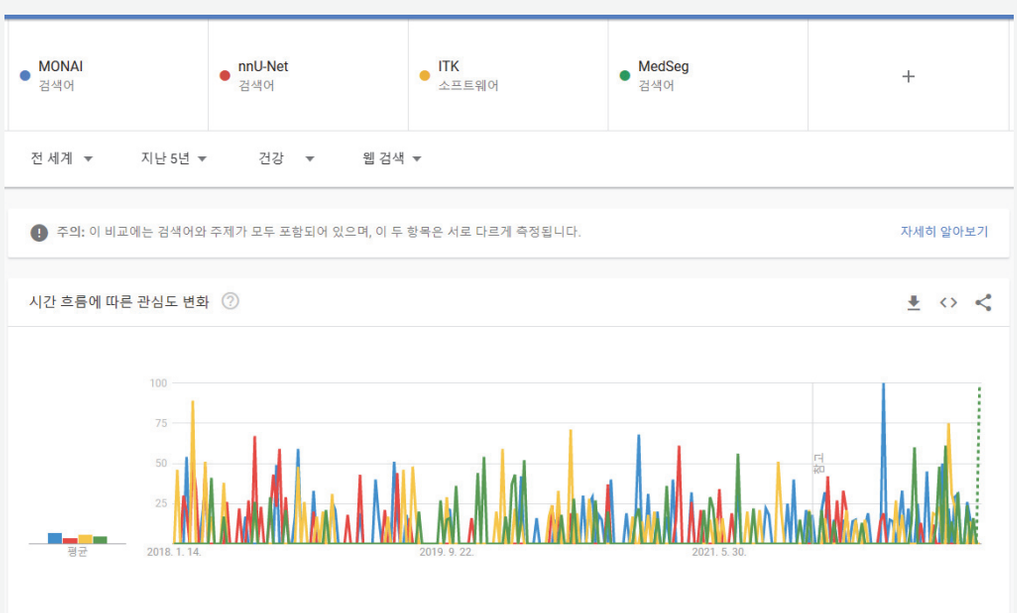
- 대표적인 기계학습 라이브러리 PyTorch, TensorFlow, Keras, SSL4MIS 및 H2O.ai는 대부분 의료 분야에서도 사용되며, 이외에도 의료 분야에 특화된 SSL4MIS, MONAI, nnU-Net, ITK, MedSeg 등이 존재한다.
  - MONAI: 'Medical Open Network AI'의 약어로, PyTorch 기반이며, 의료 영상 심층학습에 최적화됨
  - nnU-Net: 독일 암 연구 센터(DKFZ)의 의료 영상 컴퓨팅 부서에서 출시했으며, 주로 3D 생의학 이미지 작업에 사용됨
  - ITK: 'Insight Toolkit'의 약어로, 의료 영상 처리에 최적화됨
  - MedSeg: 튀빙겐 대학병원의 의료 이미지 및 데이터 분석 연구소에서 출시하였으며, 의료 이미지에 대한 분할 목적으로 3D 컨볼루션 신경망을 학습하고 평가하는 데 사용하는 것을 목표로 함
  - H2O.ai: Hadoop 및 Spark와 통합된 오픈소스 기계학습 라이브러리. 주로 위험 및 사기 성향을 분석하고자 개발되었으나, 의료 애플리케이션의 환자 분석에도 사용됨
  - SSL4MIS: 중국 전자 과학 기술 대학의 기계 및 전기 공학부, 의료 인공지능 연구소에서 출시했으며, 주로 반 지도 이미지 분할 목적으로 사용됨

오픈소스 라이브러리 / 항목	MONAI	nnU-Net	ITK	MedSeg	PyTorch	Tensor Flow	H2O.ai	SSL4MIS
오픈 이슈 개수	221	211	218	2	9,765	2,100	-	15
Pull Request 수	30	14	31	-	785	260	62	-
마지막 커밋 일시	22.01.11	22.09.19	23.01.11	21.07.14	23.01.11	23.01.11	23.01.05	23.01.12
Contributor 수	132	30	271	5	2,590	3,275	166	3
Used 수	618	62	-	-	190,540	232,284	-	-
Star 수	3,700	3,191	1,134	14	61,500	170,000	6,100	1,269
StackOverflow 질문 수	23	12	530	1	20,076	80,534	1,841	-

* 튀빙겐 대학의 라이브러리 MedSeg는 활성화 정도를 판단하는 수치는 낮을 수 있으나 병원 연구소에서 직접 개발하여 고려해 볼 수 있음

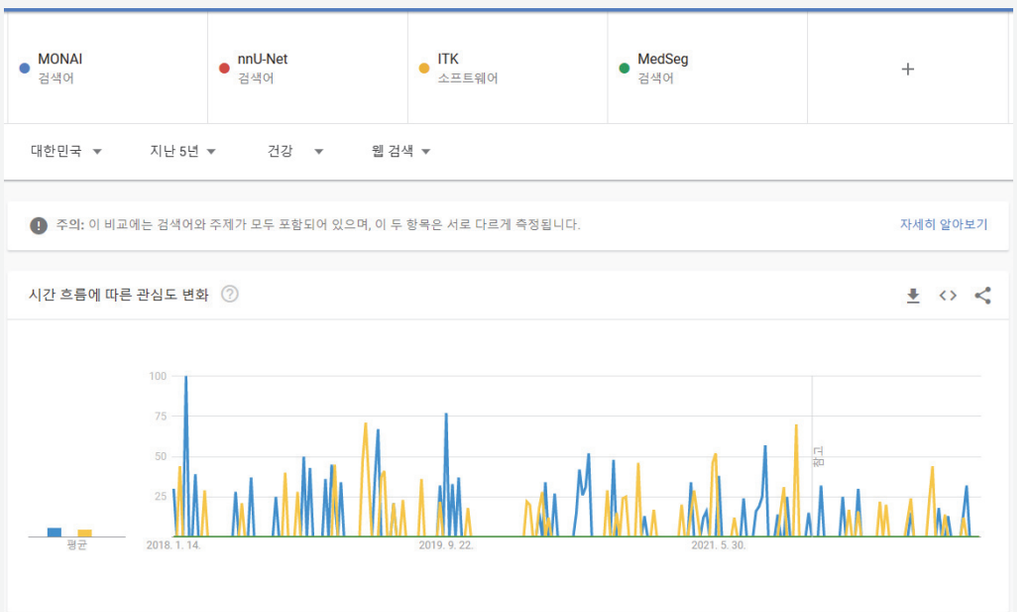
〈Google Trends를 통한 질의 분석 예시(2023.01.11. 기준)〉

- 전 세계 기준, 지난 5년, 건강 카테고리에서 MONAI, nnU-Net, ITK, MedSeg 중에서 MONAI와 ITK의 조회 수가 조금 더 높다(ITK는 동음이의어가 많아서 소프트웨어 주제 검색으로 한정하여 비교).



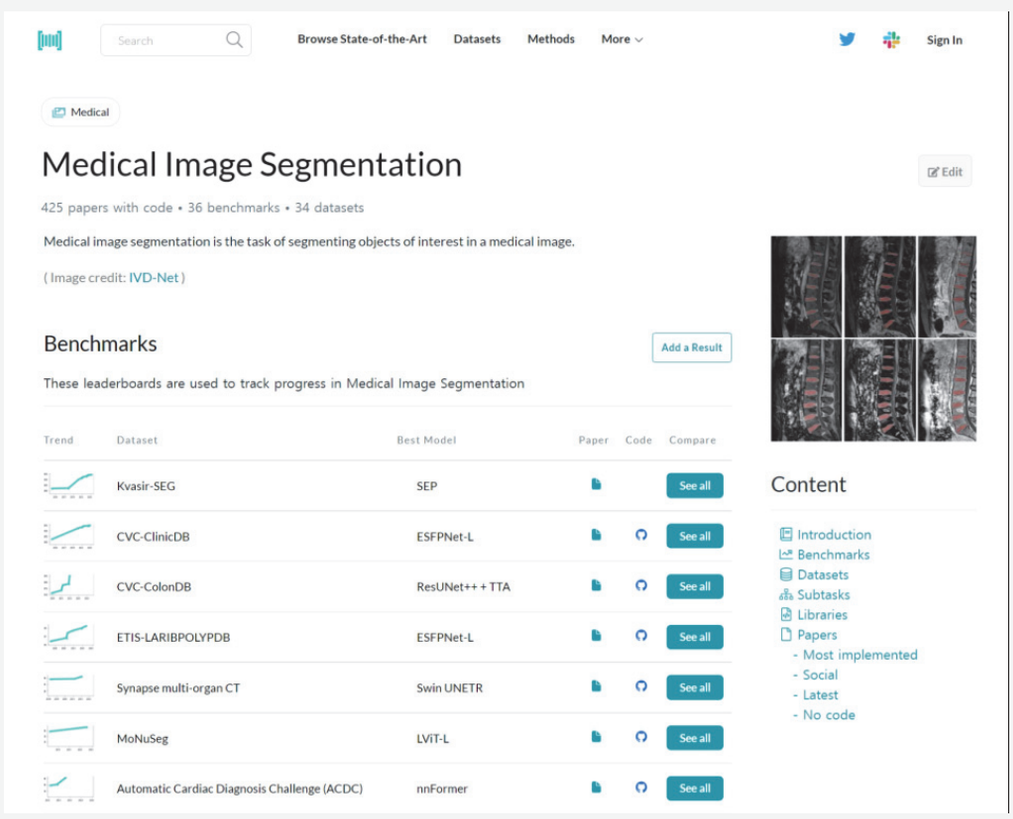
MONAI, nnU-Net, ITK, MedSeg의 전 세계 기준 지난 5년 건강 카테고리 내 검색어 조회 수(관심도) 변화

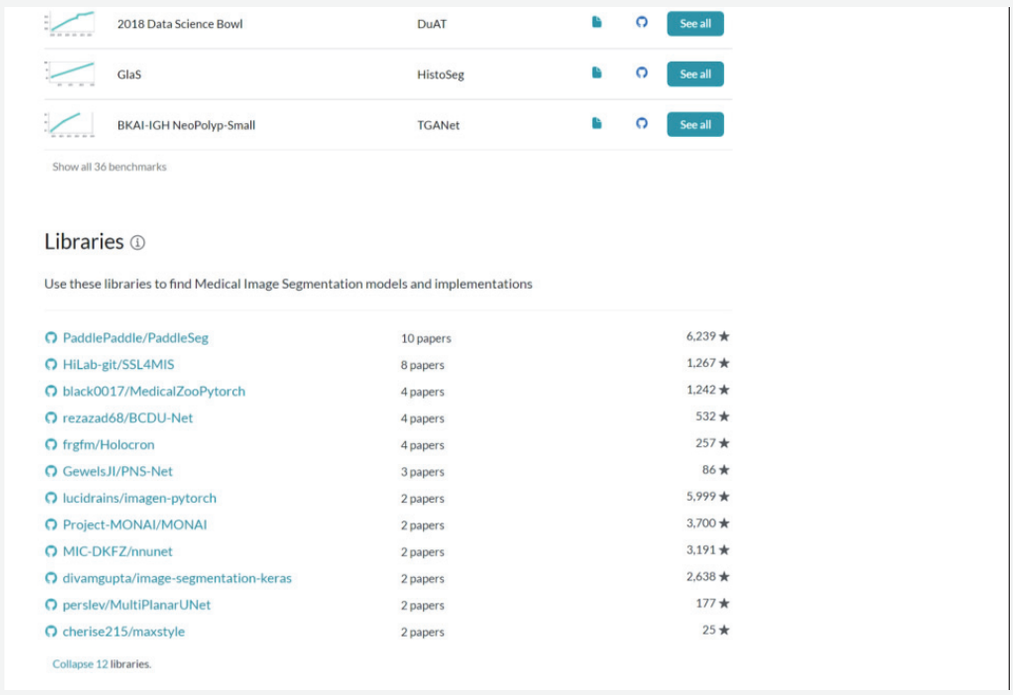
- 대한민국 기준, 지난 5년, 건강 카테고리에서 MONAI, nnU-Net, ITK, MedSeg 중에서 MONAI의 조회 수가 가장 높다(ITK는 동음이의어가 많아서 소프트웨어 주제 검색으로 한정하여 비교).



MONAI, nnU-Net, ITK, MedSeg의 대한민국 기준 지난 5년 건강 카테고리 내 검색어 조회 수(관심도) 변화

〈Papers with Codes 분석 예시(2023.01.11. 기준)〉





2018 Data Science Bowl	DuAT		See all
GlaS	HistoSeg		See all
BKAI-IGH NeoPolyp-Small	TGANet		See all

Show all 36 benchmarks

Libraries

Use these libraries to find Medical Image Segmentation models and implementations

PaddlePaddle/PaddleSeg	10 papers	6,239 ★
HiLab-git/SSL4MIS	8 papers	1,267 ★
black0017/MedicalZooPytorch	4 papers	1,242 ★
rezazad68/BCDU-Net	4 papers	532 ★
frgfm/Holocron	4 papers	257 ★
GewelsJI/PNS-Net	3 papers	86 ★
lucidrains/imagen-pytorch	2 papers	5,999 ★
Project-MONAI/MONAI	2 papers	3,700 ★
MIC-DKFZ/nnunet	2 papers	3,191 ★
divamgupta/image-segmentation-keras	2 papers	2,638 ★
perslev/MultiPlanarUNet	2 papers	177 ★
cherise215/maxstyle	2 papers	25 ★

Collapse 12 libraries.

Papers with Codes 내 Medical Image Segmentation 질의 분석 결과 –
논문과 함께 사용되는 라이브러리 목록, 참조된 논문 수 및 스타 수 확인

07-2 오픈소스 라이브러리의 위험 요소는 관리되고 있는가?

Yes No N/A
☐ ☐ ☐

해당여부 판단

인공지능 모델 또는 시스템을 개발하는 과정에서 오픈소스 라이브러리를 사용하는 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 오픈소스 라이브러리는 저작권자가 소스코드를 공개했을 뿐이며 지식재산권으로 보호받는 소프트웨어이다. 따라서, 저작권자가 제시한 라이선스(저작권) 준수 조건이 존재하며, 오픈소스 라이브러리마다 라이선스에 따라 다양한 의무 사항이 있다. 이때, 라이선스 위반 및 저작권 침해로 법적 책임을 져야 할 위험이 있어서 반드시 라이선스와 관련한 위험 요소를 분석하고 관리해야 한다.
- 오픈소스 라이브러리의 종류 및 버전 선택 시 개발 과정에서 사용된 오픈소스 라이브러리 또는 개발 환경 버전 변경에 따른 호환성을 고려하여야 하며, 이때 사용된 오픈소스 라이브러리에서 보안 취약점이 발견될 수 있으므로 이러한 이슈들을 확인하여 보안상의 위험 요소에 대한 관리도 필요하다.

07-2a 사용중인 오픈소스 라이브러리의 라이선스 준수사항을 이행하였는가?

Yes No N/A
☐ ☐ ☐

- 오픈소스는 무료로 사용할 수 있지만, 라이선스별로 준수사항은 별도로 규정된다. 그러므로 오픈소스 라이브러리를 활용하여 인공지능 모델을 개발한다면, 사용할 오픈소스의 라이선스 종류 및 라이선스 고지문을 확인하고, 허용 또는 의무 사항을 우선해서 숙지해야 향후 발생할 수 있는 법률적 위험을 최소화할 수 있다.
- 다음은 OSI[Open Source Initiative] 단체에서 정한 오픈소스 라이선스의 준수사항이다[36].
  - ✓ 자유로운 재배포(Free redistribution)
  - ✓ 소스코드 공개(Source code open)
  - ✓ 2차 저작물 허용(Derived works)
  - ✓ 저작자의 소스코드 원형 유지(Integrity of the author's source code)
  - ✓ 개인이나 단체에 대한 차별 금지(No discrimination against persons or groups)
  - ✓ 사용 분야에 대한 차별 금지(No discrimination against fields of endeavor)
  - ✓ 라이선스의 배포(Distribution of license)
  - ✓ 특정 제품에만 유용한 라이선스 금지(License must not be specific to a product)
  - ✓ 다른 소프트웨어를 제한하는 라이선스 금지(License must not contaminate other software)
  - ✓ 기술 중립적인 라이선스 제공(License must be technology-neutral)
- 각 오픈소스에서 정하는 라이선스 고지문을 잘 숙지 및 분석하여 저작권, 특허권 등의 지식재산권 위반에 따른 민형사상의 법적 분쟁 등 향후 발생할 수 있는 법률적 위험을 최소화한다. 특히, 소스코드 공개 의무 등은 추후 기업의 영업비밀 노출의 위험으로 이어질 수 있으므로 주의하여야 한다.

대표적 오픈소스 라이선스의 주요 내용

OSI 기준	Apache License 2.0	GPL General Public License 3.0	AGPL Affero GPL 3.0	LGPL Lesser GPL 3.0	MIT License	Artistic License 2.0	Eclipse License	BSD Berkeley Software Distribution License	MPL Mozilla Public License 1.1
복제, 배포, 수정의 권한 허용	○	○	○	○	○	○	○	○	○
배포 시 라이선스 사본 첨부	○	○	○	○	○		○	○	○
저작권 고지사항 또는 Attribution 고지사항 유지	○	○	○	○	○	○	○	○	○
배포 시 소스코드 제공 의무와 범위		전체코드	네트워크 서비스 포함 전체코드	2차 저작물		○ (표준 버전)	모듈 단위		파일 단위
조합저작물 작성 및 타 라이선스 배포허용	○			○	조건부	○	○	조건부	○
수정내용 고지		○	○	○		○	○		○
명시적 특허 라이선스의 허용	○	○	○	○		○	○		○
라이선시가 특허소송 제기 시 라이선스 종료	○	○	○	○		○	○		○
이름, 상표, 상호에 대한 사용 제한	○		○			○		○	
보증의 부인	○	○	○	○	○	○	○	○	○
책임의 제한	○	○	○	○	○	○	○	○	○

07-2b 사용중인 오픈소스 라이브러리의 호환성 및 보안취약점을 확인하였는가?

Yes No N/A
☐ ☐ ☐

- 라이브러리 버전의 변경 과정에서 개발 환경, 언어, 도구 및 다른 라이브러리와 호환되지 않는 문제를 초래할 수 있다. 따라서 오픈소스 라이브러리 종류 및 버전 선택 시 라이브러리 간 의존성dependency을 파악하는 등 호환성을 고려해야 한다.
- 사용 중인 오픈소스 라이브러리에서 보안 취약점이 발견되기도 한다. 보안 취약점에 따른 영향을 최소화하기 위해 보안 취약점 및 버전 변경에 따른 릴리즈 노트release note를 지속적으로 확인하여 신속히 탐지 및 대응해야 한다.
- OpenVASVulnerability Assessment Scanner, OpenSCAPSecurity Content Automation Protocol, OWASPThe Open Web Application Security Project, CVECommon Vulnerabilities and Exposures details 등 취약점 기준 및 분석 도구를 통해 오픈소스 라이브러리의 보안 취약점을 분석함으로써, 최근 발견된 보안 위협 내용이나 그에 대한 제조사의 대응 정도를 파악할 수 있다.

참고 오픈소스 라이브러리의 보안 취약점 분석 예시

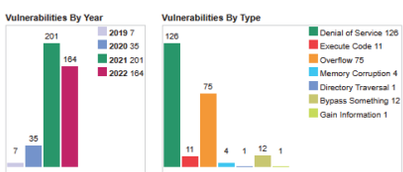
TensorFlow CVE 예시(2023.01.11. 기준)

- DoS[Denial of Service] 공격에 취약한 부분이 존재하는 것으로 분석되고(31.0%), Overflow 위협도 존재하는 것으로 분석(18.4%)
- 보고된 보안 취약점은 2021년 총 201건에서 2022년에 총 164건으로 줄어들어 제조사 측에서의 보안 위협 대응이 어느 정도 이루어지는 것으로 분석

Vulnerability Trends Over Time

Year	# of Vulnerabilities	DoS	Code Execution	Overflow	Memory Corruption	Sql Injection	XSS	Directory Traversal	Http Response Splitting	Bypass something	Gain Information	Gain Privileges	CSRF	File Inclusion	# of exploits
2019	7	1	1	4											
2020	35	6	2	8	3										
2021	201	41	6	38	1			1		8	1				
2022	164	78	2	25						4					
Total	407	126	11	75	4			1		12	1				
% Of All		31.0	2.7	18.4	1.0	0.0	0.0	0.2	0.0	2.9	0.2	0.0	0.0	0.0	

Warning : Vulnerabilities with publish dates before 1999 are not included in this table and chart. (Because there are not many of them and they make the page look bad; and they may not be actually published in those years.)



2019~2022년까지 Tensorflow 오픈소스 라이브러리의 CVE 보안 취약점 분석 결과

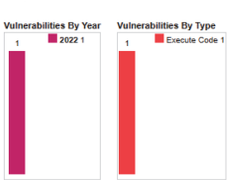
PyTorch CVE 예시(2023.01.11. 기준)

- 보안 취약점 분석 결과, 2022년 1건의 보안 위협이 발견됨
- Code Execution 보안 위협: 이전 버전에서 eval 함수가 안전하지 않게 사용되기 때문에 torch.jit.annotations.parse_type_line에서 임의의 코드 실행을 유발할 수 있음

Vulnerability Trends Over Time

Year	# of Vulnerabilities	DoS	Code Execution	Overflow	Memory Corruption	Sql Injection	XSS	Directory Traversal	Http Response Splitting	Bypass something	Gain Information	Gain Privileges	CSRF	File Inclusion	# of exploits
2022	1		1												
Total	1		1												
% Of All		0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Warning : Vulnerabilities with publish dates before 1999 are not included in this table and chart. (Because there are not many of them and they make the page look bad; and they may not be actually published in those years.)



2022년 Pytorch 오픈소스 라이브러리 CVE 보안 취약점 분석 결과

- 2022년 발견 항목 – CWE-94: Failure to Control Generation of Code('Code Injection')
  - ✓ 보안 위협 내용: 사용자의 입력에 코드 구문이 포함되도록 허용하면 공격자가 소프트웨어의 의도된 제어 흐름을 변경하는 방식으로 코드를 작성할 수 있으며, 인공지능 시스템의 오작동 등을 유도할 수 있음

다양성 존중

요구사항 08 인공지능 모델의 편향 제거

대표행위자 | 인공지능 모델 개발자 협력 대상 | 데이터 과학자 시스템 엔지니어

- 인공지능 모델을 개발하는 과정에서 모델의 종류나 시스템의 목표에 따라 편향*이 발생할 수 있으므로, 이를 제거하기 위한 기법을 고려한다.

* 요구사항 06-2 에서 언급한 바와 같이 인종차별, 성차별 등 사회적·윤리적으로 문제일 때에 한함

08-1 모델 편향을 제거하는 기법을 적용하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

의료 분야 인공지능 모델 개발 시 민감한 특성이 입력값 또는 출력값에 활용되거나 영향을 미쳐 편향 발생이 예상되는 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 인공지능 모델은 데이터에 잠재된 편향을 학습하게 되고, 심지어 편향을 더욱 증폭시키기도 한다. 따라서 데이터 정제 단계에서 데이터에 잠재된 편향을 제거하는 방법뿐만 아니라, 모델 개발 과정에서도 모델 편향을 제거 또는 완화하기 위한 기법을 적용하는 것이 바람직하다.
- 특정 성별에서만 발병하는 질병 또는 특정 인종에서 압도적으로 발병하는 질병 등 의료 분야 인공지능 모델을 적용하는 문제에서 편향은 생물학적 요인이 반영된 일반적인 상황일 수 있다. 그러나, 남녀노소 및 인종을 불문하고 발병할 수 있는 질병 또한 존재하며, 이러한 질병 진단을 목적으로 하는 인공지능 모델을 개발할 때는 사회적·윤리적 편향 문제, 인종에 따라 다른 오진율이 생기는 등 중대한 문제를 초래할 수 있어 유의하여야 한다.
- 편향 완화 기법은 이를 적용하는 단계에 따라 3가지 방식으로 나뉜다. 모델 학습 전에 적용해야 할 편향 완화 기법$^{pre-processing}$, 모델 학습 중에 적용할 기법$^{in-processing}$, 모델 학습 이후 적용할 기법$^{post-processing}$ 이다. 구현하려는 인공지능 모델 및 목표 임무에 따라서 이 중 적절한 기법을 선택하여 적용하여야 한다.

08-1a 개발하려는 모델에 맞게 편향제거 기법을 선택하였는가?

Yes ☐ No ☐ N/A ☐

- 편향된 데이터로 학습된 모델은 진단 신뢰성이 크게 떨어질 수밖에 없다. 이러한 모델 편향은 다양한 기법을 통해 완화할 수 있으므로 적절한 기법의 적용을 고려하여야 한다.
- 각 방식의 특성과 구현하려는 인공지능 모델 및 목표 임무에 맞게 적절한 기법을 선택하여 적용해야 한다.

인공지능 모델의 편향을 완화하는 기법 예시

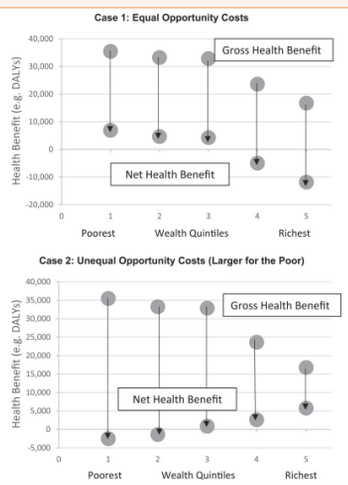
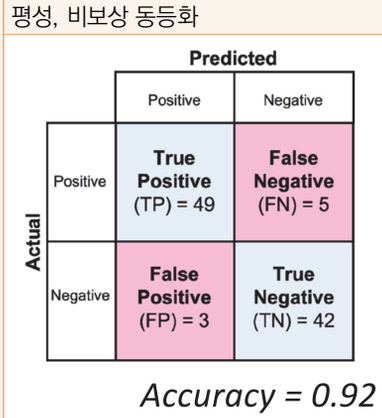
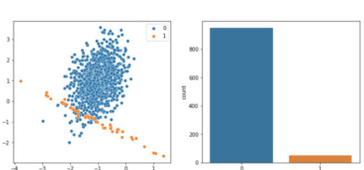
편향 유형	기법 (접근 방법)	기법 구분			설명
		Pre	In	Post	
인지 편향 cognitive bias	다양한 결정 계획 수립		✓		- 정의: 경험으로 발생하는 편향 - 예시: 질병 진단 시 데이터가 상대적으로 많이 축적된 백인 남성만을 기준으로 진단하는 편향 - 완화 방법: 다양한 팀 또는 전문가의 도움으로 완화
알고리즘 편향 algorithmic bias	가중치 재지정	✓			- 정의: 불공정한 학습 반복으로 발생하는 편향 - 예시1: 고소득층인 백인의 의료비 지출이 많으므로 이윤을 극대화하고자 백인을 먼저 치료해야 한다고 판단하는 편향 - 완화 방법1: 공정성을 높이는 가중치를 할당하여 완화 - 예시2: 인종 및 민족 그룹 전반에 걸쳐 유사한 산후우울증 비율을 나타내는 증거에도 불구하고, 백인 여성이 산후우울증 진단을 받을 가능성이 흑인 여성보다 2배 더 높은 편향 내재 - 완화 방법2: (가중치 재지정) 인종별 레이블의 조건부 확률에 따라 각 그룹-레이블 조합에 서로 다른 가중치 적용 (편향 제거) 로지스틱 회귀를 위해 편향 제거[27] 방안을 적용하여 완화
	편향 제거prejudice remover		✓		
자동화 편향 automation bias	자동화 시스템 감독			✓	- 정의: 자동화된 판단 선호로 발생하는 편향 - 예시: 모델이 의료 영상이나 이미지를 판독하는 능력이 향상됨에 따라 사용자 모르게 환자의 인종이나 성별을 예측하고 스스로 학습하는 편향[20] - 완화 방법: 모델을 지속적으로 감독해 완화

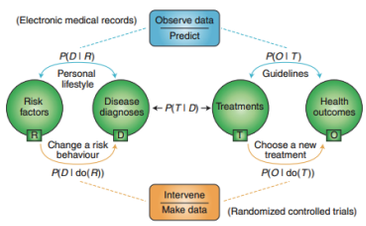
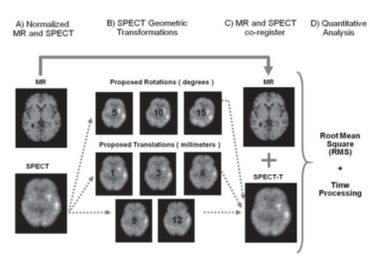
08-1b 편향성 평가 및 모니터링을 위한 정량적 지표를 선정하고 관리하는가?

Yes ☐ No ☐ N/A ☐

- 편향성을 정량적으로 측정하는 지표는 아래와 같이 5가지 분류로 나눌 수 있으며, 개발하려는 모델과 임무 목표에 맞게 지표를 선정하고, 편향 완화 여부를 지속해서 측정 및 관리하는 것이 바람직하다.

편향을 정량적으로 측정하는 지표 분류

<table>
<tr><th>분류</th><th colspan="2">지표</th></tr>
<tr><td rowspan="2">패리티parity 기반 지표[37]</td><td colspan="2">인구통계학적statistical/demographic 형평성 지표, 차등적disparate 효과 지표</td></tr>
<tr><td> 
</td><td>‘불공정’의 기준이 되는 수학적 정의가 필요하지만, 수식에 따라 공정에 관한 판단이 변한다. 대표적인 예로 ‘인구통계적 형평성’, ‘기회의 동등성’, ‘가능성의 동등성’ 등이 있다.
예를 들면, 왼쪽 그래프는 대체 자금 출처의 잠재적 건강 형평성 영향에 대한 그래프다. Case 1은 의료 기회비용이 균등하게 배분된다고 가정할 때, Case 2는 사회적 약자에게 불균형적으로 혜택을 주는 프로그램에서 자금을 조달하는 사례를 설명한다. Case 2는 대체 용도에서 자금이 건강상의 영향을 설명할 때 처음에 건강 형평성에 찬성 영향을 미칠 수 있는 것으로 보이는 프로그램이 실제로 중립적이거나 심지어 반부진일 수 있음을 보여준다.</td></tr>
<tr><td rowspan="2">혼동 행렬 confusion matrix 기반 지표[38]</td><td colspan="2">동등 기회equalized opportunity, equalized odds, 전체 정확도 형평성, 조건부 사용 정확도 형평성, 대응 형평성, 비보상 동등화</td></tr>
<tr><td> 
</td><td>모델, 검사 도구, 알고리즘의 진단·분류·판별·예측 능력을 평가하는 것으로, 오류 행렬error matrix라고도 한다. 혼동행렬을 통해 정확도accuracy, 정밀도precision, 재현율recall, 특이도specificity, F1-score를 도출할 수 있다.
- (정확도) 예측이 현실에 부합할 확률
- (정밀도) 예측 결과가 긍정적일 때 현실도 실제로 긍정일 확률
- (재현율) 현실이 실제로 긍정일 때 예측 결과도 긍정일 확률
- (특이도) 현실이 실제로 부정일 때, 예측 결과도 부정일 확률
- (F1-score) 정밀도와 재현율을 활용하는 평가용 지수로 분류 클래스의 데이터 불균형을 평가할 수 있다.</td></tr>
<tr><td rowspan="2">점수 기반 지표[39]</td><td colspan="2">양성 및 음성 클래스 균형 지표</td></tr>
<tr><td>  </td><td>왼쪽 그래프(다수 클래스 데이터 950개, 소수 클래스 데이터 50개)는 다수의 클래스로 인해, 다수의 클래스를 예측할 때 높은 정확도를 얻을 수 있다. 하지만, 소수 클래스를 예측할 때 낮은 정확도를 보인다. 따라서, 점수 기반 지표인 ROC-AUC, F1-score를 통해 불균형 평가를 통한 데이터의 편향 여부를 확인하여 편향 완화를 고려할 수 있다.</td></tr>
</table>

분류	지표	
사후가정 counterfactual 기반 지표[40]	사후가정 공평성	
		왼쪽 그림은 환자의 습관 변경, 치료 방법 변경 등으로 인해 결과를 달리 설명할 수 있음을 나타낸다. 이처럼 반사실적conterfactual 설명으로 개별 인스턴스 예측 결과를 설명할 수 있다. 반사실적이란 관찰된 사실과 모순되는 가상의 사실이다. 따라서, 예측에 대한 반사실적 설명을 통해 원인이 되는 입력값의 변화를 설명할 수 있으므로 입력 데이터의 편향을 완화하는 데 참고할 수 있는 지표다.
개인 공평성 지표[41]	일반화 엔트로피 지수, 셰일 지수	
		개인 공평성 지표 중 하나인 엔트로피 지수(수치: 0~1)는 불순도를 수치화한 지표로, 확률 변수의 불확실성을 수치로 나타낸 값이다. 0에 가까울수록 불순도가 낮고, 1에 가까울수록 불순도가 높아 데이터의 편향 여부를 확인할 수 있다. 왼쪽 그림은 TSallis 뇌 MRI 이미지 데이터를 분석하면서 엔트로피를 상호 정보를 결정하는 비용함수cost-function로 사용하였다.

안전성

요구사항 09 인공지능 모델 공격에 대한 방어 대책 수립

대표행위자 | 인공지능 모델 개발자 협력 대상 | 시스템 엔지니어

- 인공지능 모델은 적대적 의도를 가진 사용자에 의해 학습 데이터 및 기능을 도용당하거나 다른 방식의 공격으로 악용될 수 있으므로 이를 방지 또는 완화하기 위한 대책을 수립한다.

09-1 모델 추출 공격model extraction attack에 대한 방어 방안을 수립하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

인공지능 시스템이 외부 접근을 통한 공격의 위험이 있는 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 일반적인 의료기기나 소프트웨어에 대한 인공지능 모델의 실제 공격 사례는 아직 없지만, 넓은 의미의 헬스케어 전 분야에서는 인공지능 모델 공격이 이론적으로 가능하므로 추후 공격 사례가 발생할 가능성이 있다. 따라서, 최신 연구 동향을 파악하여 다양한 유형의 인공지능 모델 공격에 대한 방어 기법의 도입을 고려하여야 한다.
- 모델 추출 공격은 학습된 모델의 다양한 입력에 대한 예측 결과를 분석하고 분류 기준을 추출하여 서비스 중인 학습 모델과 유사한 성능의 대체 모델을 구성하는 방법과 학습된 모델의 입력 데이터, 모델의 초매개변수hyperparameter 정보, 계층 구조 등을 추출하는 공격 방식이 존재한다. 이러한 인공지능 모델에 대한 공격을 완화하기 위해 질의query 횟수 제한, 예측 결과 난독화 등의 방법들을 적용할 수 있다.

참고 클라우드 기반 기계학습 서비스에 대한 모델 추출 공격 결과

Service	Model Type	Queries	Time(s)
Amazon	Logistic Regression	650	70
BigML	Decision Tree	1,150	631

Stealing Machine Learning Models via Predictions APIs, Usenix, 2016

- 모델 추출 공격 방법으로 70초 동안 650번의 질의로 아마존 클라우드에서 제공하는 기계학습 모델(logistic regression)과 유사한 모델을 만들어 낸 연구 결과
- 선형 분류기의 회귀계수 및 결정트리의 경로에 대한 정보를 획득해 유사한 인공지능 모델을 만들 수 있음

09-1a 모델 추출 공격에 대비하는 방어 기법을 적용하였는가?

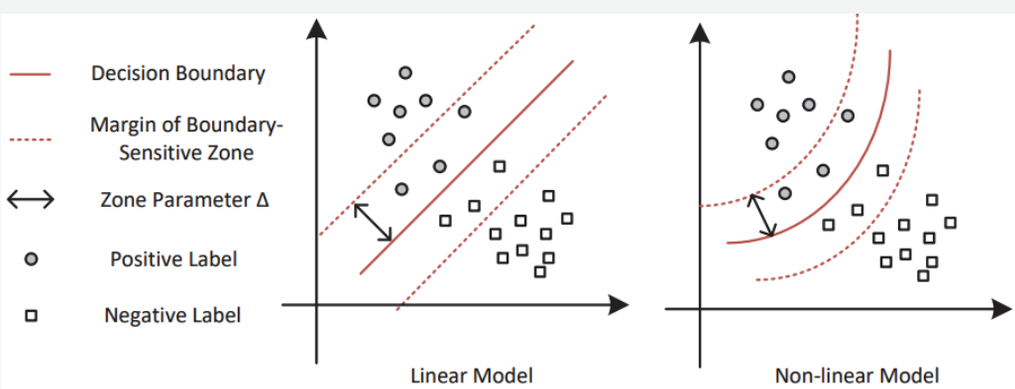
Yes No N/A
☐ ☐ ☐

- 인공지능 모델 공격에 대한 주요 완화 방법에는 특정 시간 간격당 인공지능 서비스에 대한 질의의 수를 제한, 의심스러운 질의에 대한 탐지 및 경고, 예측 결과의 난독화obfuscation 등이 있다.

인공지능 모델 추출 공격에 대한 방어 기법 예시

방어 기법 분류	방어 기법 내용
질의 횟수 제한	• 특정 기간 내에 수행할 수 있는 질의 수를 제한하여 모델 공격을 위한 반복적인 질의를 방어하는 기법
학습 기반 모니터링	• 기계학습을 활용하여 모델 공격에 대해 사전 탐지 및 경고 알림, 상응하는 방어 기법을 실행하는 등 능동적으로 방어하는 기법
예측 결과 난독화	• 예측 결과가 결정경계에 가까운 경우 예측 결과의 정확도를 임의로 낮춰 모델의 세부 속성에 대한 추출을 방해하는 기법

참고 예측 결과 난독화 관련 방어기법 – BDPL Boundary Differentially Private Layer[42]



BDPL 기법은 분류를 결정짓는 기준 및 그 주변 영역을 boundary sensitive zone으로 지정하고 이 영역을 보호하여 외부에서 들어오는 입력이 민감한 영역에 가까울수록 결과에 잡음을 섞어 모델을 추출하기 어렵게 만든다.

09-2 모델 회피 공격model evasion attack에 대한 방어 방안을 수립하였는가?

Yes ☐ No ☐ N/A ☐

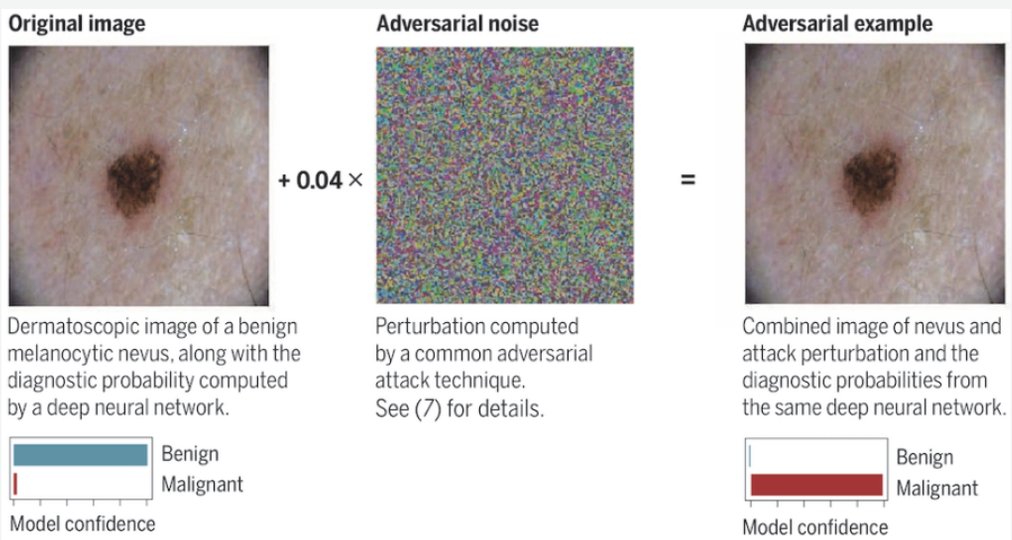
해당여부 판단

인공지능 진단 서비스 개발 시 의료보험 부정수급, 약물 임상시험 대상자 선별, 응급환자 선별 등에서 편법이 예상되는 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 모델 회피 공격은 적대적 공격의 한 방법으로, 입력 데이터에 최소한(인간이 알아차릴 수 없을 정도)의 변조를 가해 인공지능 모델의 성능을 교란시키는 기법이다. 의료 분야 서비스의 시나리오로는 행위별 수가제free-for-service 상황에서 부정수급을 노린 적대적 공격, 임상시험 참가자 선별을 위한 적대적 환자adversarial patient 생성, 응급환자 진찰 순서의 변경을 노린 적대적 공격 등을 예상할 수 있다[43].
- 특히, X선 영상, MRI 영상, 초음파 영상 등 이미지를 활용하는 인공지능은 모델 회피 공격에 매우 취약할 수 있어 이를 완화하는 방어 기법으로 적대적 학습, Gradient Masking/Distillation, Feature Squeezing 등의 기법을 고려할 수 있다.

참고 정상 피부 점 이미지에 대한 모델 회피 공격 예시[44]

피부의 점 이미지에 대하여 인간이 알아차릴 수 없는 수준의 변조를 가하면 인공지능 모델이 피부의 점을 악성종양(암)으로 진단할 수 있다.



09-2a 모델 회피 공격에 대비하는 방어 기법을 적용하였는가?

Yes ☐ No ☐ N/A ☐

- 의료 분야의 현미경 검사, X선 영상 분석, 안저 영상 분석 등 다양한 인공지능 알고리즘을 대상으로 하는 모델 회피 공격 사례가 다수 발생하고 있어 이를 방어하기 위한 적절한 대응 과정이 필요하다.
- 모델 회피 공격에 대한 주요 완화 방법에는 적대적 공격 데이터를 역으로 활용한 적대적 학습, 적대적 공격 여부를 판단하는 모델을 추가하는 방법, 자석 방법magnet method 등이 있다.

모델 회피 공격에 대한 방어 기법

방어 기법 분류	방어 기법 내용
적대적 학습	• 가장 잘 알려진 방법의 하나로, 모델을 학습시킬 때 적대적 사례를 모방한 적대적 샘플 훈련 데이터셋을 학습 데이터셋에 포함하는 방법이다. • 하지만, 적대적 샘플 훈련 데이터셋이 충분한 수와 다양성이 보장되지 않으면, 즉 모든 적대적 사례의 경우의 수를 고려하지 않는다면 적대적 학습이 방어 기법으로서 성능을 보장하지 못한다. • 적대적 학습 방법 종류: FGSMFast Gradient Sign Method 적대적 학습, PGDProjected Gradient Descent 적대적 학습, ALPAdversarial Logit Pairing
적대적 공격 여부를 판단하는 모델 추가	• 적대적 공격 탐지 기법으로, 동일한 두 모델(원래의 모델과 적대적 공격을 판단하는 모델)의 추론 결과를 비교하여 두 결과 간에 차이가 발생했을 때, 적대적 공격으로 판단하는 기법이다. • 또한, 특정 모델에 적용되는 적대적 공격을 불가능하게 만들고자 다수의 학습 모델을 조합하는 방법도 있다.
자석 방법[45]	• 정상 데이터의 다양함을 근사화하여 정상 예와 적대적 예를 구분한다. 적대적 예를 다양체(매니폴드) 근처로 이동하여 재구성하며, 이는 섭동perturbation이 작은 적대적 사례를 올바르게 분류하는 데 효과적이다.
질의 조회 차단	• 반복적인 질의를 시도하는 inversion attack이나 model extraction attack을 방지하고자 모델 횟수를 제한하는 방식이다.

책임성 투명성

요구사항 10 인공지능 모델 명세 및 추론 결과에 대한 설명 제공

대표행위자 | 인공지능 모델 개발자 협력 대상 | 데이터 과학자 시스템 엔지니어 시스템 운영자 전문 의료진

- 인공지능 모델의 추론 결과만으로는 예측된 결과가 어떤 요소에 의해 도출되었는지 알기 어렵다. 또한, 시스템의 최종 결과를 얻기 위해 다수의 인공지능 모델이 사용될 수 있다. 이러한 과정에서 인공지능 모델의 예측 결과에 대한 사용자 신뢰를 확보하기 위해서 사용된 모델 정보, 결과 도출 과정에 대한 설명*, 추론 결과에 대한 설명을 제공한다.

  * 사람이 인공지능 모델의 의사결정 방식을 파악할 수 있도록 돕는 모델의 작동 방식에 대한 유용한 정보(예: 의사결정 메커니즘, 의사결정의 기초를 이루는 학습 데이터, 인공신경망 내에서 사용된 변수와 가중치)

참고 설명가능성explainability 적용 전 고려해야 할 사항

1. **제품 및 서비스의 다양성에 대한 고려:** 설명가능성이 모든 인공지능 모델과 제품 및 서비스에 필요한 것은 아니다. 사용자가 제품 및 서비스를 이용하면서 시스템 동작 및 모델의 추론 결과에 관해 설명을 요구하는 분야가 있지만, 그렇지 않은 분야도 있다. 관련하여, UNESCO에서는 일시적이지 않거나, 쉽게 되돌릴 수 없는 인공지능 시스템의 경우에는 출력된 결과의 투명성이 보장되도록 사용자에게 의미 있는 설명이 제공되어야 한다고 언급한다. 따라서 이러한 사항들을 고려하여 본 요구사항을 선택적으로 적용할 수 있다.
2. **설명가능성이 미치는 영향에 대한 고려:** 설명가능성은 아직도 기술적으로 연구 및 개발이 활발하게 이루어지고 있는 분야로서, 여전히 기술적 한계가 존재함과 동시에 설명가능성 외 다른 속성과도 상호 연관성이 있어 신중히 접근해야 한다. 일례로, 과도하게 설명가능성을 구현하는 경우, 모델 성능 및 프라이버시 등에 부정적인 영향을 초래한다는 의견도 존재한다. 따라서 본 요구사항은 개발과정에서 진단 보조 측면의 인공지능 활용도와 설명이 필요한 정보에 관해 전문 의료진과 충분히 논의하여 적절한 설명 수준을 마련하여야 한다.

10-1 사용자가 모델 추론 결과의 도출 과정을 수용할 수 있도록 근거를 제공하는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

의료 인공지능 모델의 추론 과정이나 추론 결과가 도출된 이유에 대해 사용자 대상의 설명이 요구되는 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

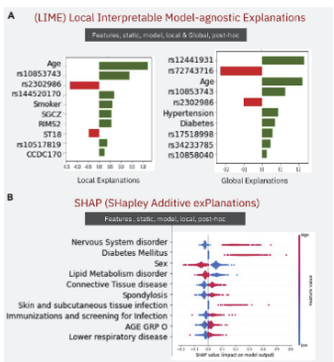
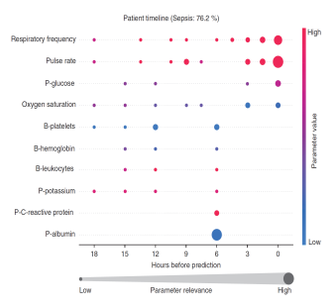
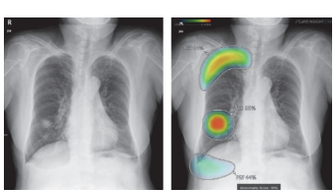
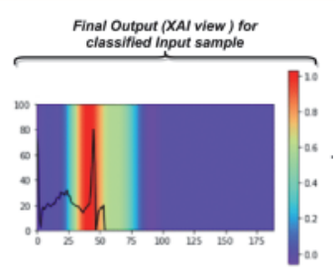
- 의료진과 환자가 인공지능 모델의 추론 결과 및 인공지능 시스템의 동작을 사용자가 신뢰하기 위해서는 시스템 사용자가 인공지능 모델이 제공하는 추론 결과의 도출 과정을 이해할 수 있어야 하며, 이에 대한 설명 및 근거를 사용자에게 제시하는 것이 바람직하다.
- 사람이 이해할 수 있는 방식의 모델 판단 근거를 제시할 수 있는 설명 가능한 인공지능XAI 기술에 대한 검토 및 적용을 고려해야 하며. 설명이 필요한 요소 및 인공지능 모델 특성에 따라 CAMClass Activation Map, MPMeaningful Perturbation 등의 XAI 기법 적용을 검토할 수 있다.
- XAI 기술을 통해 모델 추론 결과를 설명할 수 있는 경우 임상적 의사결정 측면에서 의료진에 도움을 줄 수 있다. 임상에서는 인과관계를 이해하는 것이 필수적이므로, 모델의 설명력은 의학에서 특히 중요하다고 할 수 있다. 그러나 의료진이 이해하기 어려운 정도의 설명은 오히려 의사결정에 부정적 영향을 끼칠 수 있으므로 기법 적용 시 의료진과 충분한 논의가 필요하다.
- 또한, 인공지능 모델 추론 결과의 근거를 설명하는 것이 항상 가능한 것은 아니므로 XAI 기술 적용 이외의 대안을 활용하여 인공지능 시스템의 투명성 확보가 필요할 수 있다. 따라서, XAI 기술 적용 가능 여부를 검토한 후, 검토 결과 XAI 기술 적용이 가능하다면 본 세부 요구사항을 활용하고 적용이 어렵다면 10-1b를 활용할 수 있다.

10-1a XAIeXplainable AI 기술 적용이 가능한 경우, 인공지능 모델의 추론 결과를 설명하기 위한 기법 적용에 대해 검토하였는가?

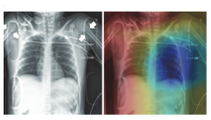
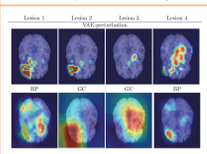
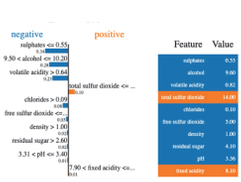
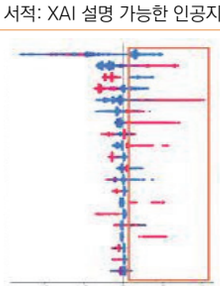
Yes ☐ No ☐ N/A ☐

- XAI 기술 도입을 고려한다면 의료 인공지능에서 주로 사용되는 3가지 유형의 데이터 종류에 따라서 도입이 가능한 기법 탐색이 필요하다.
  - ① 전자의무기록이나 차트에 저장된 환자 진료 기록, 유전체 데이터 등 복잡한 의료 데이터를 분석하는 인공지능: 텍스트(자연어) 데이터 사용
  - ② X-Ray, CT, MRI 등 의료 영상을 판독하는 인공지능: 이미지 데이터 사용
  - ③ 환자의 임상데이터 등 연속적 의료 데이터를 모니터링하여 질병을 예측하는 인공지능: 시계열 데이터 사용
- 의료진이 사용하는 인공지능 시스템에 XAI 기법을 도입하는 예로 CAM 적용을 들 수 있다. 의료 인공지능에서 현재 사용 빈도가 높은 CNNConvolutional Neural Network 계열 모델을 활용할 때 CAM 기법을 통한 히트맵heatmap 표현을 통해 추론 결과에 미친 중요도에 따라 해당 위치를 강조하여 나타낼 수 있다.
- 이처럼 XAI 기술은 의료진의 진단을 보조하는 의료 인공지능 시스템이 어떤 과정과 근거로 추론 결과를 도출하였는지 이해하는 데 도움을 줄 수 있다.

참고 의료 분야 데이터 형식별 활용 사례

데이터 종류	적용기법 예	
텍스트(자연어) 데이터		COVID-19와 연관된 임상 유전체 요인을 인공지능에 입력 데이터로 주어 COVID-19 중증도severity를 설명하였다. 제시된 그림에서는 양성 또는 음성 여부에 따라 초록색으로 표현하여 인공지능이 어떤 유전체 요인(텍스트)에 주목하는지 설명할 수 있다[46].
		P-creatinine, Kindney eGFR 등 임상 매개변수를 활용한 급성 중환자 조기 발견 인공지능 모델에서 조기 경고 점수EWS, Early Warning Scores를 이용하여 추론 결과에 대한 시각적 설명을 제공하였다. 해당 연구에서는 시간에 따라 환자의 패혈증과 연관된 요인을 색상별로 나타내어 연관성이 짙은 순으로 상위 10개를 나열하였다. 이처럼 임상 진단 상황에서 예측 결과를 시각적으로 설명하면 의료진들이 효과적으로 진단 보조에 활용할 수 있다[47].
이미지 데이터		인공지능 모델이 주어진 이미지를 기반으로 환자의 상태를 예측할 때 모델 예측에 활용되는 판단 영역 및 중요도는 다를 것이다. 따라서 왼쪽 그림과 같이 히트맵을 사용하여 모델이 예측에 활용한 영역을 시각화한다면, 사람이 모델 예측 근거를 이해하는 데 도움이 될 수 있다.
시계열 데이터		정제된 심전도 신호 데이터를 입력 데이터로 삼는 인공지능이 어떤 파형에 주목하는지 시각적으로 나타내고 있다. 심장 박동에서 중요한 의미를 갖는 부분의 파형을 빨간색으로 표시한다[48].

참고 의료분야 데이터 형식별 XAI 기술적용 방안

데이터 종류	적용기법 예		
	알고리즘	내용	
이미지 데이터	CAM		Class Activation Map 알고리즘은 CNN이 입력으로 들어온 이미지를 분류할 때 "어떤 부분을 보고" 예측했는지를 알려주는 역할을 한다. 입력 이미지에 히트맵을 씌워 주어진 질병을 예측할 때 중요한 부분에 가까워질수록 빨간색으로 변해감을 확인할 수 있다.
		참조논문: https://doi.org/10.1109/CVPR.2016.319 참조코드: https://github.com/zhoubolei/CAM 출처: https://doi.org/10.1148/radiol.2018181422	
	MP		Meaningful Perturbation은 입력 이미지에 대해 블러 처리나 부분 삭제 등 학습에 방해되는 요소가 존재하더라도 사물을 감지할 수 있는 기술이다. 이를 의료 분야에 적용하면 CT 이미지 내의 텍스트 등의 요소에 방해받지 않으면서 의료진에게 병변과 관련된 위치를 주목하도록 할 수 있다.
		참조논문: https://doi.org/10.1016/j.media.2022.102470 출처: https://doi.org/10.1117/12.2511964	
텍스트(자연어) 데이터	알고리즘	내용	
	LIME		LIME[Local Interpretable Model-agnostic Explanation]은 이미지뿐만 아니라 텍스트(자연어) 데이터에 대한 입력도 지원한다. 위 그림은 LIME의 사용 예시를 보여준다. 환자가 독감에 걸린 것으로 모델이 예측하면, LIME은 모델이 예측에 사용한 변수 중 높은 기여도를 보이는 변수를 추출한다. 이를 통해 의료진은 모델의 예측에 대한 신뢰 여부 정보에 입각한 결정을 내릴 수 있다.
		출처: https://doi.org/10.48550/arXiv.1602.04938 유튜브: https://www.youtube.com/watch?v=d6j6bofhj2M 서적: XAI 설명 가능한 인공지능, 인공지능을 해부하다	
	SHAP		SHAP[SHapley Additive exPlanations]은 특정 변수가 예측에 얼마나 이바지하는지 파악하고자 특정 변수에 대한 모든 변수 조합을 입력하고 결괏값을 비교하여 변수의 기여도를 계산한다. 기여도가 높은 변수를 파악할 수 있다는 점에서 LIME과 유사하나, 데이터에 가중치를 부여한다는 점에서 차이가 있다. 위 그림에서 파란색과 빨간색은 예측 결과에 높은 기여도(파랑)와 낮은 기여도(빨강)를 나타내며, 막대의 크기는 영향을 미친 정도를 나타낸다.
		출처: https://christophm.github.io/interpretable-ml-book/shap.html#fnref44	

10-1b XAI 기술 외에도 수용가능한 모델 추론 결과의 근거를 제공하는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 모델의 추론 결과 및 결정의 근거를 설명하는 것이 항상 가능한 것은 아니다. 또한, 인공지능 기술을 임상에 적용하려면 충분한 임상 유용성과 안전성이 검증되어야 하므로 XAI 기술을 적용하더라도 추론 결과의 설명이 불충분할 수 있다[49].
- 특히, 투명성은 의료진과 환자의 기계학습 가능 의료기기에 대한 신뢰 확보 방안으로 필수적으로 요구되고 있는 현실이며, 임상시험 방법 설계 시 임상 유효성 평가변수, 임상시험 평가 결과에 대한 성공 기준의 설정 사유 및 근거 제시를 요구하고 있다[50].
- 모델 추론 결과 근거를 제공하는 유효성 검증 과정에서 임상의의 판독 정확도와 비교가 수행될 수 있으며, 유효성을 검증하는 참조 표준ground truth 설정 과정에서도 숙련된 임상의의 참여가 동반될 수 있다. 이처럼 임상시험에서 숙련된 임상의가 참여한 인공지능 모델의 유효성 검증 과정을 설명함으로써 추론 결과에 대한 수용 가능성을 높여야 한다[8,51].

10-2 인공지능 모델 상세 문서를 통해 모델의 명세를 투명하게 제공하는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

신뢰성을 확보하기 위해 인공지능 알고리즘 또는 모델이 반영된 시스템의 명세 정보를 투명하게 제공하고자 하는 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 인공지능 시스템의 투명성을 확보하는 방안 중 하나는 인공지능 모델 또는 서비스의 개발, 테스트 및 배포 과정에서 발생한 다양한 결과를 문서로 작성하는 것이다. 모델의 명세를 작성한 상세 문서가 확보될 경우, 사용자가 인공지능 모델과 관련된 정보를 요구했을 때 모델의 목적, 입·출력 정보, 성능, 편향 여부 및 신뢰도 등의 결과들을 투명하게 공개할 수 있다.
- WHO 및 유럽의회조사처EPRS, European Parliamentary Research Service 등은 인공지능 모델 설계 및 구축 전·후의 충분한 정보의 명세 및 문서화를 요구하고 있다[51,52]. 인공지능 모델의 주요 정보 및 구성 요소를 상세하게 문서화하는 것은 의료 인공지능의 투명성뿐만 아니라 잠재적인 오류 발생 시 추적 가능성 확보 측면에서도 중요한 요소이다.

10-2a 시스템 개발 과정과 모델 작동 방식에 대한 세부 정보가 설명된 문서를 작성하였는가?

Yes ☐ No ☐ N/A ☐

- IBM과 WEF에서는 모델의 명세를 작성한 문서를 통해 인공지능 시스템의 투명성을 확보하는 방안을 제시하고 있다. 그러나, 인공지능 모델 개발 과정에서 임상적 의사 판단 정보가 포함되는 의료 분야는 명세 작성 시 인공지능 모델의 메커니즘 측면을 포함하여 추가 정보를 명시하여야 한다.
- 추가 정보의 예는 의도된 임상 용도, 주요 가정, 명세 해석 시 주의사항 등이 있으며, 의료 분야 특성에 맞는 모델 명세 작성 정보에 대한 다양한 방안이 제안되고 있다. 다음은 유럽의회조사처가 제안한 'AI 여권'의 모델 명세 작성 시 필요 정보이다[51,52].

명세 종류	명세 내용
모델 관련 정보	모델 소유자, 기술 성숙도TRL, Technology Readiness Level, 라이선스, 생성 데이터 등
사용 목적	주요 용도, 2차 용도, 사용자, 반대 적응증counter-indications, 윤리적 고려사항 등
모델 세부 정보	모델 디자인, 초매개변수 정보, 목적함수, 공정성 제약조건 등
사용 데이터 정보	데이터 출처, 인구 그룹, 변수, 전처리 방법 등
평가 정보	평가 데이터, 지표, 결과, 한계 등
모니터링 정보	마지막 평가 정보, 확인된 실패 정보, 버전명 등

참고 패혈증 감시 모델의 모델 명세 예시[53]

패혈증 감시 인공지능 모델에 대한 관련 정보 명세를 위해 의약품 라벨 정보를 참조하여 메커니즘, 성능, 주의사항 등의 정보를 전달하고자 다음과 같이 모델 명세를 작성하여 제시하였다.

Model Facts

Model name: Deep Sepsis
Version: 1.0

Summary
This model uses EHR input data collected from a patient's current encounter at the hospital to estimate the probability that the patient will meet sepsis criteria within the next 4 hours. It was developed in 2016-2019 by the Duke Institute for Health Innovation. The model was licensed to Cohere Med in July 2019.

Mechanism
- **Outcome** .....Sepsis within the next 4 hours. See (1) for sepsis criteria
- **Output** .....0% - 100% probability of sepsis occurring in the next 4 hours
- **Patient population** .....all adult patients >18 y.o. presenting to DUH ED and admitted
- **Time of prediction** .....every hour of a patient's encounter
- **Input data source**.....Epic data extracted from Clarity
- **Input data type** .....demographics, analytes, vitals, medication administrations
- **Training data location and time-period** .....DUH, 10/2014 – 12/2015
- **Model type**..... Recurrent Neural Network

Validation and performance
- **Retrospective:** 20% random held out set from 10/2014 – 12/2015, AUROC 0.882
- **Temporal:** 6-month temporal validation set of ED visits at DUH between 03/2018 – 08/2018, AUROC 0.943
- **Prospective:** TBD
- **External:** TBD

Uses and directions

10-3 필요 시, 인공지능 모델 추론 결과에 대한 설명을 제공하는가?

Yes No N/A
☐ ☐ ☐

해당여부 판단

신뢰성을 확보하기 위해 인공지능 알고리즘 또는 모델이 반영된 시스템의 추론 결과에 대한 설명을 함께 제공하고자 하는 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 사용자에게 인공지능 모델의 추론 결과에 대한 설명을 제공하면, 사용자는 단순히 해당 인공지능 모델의 최종 결과뿐 아니라 그 결과가 도출된 수치적인 근거로 확률값, 불확실성 등을 제공받을 수 있다. 이러한 정보는 사용자의 의사결정에 도움이 되지만, 오히려 사용자의 혼란을 유발할 수도 있으므로, 정보 제공의 필요성을 사전에 검토하는 것이 필요하다.

10-3a 모델 추론 결과에 대한 설명이 필요한지 검토하였는가?

Yes No N/A
☐ ☐ ☐

- 인공지능 시스템이 도출한 결과에 대한 설명을 제공하는 것은, 사람들이 인공지능을 활용하여 의사결정하는 데 도움이 될 수 있지만, 오히려 방해될 수도 있다. 따라서 모든 경우에 모델의 추론 결과에 대한 설명을 제공하기보다는, 설명이 꼭 제공되어야 하는지를 확인하는 과정이 선행되어야 한다.
- 모델의 추론 결과에 대한 설명을 제공하지 않는 편이 더 나을 때의 두 가지 예시는 다음과 같다.
  - ✓ 첫째, 모델의 추론 결과에 대한 설명 제공 자체가 사용자의 의사결정에 크게 영향을 미치지 않을 것으로 판단되는 경우이다. 설명 제공으로 인해 미치는 영향을 명확하게 분석하지 않은 경우, 자세한 설명을 제공하면 사용자의 의사결정에 더 도움이 된다고 생각할 수 있지만, 예상과는 다르게 혼란을 초래할 수 있다. 예를 들어, 인공지능 시스템이 도출한 두 가지 결과가 있고, 각각의 예측 확률이 85.8%, 87.0%라면, 사용자는 어떤 결과를 활용하여 의사결정을 할지 혼란스러울 수 있다.
  - ✓ 둘째, 예측 확률이 너무 높거나 낮은 경우에도 모델의 추론 결과에 대한 자세한 설명을 제공하지 않는 것이 낫다. 만약 시스템의 출력 결과에 대해 예측 확률값이 100%라고 사용자에게 알릴 경우, 사용자가 시스템의 출력 결과를 맹목적으로 수용하게 만들 수 있다.

10-3b 사용자에게 인공지능 모델 추론 결과에 대한 설명을 제공하였는가?

Yes No N/A
☐ ☐ ☐

- 진단 보조 목적의 인공지능은 의료진의 빠르고 정확한 진단에 활용될 수 있다. 하지만 모델 추론 결과만을 제시하면 입력 데이터의 잡음, 새로운 관측치의 입력, 의료진의 자동화 편향과 같은 인적 요인 등 예상치 못한 문제 발생 시 진단 오류를 발생시킬 위험이 있다[52,54].

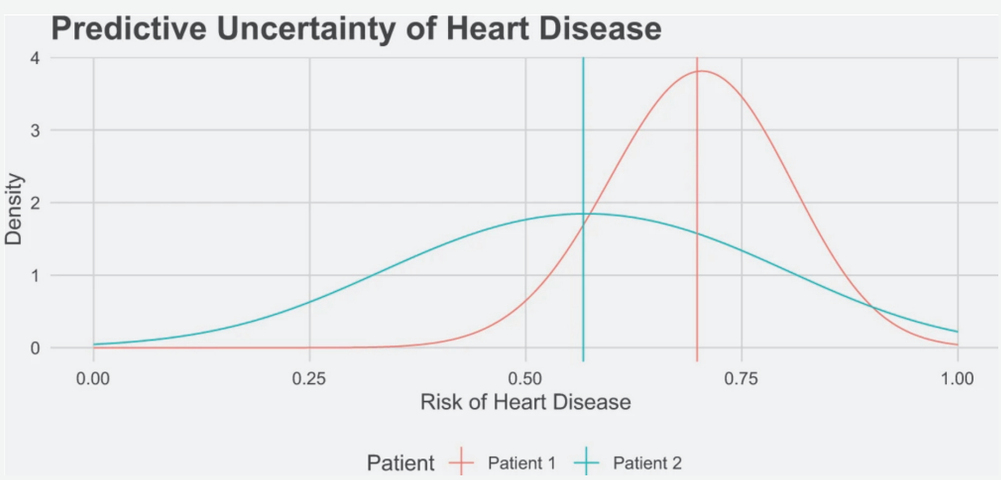
- 이러한 문제를 해결하고자 인공지능 모델 추론 결과에 대한 확률값을 수치화하고, 추론 결과를 얼마나 확신하는지를 나타내는 불확실성을 정량화하여 설명하는 방법을 고려해 볼 수 있다. 다음은 추론 결과에 대한 확률과 추론 결과의 불확실성에 따른 설명 예시이다.

추론 확률(0~1)	불확실성(0~1)	설명 예시
0.98	0.01	• 모델의 추론 확률이 98%로 높고, 추론에 대한 불확실성이 1%로 낮아서, 모델 추론 결과를 신뢰할 수 있음
0.98	0.90	• 모델의 추론 확률이 98%로 높지만, 추론에 대한 불확실성이 90%로 높아서 모델 추론 결과를 신뢰하기 어려움
0.20	0.01	• 모델의 추론 확률이 20%로 낮고, 추론에 대한 불확실성이 1%로 낮아서 모델 추론 결과를 신뢰할 수 있음
0.20	0.90	• 모델의 추론 확률이 20%로 낮지만, 추론에 대한 불확실성이 90%로 높아서 모델 추론 결과를 신뢰하기 어려움

- 인공지능 모델의 추론 확률이 임계치보다 낮거나, 새로운 유형의 관측치 등의 요인으로 불확실성이 높을 때는 추론 결과를 출력하지 않고 '판단 불가'와 같은 설명을 통해 의료진의 전적인 임상 판단이 이루어지도록 할 수 있다[54].

참고 심장질환 위험 예측 불확실성을 계산하는 경험적 분포 표준편차 계산 예시[54]

심장질환 위험 예측 모델을 통해 불확실성을 설명하는 예로 UCI[University of California Irvine] 심장질환 데이터셋을 활용한 모델에서 위험 분포를 계산하였다. 환자 1과 환자 2에 대한 경험적 분포의 표준편차를 계산한 결과 각각 7.6%, 15.3%의 모델 추론 결과 불확실성을 보여 환자 2의 추론 결과에 대해 의료진의 더 많은 주의가 필요함을 알 수 있다.



04 시스템 구현

다양성 존중

요구사항 11 인공지능 시스템 구현 시 발생 가능한 편향 제거

대표행위자 | 시스템 엔지니어 협력 대상 | 시스템 운영자 인공지능 모델 개발자

- 의료 인공지능 시스템 구현 단계에서 개발자의 배경지식이나 편견 또는 의사결정 규칙을 정의하는 과정에서 인공지능 시스템이 전문 의료진에 의해 편향될 수 있다. 따라서 발생 가능한 편향을 식별하고 이를 제거하는 방안을 고려하여 인공지능 모델을 설계한다.

11-1 소스 코드 및 사용자 인터페이스로 인한 편향을 제거하기 위해 노력하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

의료 인공지능 시스템 개발 단계에서 소스 코드를 통해 데이터 접근 방식을 설계하거나 사용자 인터페이스 및 상호작용을 통해 정보를 제공하는 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 의료 인공지능 시스템의 최종 결정은 사용자가 수행하므로, 특정 선택을 암묵적으로 유도하는 사용자 인터페이스 등을 통한 편향이 발생할 수 있다.
- 편향 방지를 위해 의료 인공지능 시스템의 구현 단계에서 코드를 주기적으로 검토하여 개발자의 제한된 배경지식이나 편견이 코드에 반영되지는 않았는지 등을 확인하여야 한다.
- 사용자 인터페이스user interface 및 상호작용interaction 측면에서는 표현 편향presentation bias이나 순위 편향ranking bias 등이 발생하지는 않는지 미리 확인하여 편향을 방지할 수 있도록 시스템을 설계하는 것이 바람직하다.

11-1a 데이터 접근 방식 구현과정 등 소스 코드에서의 편향 발생 가능성을 확인하였는가?

Yes ☐ No ☐ N/A ☐

- 프로그래밍 과정에서 개발자의 의식적 및 무의식적 편견으로 인한 편향 문제가 제기되고 있으며, 인공지능 시스템은 모델에서 활용할 데이터에 접근하는 방식, 알고리즘 규칙, 사용할 변수 등을 구현하는 과정에서 개발자로부터 편향이 발생할 수 있다.
- 개발자가 프로그래밍 과정에서 편향을 갖지 않도록 권고하는 지침은 조직 문화적인 부분에서 도움이 될 수 있다. 마이크로소프트에서는 책임 있는 대화형 AI 개발을 위한 안전 지침을 공개하였으며, 아래는 의료용 챗봇 구축 시 개발자가 준수해야 하는 사항이다.

참고 마이크로소프트 개발자 안전 지침에서의 의료용 챗봇 구축 사례

- 의료 회사에서 근무하는 개발자가 의사와 편안하게 의논하기 어려운 내용을 환자에게 질문하는 봇을 개발한다면, 개발자에게는 환자 정보를 비밀로 유지하고 위험한 의료 조언을 제공하지 않을 윤리적·법적 의무가 있다. 또한, 이 의료 봇을 디자인하는 동안 개발자는 환자 데이터를 안전하게 유지하면서 정규 의료 전문가가 검토해야 하는 유사한 상황에 플래그를 지정하는 방법을 결정해야 한다.

- 또한, 의료 분야 서비스를 위한 인공지능 기반 시스템 구축 시, 의사결정 규칙은 전문 의료진의 지식을 기반으로 정의해야 하며, 개인의 편향을 줄이려면 다양한 전문가를 선정해야 한다.
- 그 외에도 오픈소스 도구(예: FairML, Google What-If Tool)를 활용하여 주기적으로 출력 데이터의 통계를 분석하여 알려지지 않은 편향을 발견하거나, 미리 지정한 공평성 평가지표에 따라 기능의 위험 여부를 알리는 등의 기능을 수행해야 한다. 이 도구들을 활용함으로써 시스템 구현 단계에서 편향을 빨리 발견하여 대응할 수 있다.

11-1b 사용자 인터페이스user interface 및 상호작용interaction 방식으로 인한 편향을 확인하였는가?

Yes ☐ No ☐ N/A ☐

- 의료 인공지능 시스템은 의사결정의 보조적인 용도로 사용되며, 최종 의사결정은 전문 의료진이 수행하는 것이 일반적이다. 이처럼 사용자의 상호작용을 통해 최종 의사결정을 수행하는 제품은 사용자 인터페이스 설계 시 편향 발생 가능성을 면밀하게 확인하여야 한다.
- 도널드 노먼이 언급한 인터렉션 기초 원칙 중 하나인 행동 유동성affordance은 인공지능 시스템에도 적용 가능하다. 이는 인터페이스 디자인 측면에서 사용자가 오류를 최소화하면서 신속하게 작업을 완료할 수 있도록 모양, 색상, 그림자, 대비, 애니메이션 등의 디자인적 요소를 활용하여 상호작용 방법에 대한 단서를 제공한다.
- 그러나, 의도치 않은 행동 유동성은 사용자 상호작용의 편향을 발생시킬 수 있으며, 사용자 인터페이스 설계 및 구현 시 편향 발생 가능성이 있는 요소(예: 표현 편향, 순위 편향)는 확인한 후 개선해야 한다.
  - ✓ 표현 편향: 정보가 표현되는 방식에 따라 발생하는 편향이다. 사용자는 제일 눈에 띄는 콘텐츠를 우선해서 관심을 표현하는 경향이 있다. 예를 들어 인공지능 시스템이 추천하는 진단 결과 몇 가지를 의료 전문가가 검토 후 최종 선택할 때, 실제의 진단 정확도와는 상관없이 가장 크고 선명하게 표시되는 의료 영상이 포함된 진단 결과를 무의식적으로 선택할 수 있다.
  - ✓ 순위 편향: 정보가 노출되는 순서에 따라 발생하는 편향이다. 사용자는 최상위 결과가 관련성이 가장 높고 중요하다고 생각하는 경향이 있다. 예를 들어 의료 전문가가 인공지능 시스템이 추천하는 진단 결과 몇 가지를 검토 후 최종 선택할 때, 최상위 또는 가장 왼쪽에 노출된 진단 결과를 선택하는 빈도가 나머지 결과를 선택하는 빈도보다 높을 수 있다.

책임성 안전성 투명성

요구사항 12 인공지능 시스템의 안전 모드 구현 및 문제발생 알림 절차 수립

대표행위자 | 시스템 엔지니어 협력 대상 | 시스템 운영자 인공지능 모델 개발자 품질 관리자

- 인공지능 시스템을 통해 생성되는 결과나 의사결정은 개인 혹은 사회에 부정적인 영향을 미칠 수 있으므로, 이에 대한 대응이 가능하도록 안전 모드를 구현하고, 문제 발생 알림 절차를 수립한다.

12-1 공격, 성능 저하 및 사회적 이슈 등의 문제 발생 시 대응 가능한 안전 모드를 적용하는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

의료 분야 인공지능 서비스 중에서 문제 발생에 대한 대응이 중요한 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 고장 안전fail-safe은 산업 전반에서 사용되는 일반적 개념으로, 고장이나 오류로 문제가 발생하더라도 안전한 상태를 유지하는 방법 및 기능을 의미한다. 이는 인공지능 시스템에도 적용될 수 있다. 인공지능 시스템에서도 외부의 공격, 인적 오류, 인공지능 모델의 성능 저하, 편향 발생으로 인한 사회적 물의, 사고 등이 예상되는 경우, 이의 발생 원인을 파악하고 해결하거나 사용자에게 정상적인 기능으로 복구할 수 있는 방법을 제시하여야 한다. 이러한 대처 방법이 작동하는 상태를 안전 모드라고 한다.
- 특히, 의료 분야는 인공지능 시스템의 오류 발생 시 환자에게 직접적인 피해를 줄 수 있어 기술적인 안전 모드 구현 외에도 의사결정 과정에 의료진을 포함해 안전성을 확보해야 한다.
- 안전 모드를 구현하는 방법과 예시는 아래와 같다.
  - ✓ 시스템에 문제 발생 시 기능 정지 및 피드백 제공 화면으로 전환
  - ✓ 시스템에 문제 발생 시 서비스 제공 초기 화면 혹은 상태로 복구
  - ✓ 인공지능 판단 결과의 불확실성이 높거나 문제 발생 가능성이 높은 경우, 이에 대한 의사결정을 회피하거나 사용자에게 상황에 대한 안내 제공
  - ✓ 사용자의 악의적인 의도를 파악하고 이에 대한 입력을 거절
  - ✓ 자동 및 자율 운영 중 시스템에 문제 발생 시 사람의 개입 유도
  - ✓ 예상되는 사용자 오류에 대해 안내 및 대응 제공

12-1a 문제 상황에 대한 예외 처리 정책이 마련되어 있는가?

Yes ☐ No ☐ N/A ☐

- 시스템에 문제가 발생하는 상황에서 기능 정지, 화면 전환 및 서비스 제공 초기 상태로의 복구, 입력 거절, 의사결정 회피 등의 예외 처리가 이루어지는지 확인해야 한다.
- 이러한 예외 처리가 이루어지는 경우, 의료진에게 시스템 운영이 적절치 않은 이유와 시스템 대응에 관해 설명을 제공해야 한다.
  - ✓ 예를 들어, 시스템 사용자가 전문 의료진이면 추론 결과에 오류가 발생하면 의료진이 직접 진단이나 처방을 내릴 것을 권고하여야 한다. 모델 추론 결과의 불확실성이 높을 때 정확도나 신뢰도가 낮으므로 주의하여야 한다는 알림을 제공할 수 있다.
  - ✓ 시스템 사용자가 환자일 때 의료 데이터 입력 시 문제 상황이 발생하면 담당 의료진에게 도움을 요청하거나 가까운 의료 기관을 방문할 것을 시스템을 통해 공지하여야 한다. 사용 부주의 또는 기기의 오류로 인해 데이터 입력이 부정확하게 이루어지면 입력을 거절하는 등의 조처를 할 수 있다.

12-1b 인공지능 시스템의 보안 강화를 위한 보안 기법을 적용하였는가?

Yes ☐ No ☐ N/A ☐

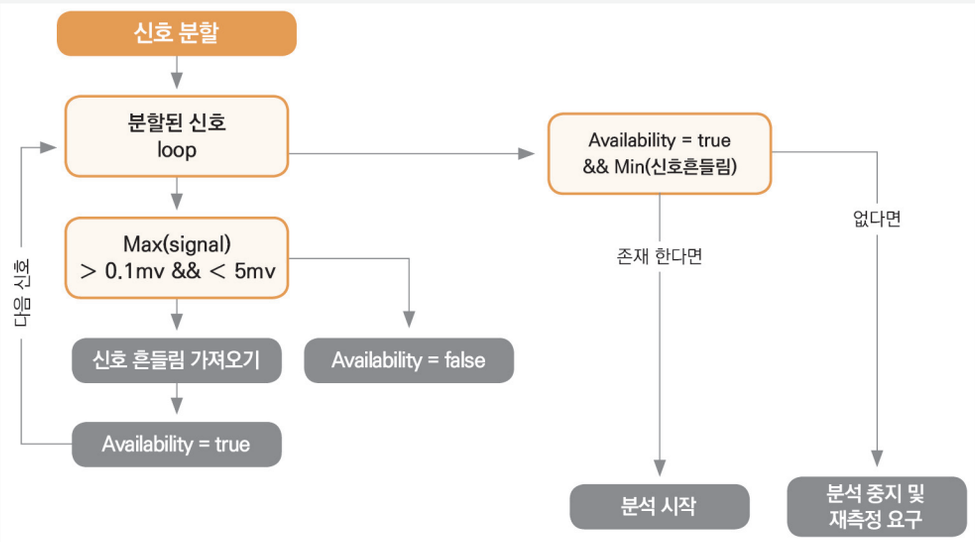
- 대부분 의료기기가 환자 데이터의 교환, 수집 및 저장 등을 위해 원격지 또는 네트워크로 연결된 병원 내 서버, 환자 단말기 등으로 데이터를 전송하므로 일반적인 인터넷 환경에서 발생 가능한 보안 위험과 인공지능 시스템을 고려한 보안 기법을 적용해야 한다[55].
- 인공지능 시스템을 개발할 때 격리 및 탐지 등 보안 기법을 활용한 인공지능 보안 아키텍처와 구축 솔루션을 적용함으로써 인공지능 데이터 및 모델에 대한 보안성뿐만 아니라 인공지능 시스템의 전반적인 보안성을 확보할 수 있다.

인공지능 시스템의 보안 기법 예시

보안 기법	설명 및 예시
격리	의사결정에 활용되는 주요 기능을 모듈 단위로 분리하고 모듈 간 접근제어 기법을 설정하여 인공지능 시스템에 대한 보안성을 확보할 수 있다.
탐지	인공지능 시스템에 대한 공격을 지속해서 모니터링해 네트워크 보안 상태를 종합적으로 분석하고, 현재의 위험 수준을 측정할 수 있다.

Use case M사의 심전도 진단 예외 처리 사례

- 사용자의 제품 사용과정에서 발생할 수 있는 측정 데이터의 불량, 제품 목적을 벗어난 인공지능 모델 추론 결과, 추론 결과의 불확실성 등 문제 상황을 상정하여 예외 처리 정책을 마련함
- 기술적으로는 모델의 성능 임계치를 계산하여 문제 상황에 대한 사용자 안내가 필요한지를 판단함



- 문제 상황으로 판정될 때 사용자에게 모델의 추론 결과, 제품·서비스의 목적, 문제 상황에 대한 안내 및 전문의료진 진단 권고 등의 정보 안내를 제공함

(a) 정상 상황 시 사용자 화면

(b) 문제 발생 시 사용자 화면

상황별 사용자 안내 정보

참고 부정적 영향을 줄 수 있는 입력 데이터 감지 솔루션의 필요성[56]

- 치료 지침을 제시하는 왓슨 포 온콜로지watson for oncology의 학습 과정에서, 수많은 자료 중 학습 자료로 활용할 자료를 선별하고자 많은 시간을 할애하여 수작업이 수행되었다. 정확하지 않은 자료가 입력되면 출력 결과에 혼선을 초래할 위험성이 있으며, 부정적인 영향을 줄 수 있는 공격성 입력 데이터가 유입되면 시스템에 악영향을 끼치기 때문이다. 따라서, 부정확하거나 공격성이 있는 데이터를 감지하는 솔루션을 도입하면 의료 인공지능 시스템의 보안성을 크게 강화할 수 있다.

12-1c 인공지능 시스템의 불확실성을 완화하기 위해 사람을 포함한 의사결정 방식을 고려하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템의 활용 시 진단을 목적으로 한 의학적 의사결정은 의료진이 전면적인 통제권을 유지하도록 하고 있다[52]. 또한, 의학 지식은 지속해서 변화하므로 시스템의 불확실성을 완화하는 방법 가운데 하나로 시스템-의료진 간 상호작용을 통한 의사결정 방식을 고려할 수 있다[57].
- 이는 대화형 모델을 구현하여 의료진의 도움을 받아 최종 추론 결과를 출력함과 동시에 추론 결과의 불확실성을 보완할 수 있는 인공지능 모델 측면의 접근 방식이 있을 수 있으며, 올바른 데이터가 모델에 입력될 수 있도록 의료진의 확인을 유도하는 시스템적 접근 방식이 있을 수 있다[57].
- 이처럼 예외 처리 및 보안 기법 외에 사람이 직접 또는 부분적으로 의사 결정 과정에 포함하여 인공지능 사용으로 인한 불확실성을 완화하는 방안을 고려하여야 한다.

12-1d 예상되는 사용자 오류에 대한 안내 및 대응을 제공하는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 기술이 적용된 진단 보조 의료기기의 사용자는 의료진과 환자로 구분할 수 있다. 의료진은 의료 행위의 주체로서 진단 보조 의료기기를 사용할 수 있으며, 환자는 진단에 사용될 데이터 수집을 위해 일정 기간 데이터 수집 장치를 대여하여 사용할 수 있다.
- 이때 인공지능에 대한 기술적 이해도가 낮은 사용자로 인한 인적 오류가 발생할 수 있다[51,52]. 서비스 담당자는 다양한 사용자 오류 유형을 이해하고, 이와 관련하여 발생할 수 있는 오류를 사전에 정의하고, 분석하여야 한다. 의료 인공지능 시스템에서 발생할 수 있는 사용자 오류의 예는 다음과 같다.
  - ✓ 사용자가 의료진일 때: 데이터 입력 또는 측정 오류, 시스템 출력의 오인식(음성 진단 결과를 양성으로 인식) 등
  - ✓ 사용자가 환자일 때: 잘못된 측정 자세나 사용법 미숙지로 인한 측정 데이터 오류, 기기 관리 미숙으로 인한 하드웨어 오류 유발 등
- 사용자 오류에 따른 사전 대응 방안의 예시는 다음과 같다.

✓ 제약조건 설정: 잘못된 사용자 입력을 막고자 사용자의 선택을 어느 정도 제약하거나 수용 가능한 옵션을 정의하여 보여 주는 것을 말한다.

✓ 시스템 제안·정정: 자주 발생하는 사용자의 실수를 수집하고, 실제 서비스 시 유사한 사용자 실수가 발생한다면 시스템에서 정정을 유도하거나 올바른 입력을 제안한다. 예를 들어 측정 실수가 빈번한 데이터를 수집할 때 이상치를 설정하여 재측정을 제안할 수 있다.

✓ 기본값 설정: 시스템에서 필수적으로 자주 사용되는 값을 기본값으로 먼저 제공하거나 관련 예시를 제공하여 사용자 실수를 줄일 수 있다. 예를 들어 의료진을 위한 진단 보조 의료기기는 의료진이 진단 결과를 정확하게 이해할 수 있도록 양성인 부분만 다른 색상이나 그림으로 표현하여 기본값(음성)에서 벗어난 부분을 표시하거나, 가장 신뢰도가 높은 진단 결과를 최상단에 제공할 수 있다.

✓ 재확인·결과 제공·실행 취소: 사용자에게 전달받은 입력 등을 재차 확인하고 그에 대한 예상 결과를 미리 전달한다. 또한 잘못된 결과에 대해 실행을 취소하는 등의 기능을 포함해 예방할 수 있다. 예를 들어 투약 처방을 보조하는 의료기기는 의료진의 최종 처방이 정확한지 재확인하는 알림을 출력할 수 있다.

12-2 인공지능 시스템에서 문제가 발생할 경우, 시스템은 이를 운영자에게 전달하는 기능을 수행하는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

예측/판단을 수행하는 인공지능 서비스의 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 의료 인공지능 시스템은 성능 저하, 공격 등의 문제 발생 시 환자의 건강과 생명에 직접적인 영향을 미칠 수 있다. 따라서 시스템 운영자가 이를 파악할 수 있도록 하는 자체적인 점검 기능이나 문제 상황에 대한 의견을 전달할 수 있는 기능을 제공하여야 한다.
- 시스템의 자체적인 점검 기능은 서비스 성능 저하나 외부 공격에 대한 검사 등을 수행한 후 가능한 범위 내에서 이에 대응하고, 해당 사실을 시스템 운영자에게 전달할 수 있는 체계를 갖춰야 한다.
- 사용자 의견 전달 기능은 시스템의 일시적인 오류나 도출 결과에 편향이 발생하는 등 문제가 생길 때, 사용자가 해당 사실을 시스템 운영자에게 전달할 수 있는 체계를 갖춰야 한다.

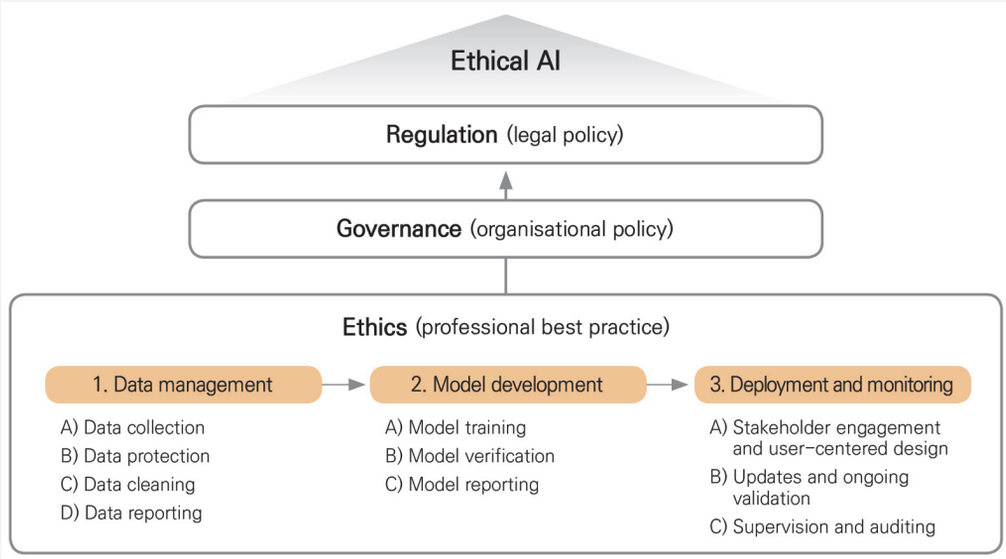
12-2a 편견, 차별 등 윤리적 문제에 대한 알림 절차를 수립하였는가?

Yes ☐ No ☐ N/A ☐

- 의료 인공지능 시스템에서 편견 혹은 차별 등 윤리적 문제의 발생 가능성을 확인하고, 문제 발생 시 이를 위한 알림 기능 혹은 절차가 수립되었는지 점검한다.
- 윤리적 문제 알림 절차의 경우, 먼저 인공지능 시스템에서 자체적인 신뢰 정도를 평가할 수 있는 기준과 점검 항목을 만든다. 주요 점검 항목의 예시는 다음과 같다.
  - ✓ 인권 보장, 사생활 보호, 다양성 존중, 침해 금지, 공공성, 연대성, 의료 데이터 관리, 책임성, 안전성, 투명성, 라이선스 관리, IRB 승인 등
  - ✓ IRB의 생명윤리 및 안전에 관한 법률에서 살펴볼 수 있는 항목은 다음과 같다. 〈인간대상 연구〉, 〈인체 유래물 연구〉, 〈배아 등의 생성 및 관리〉, 〈배아 등을 이용한 연구〉, 〈유전자 검사 및 유전자 치료〉 등이다.
  - ✓ 한국인터넷진흥원의 EU 인공지능 윤리 가이드라인 연구에서 신뢰할 수 있는 AI로서의 구성 요소와 근거, 실현 요건, 평가 등을 제시하고 있다.
- 시스템 자체 점검 기능 외에도, 시스템 운영 중 사용자가 윤리적 문제를 발견할 경우 시스템 운영자에게 의견을 전달할 수 있는 기능도 개발되어야 한다.

참고 의료 인공지능의 윤리적 운영을 위한 프레임워크 – 윤리 문제 알림 절차 예시[58]

- 의료 인공지능은 의료 서비스 개선하는 많은 가능성을 제공한다. 하지만, 제대로 관리되지 않으면 소수 집단 간의 불평등을 악화시키거나 민감한 의료 데이터의 기밀성을 손상할 수 있다. 다음은 의료 인공지능에서 윤리적 문제가 발생한 사례이다.
  - ✓ 흑색종이 아프리카인들에게 더 치명적인데도 불구하고, 현재의 흑색종 감지 인공지능 알고리즘은 주로 흰 피부 이미지 데이터에 의해 학습되어 피부가 어두운 사람들의 흑색종을 진단하는 데는 부정확한 결과를 보인다.
- 따라서, 발생 가능한 윤리적 문제를 사전에 고려하여 의료 인공지능을 개발해야 하며, 윤리적 문제 발생 시 이를 사용자 또는 관리자에게 알릴 수 있는 절차가 필요하다. 인공지능 생명주기에 따른 의료 인공지능 윤리 프레임워크로 아래의 예시를 참고할 수 있다. 본 프레임워크에서는 3단계(데이터 관리, 모델 개발, 배포 및 모니터링)에 걸쳐 윤리적 문제를 관리할 수 있는 절차를 제시하고 있다.



의료 인공지능 윤리 프레임워크 및 단계별 윤리 문제 솔루션/가이드라인

단계	세부단계	솔루션/가이드라인
1. 데이터 관리	데이터 보고	• 데이터 특성을 보고하여 투명성을 높이고 사회적 편견을 조명
		• 데이터 배포 조건에 부합하게 데이터 특성을 사용하도록 하여 투명성 증진과 데이터 사용 장려
		• 데이터셋을 인프라로 취급하여 소프트웨어 엔지니어링에서 사용되는 것과 유사한 문서 개발
		• 의료 데이터 공개 시 개인정보보호 알고리즘 사용
		• 투명성을 증진하고 대중 인식을 개선하기 위해 적절한 템플릿을 사용하여 의료 인공지능 서비스 정보 공유
		• 데이터셋 설명을 통해 데이터 편향을 보고하고 투명한 모델 개발
2. 모델 개발	모델 보고	• 모델의 성능과 의도된 사용 사례 등을 보고하여 투명성 보장
		• 모델 보고 프레임워크 도구를 사용하여 사용자 및 기타 이해관계자의 투명성 요구에 응답
3. 배포 및 모니터링	감독 및 감사	• 잠재적·부정적 사회적 영향을 피하기 위해 지속해서 내부 감사를 실시하여 윤리적 거버넌스 보장
		• 시정 조치를 위한 윤리 감사 실시

12-2b 시스템 성능 저하를 평가하기 위한 지표 및 절차를 설정하고 알림 절차를 수립하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템의 경우, 서비스 배포 및 운영 단계에서 일반적인 소프트웨어와 달리 지속적인 데이터 축적, 서비스 기능 확장, 환경의 변화 등의 이유로 성능 변화가 생길 수 있다.
- 인공지능 시스템은 실제 서비스 운영 중 갑자기 성능이 저하되었을 때 원인을 바로 알기 어려우므로, 시스템의 성능 저하를 지속해서 평가, 관리하기 위한 지표와 절차가 시스템에 포함되어야 한다. 시스템 성능 점검 결과 성능 저하 발견 시, 사용자와 시스템 운영자에게 관련 정보를 알리는 절차를 마련하여야 한다.
- 의료 분야에서 선정할 수 있는 대표적인 성능 지표로는 F1-score, IoU[Intersection over Union], mAP[mean Average Precision], accuracy, recall, specificity, precision, threat score, true positive, true negative, false positive, false negative 등이 있다. 사용한 성능 지표에 대한 해석이 가능하도록 의료진 교육이 필요하다. 평가 결과 성능 저하가 확인되면 이를 시스템 운영자에게 전달하고, 운영자는 성능 저하 원인을 찾아 개선을 진행하는 등의 절차를 마련해야 한다. 의료 분야 특성상 시스템 구현 시 판단 오류나 판독 불가 등 성능 저하가 일어나면 의료진이 직접 데이터를 다시 검토하여 재평가를 시행한다.
- 성능 개선을 통한 업데이트 후에는 식품의약품안전처의 〈인공지능 의료기기의 허가·심사 가이드라인〉 내 변경 허가 및 인증 관련 규정을 참고하여 재심사 여부를 검토하여야 한다. 알고리즘상의 변수 변경 필요 시 데이터의 활용 목적, 활용 기간, 데이터셋 유형, 데이터셋 규모 및 용량, 데이터셋 특성 등을 고려하여 변수 추가 또는 특성 변경에 대한 충분한 설명과 함께 학술적 근거를 명시하여야 한다.

참고 인공지능 학습 모델의 성능 저하 방지를 위한 유의사항 예시

- 현재 인공지능 모델 성능을 모니터링할 수 있는 다양한 도구가 개발되어 있고, 이러한 도구는 모델 성능을 개선하는 데 도움이 되는 유용한 통계 및 모델 성능에 대한 세부 정보를 제공한다.
- 지속 변화하는 의료 데이터에 대한 인공지능 모델의 성능을 유지 및 개선하도록 정기적인 성능 모니터링을 통해 모델의 성능을 점검하고 업데이트하여야 한다.
- 하지만 이미 학습되고 테스트가 완료된 모델을 다시 업데이트하는 것은 시간 및 비용 문제를 수반하며, 업데이트하더라도 모델의 정상 작동 및 성능 향상을 완벽히 보장할 수는 없다는 위험도 존재한다. 따라서 모델의 취약점을 면밀히 파악하여, 모델의 성능을 해치지 않는 적절한 전략적 모니터링 방법의 설계가 필요하다.
- 예를 들어, 모니터링 결과 성능 저하의 원인이 과적합일 때 정규화 초매개변수를 조정하여 이미 설계된 모델을 다시 학습할 수 있다. 이때는 기계학습 알고리즘 선택, 비용 함수 등의 문제를 고려할 필요가 없다.
- 또한, 다수의 기본 모델로 구성된 앙상블 모델일 때는 기존에 설계한 앙상블 접근 방식에 따라 기본 모델 중 일부를 삭제하고 새로 학습된 기본 모델로 교체하여 재학습함으로써 업데이트할 수 있다.

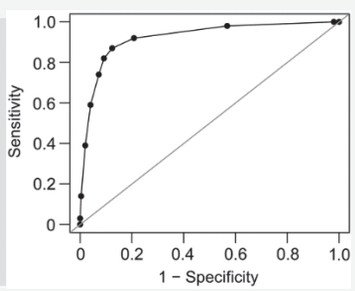
참고 인공지능의 임상 성능 및 유용성 검증 방안 예시[59]

- 의료 인공지능은 허위 진술이 발생할 수 있어 임상 성능 및 유용성을 검증해 기술적 건전성을 지속해서 입증하여야 한다.

□ 진단 또는 예측 성능평가를 위한 통계적 방법

- 진단 테스트나 알고리즘의 결과가 이분법적일 때, 판별 성능은 일반적으로 민감도(모든 질병 양성 피험자 중 검사 양성 피험자의 비율)와 특이도(전체 질병 음성 피험자 중 검사 음성 피험자의 비율) 측면에서 측정된다.
- 알고리즘 결과를 이분법적으로 제시하더라도 서로 다른 임곗값threshold을 적용하여 연속 출력을 범주형 결과로 축소한다. 즉, 연속 출력에 서로 다른 임곗값 수준을 적용하여 특이도, 민감도 값의 쌍을 얻을 수 있다.
- 특이도, 민감도를 각각 (x, y) 좌표로 사용하여 ROC 곡선을 추정한다.

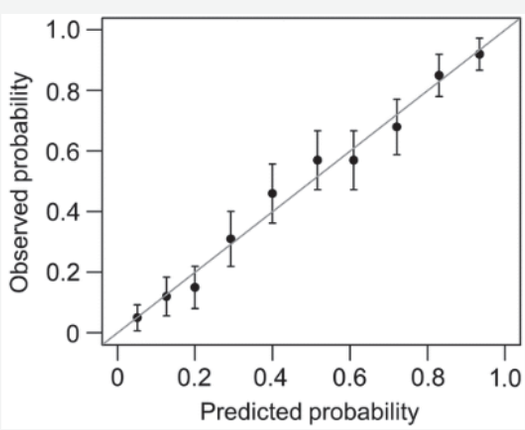
Criterion to Diagnose Lung Cancer	Sensitivity	Specificity	1 – Specificity
≥0.1	100.0	2.0	98.0
≥0.2	98.0	43.2	56.8
≥0.3	92.0	79.2	20.8
≥0.4	87.0	87.6	12.4
≥0.5	82.0	90.8	9.2
≥0.6	74.0	92.8	7.2
≥0.7	59.0	96.0	4.0
≥0.8	39.0	98.0	2.0
≥0.9	14.0	99.6	0.4
≥1.0	3.0	100.0	0.0



- ROC 분석을 사용하여 인공지능 알고리즘의 연속 출력을 이분법적 결과로 전환하는 최적의 컷오프cut-off 값을 찾을 수 있다. 컷오프 값은 민감도와 특이도에 동일 가중치가 부여되고, 질병 유병률이 50%이며, 다양한 결정에 따른 비용이 무시될 때만 최적이라는 점에 유의하여야 한다.

□ 보정 플롯calibration plot을 활용한 예측 확률과 실제 확률 사이의 적합도 평가 방법

- 각 피험자(피험자는 질병 또는 비질환 상태)를 유사한 예측 확률에 따라 십분위수로 그룹화한다. 그런 다음 각 십분위수의 평균 예측 확률을 x좌표로 하고, 동일 십분위수의 실제 확률(실제로 질병을 앓았던 십분위수의 피험자 수를 해당 십분위수의 모든 피험자 수로 나눈 값)을 사용하여 예측 확률(x 축), 실제 확률(y 축) 플롯을 그린다.



- x 축은 각 십분위수에 대한 평균 예측 확률값을 나타내고, y 축은 각 십분위수에서 해당하는 관측 확률을 나타낸다. 오차 막대는 평균 예측 확률의 95% 신뢰 구간을 나타낸다.

□ 모델 성능 평가를 위한 내·외부 데이터셋 사용

- 모델을 개발하는 데 사용된 데이터로 진단 또는 예측 모델의 성능을 평가하는 것을 내부 검증이라고 한다면, 모델 개발에 사용되지 않은 별도의 데이터로 성능을 평가하는 것을 외부 검증이라고 한다.
- 내부 검증은 이미 사용한 데이터로 다시 테스트하므로 성능을 과대평가할 수 있다. 따라서 외부 검증을 통해 진단 예측 모델을 검증하는 것이 중요하다.
- 즉, 모델 성능 검증 과정에서 과적합 및 스펙트럼 편향으로 인한 결과의 과대평가를 피하려면 대상 환자 모집단을 적절하게 나타내는 임상 코호트의 외부 데이터를 사용해야 한다.

참고 SaMD로서의 AI/ML 기반 소프트웨어 수정을 위한 규제 프레임워크[5]

- 적응형 학습Adaptive Learning 접근 방식에 의존하는 SaMD 제품은 새로운 데이터를 통합하고 실시간으로 학습할 수 있는 잠재력을 가져 제품의 위험 수준이나 성능도 빠르게 변할 수 있다. 이러한 변화의 속도와 예측할 수 없는 특성을 감안할 때, SaMD의 알고리즘이 의도된 용도에 대해 여전히 안전하고 효과적인지 확인하기 위해 기관의 추가 검토가 필요할 수 있는 시기를 결정하기 어려울 수 있다.
- 따라서 FDA는 적응형 학습 접근 방식에 의존하는 SaMD 제품의 문제를 해결하는 잠재적 접근 방식으로, 다음과 같이 네 가지 원칙의 프레임워크를 백서로 소개하였다.

❶ 품질 시스템과 우수한 기계 학습 관행에 대한 명확한 기대치
: SaMD 개발자가 해당 제품이 관련 품질 표준을 충족하고 규정을 준수하는지 확인 가능한 시스템을 갖출 것을 제안한다. 또한, 개발자는 GMLPGood Machine Learning Practices로 알려진 알고리즘 개발 시 확립된 모범 사례를 구현해야 한다.

❷ SaMD 제품의 시판 전 평가
: 개발자는 AI/ML에 의존하는 SaMD의 초기 시판 전 검토의 일부로 향후 수정 계획을 제출할 수 있다. 이 계획에는 발생할 수 있는 예상 수정 유형과 개발자가 관련 위험을 줄이기 위해 사용할 접근 방식이 포함된다.

❸ 알고리즘 변경에 FDA 검토가 필요한 시기를 결정하기 위한 제조업체의 SaMD 제품에 대한 정기적 모니터링
: 기존 SaMD 제품이 변경제어계획의 범위를 벗어나지만 새로운 용도로 사용되지 않을 때(예: 개발자가 SaMD를 다른 데이터 소스와 호환되도록 만들거나 다른 유형의 데이터를 통합) FDA는 변경 관리 계획만 검토하고 새 버전을 승인하며, 수정 사항이 새로운 용도로 이어질 때(예: 대상 환자 모집단을 성인에서 어린이로 확장) FDA는 추가 시판 전 검토를 요구할 수 있다.

❹ 투명성 및 실제 성능 모니터링
: FDA는 SaMD 개발자가 특정 투명성 원칙을 준수하고 지속적인 모니터링을 요구하며, 개발자는 다른 요구사항 중에서 구현된 업데이트 및 성능 지표를 정기적으로 보고해야 한다.

투명성

요구사항 13 인공지능 시스템의 설명에 대한 사용자의 이해도 제고

대표행위자 | 시스템 엔지니어 협력 대상 | 시스템 기획자 시스템 운영자 인공지능 모델 개발자 비즈니스 결정권자 전문 의료진

- 모델의 추론 결과에 대해 설명을 제공하는 기법을 적용하여도 사용자가 바로 이해해 해석하기 어려운 경우가 많다. 따라서 인공지능 시스템의 운영자 혹은 서비스 제공자는 사용자에게 제공되는 결과가 이해 가능한지understandability, 해석 가능한지interpretability, 설명 가능한지explainability를 평가한다. 의료 인공지능 시스템은 사용자가 의학적 배경지식을 갖춘 전문 의료진일 수도 있고, 그 반대인 환자나 일반인일 수도 있으므로, 여러 사용자의 경우를 종합적으로 고려한다.

13-1 인공지능 시스템 사용자의 특성user characteristics과 제약사항을 분석하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

의료 인공지능 시스템의 구현 시 다양한 사용자에게 정보를 제공하는 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 인공지능 시스템의 결과가 적절한지 평가하기 위해서는 먼저 해당 결과를 읽는 사용자를 고려해야 한다. 사용자가 누구인지에 따라 결과(설명)의 수준, 깊이 그리고 맥락이 정해지는 만큼 사용자에 대한 자세한 분석이 수행되어야 한다.
- 의료 인공지능 시스템의 사용자는 시스템의 개발 목적에 따라 전문 의료진 그리고 환자를 포함한 일반인 등의 두 부류로 나뉠 수 있으므로 각 사용자의 특성 및 제약사항을 분석하여야 한다. 즉, 사용자가 의학적 배경지식을 갖추었는지 그리고 그에 따른 시스템의 설명 내용에 일정 정도의 의학적 전문성이 요구되거나 불필요한지 등 여러 사용자의 경우를 종합적으로 고려하여야 한다.

13-1a 사용자 특성에 따른 세부 고려사항을 분석하였는가?

Yes ☐ No ☐ N/A ☐

- 의료 서비스 기획 단계에서 사용자의 선호도와 요구사항needs에 집중했다면, 설명의 적절성을 평가하기 위해서는 각 사용자의 다양한 특성을 고려하여야 한다. 예를 들어, 서비스 사용자의 의학적 배경지식이나 장애 유무에 따른 설명 이해력의 차이를 고려해야 한다.
- 사용자 특성 분석을 위해 고려해야 할 요소의 예시는 다음과 같다.

사용자 특성 분석을 위한 고려 사항 예시

구분	상세 구분	고려 사항
연령	아동, 성인, 노인 등	아동 또는 노인은 성인과 비교해 이해할 수 있는 어휘, 단어에 한계가 있을 수 있음
성별	남성, 여성 등	자신이 해당하지 않는 성별의 의료 정보에 대한 이해도가 낮을 수 있음
인종	아시아인, 유럽인 등	인종마다 피부색이나 신체 크기를 인식하는 평균적인 기준이 다를 수 있음
장애 유무	장애인, 비장애인	신체 크기, 신체 능력, 인지능력 등의 차이 또는 제약으로 인해 의료 정보에 접근하거나 의료 서비스를 받는 데 한계가 있을 수 있음
지식수준	환자, 의료진 등	관련 의료 서비스의 경험 여부와 의학적 배경지식의 차이로 이해도가 다를 수 있음

13-2 사용자 특성에 따른 충분한 설명을 제공하는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

의료 인공지능 시스템의 상태 및 출력 결과에 대해 사용자 대상의 설명이 요구되는 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 서비스를 이용하는 사용자는 다양하여 인공지능 시스템의 결과가 서로 다른 입장에서 설명이 해석되고 오해가 생길 수 있다. 따라서 13-1 에서 분석된 사용자 특성을 고려하여 설명을 평가할 수 있는 기준 항목을 수집한다. 설명 평가의 기준으로는 명확성, 구체성, 정확성 등을 고려할 수 있다.
- 의료 인공지능 시스템의 사용자가 전문 의료진일 때, 인공지능 시스템의 예측 결과에 관하여 전문적인 설명을 제공할 수 있다.
- 그러나 사용자가 환자 또는 일반인이라면 연령, 성별, 인종, 장애 유무, 지식수준 등에 따라 사용자의 특성을 세분화하여 설명 평가 기준에 따라 설명을 제공하여야 한다. 이를 위해서는 다양한 사용자의 개인 정보를 수집해야 하는데, 장애 유무나 지식수준 등에 대한 개인 정보를 얻기는 어려운 현실이다. 따라서, 일반 성인을 기준으로 설정한 설명이 주로 제공되고 있다.

13-2a 사용자 특성에 따른 설명 평가의 기준을 수립하였는가?

Yes ☐ No ☐ N/A ☐

- 다양한 사용자가 서비스를 이용하는 만큼 설명을 포괄적으로 평가할 수 있는 기준과 세부 항목을 정하는 단계가 필요하다. 설명의 평가 기준은 명확성, 구체성, 적절성, 정확성 등의 항목이 될 수 있다. 세부 항목으로 데이터 유형data type이나 모달리티modality에 따라 각 항목에서 고려되어야 할 내용이 달라질 수 있다.

- 예를 들어 사용자가 의료진일 때, 인공지능 시스템이 제공한 진단 결과를 의료진이 수용할지 거부할지 언제든 즉각 판단할 수 있도록 분석 데이터 정보, 가능성이 큰 진단 예측 결과 정보, 진단 가능성이 적다고 판단하여 진단 예측 결과에서 제외한 항목 정보 등에 관해 설명 기준을 수립하여 평가하여야 한다[60].

사용자 특성에 따른 설명 평가 항목 예시

구분	평가 항목
명확성	• 사용자에게 다른 오해를 불러일으킬 만한 표현·단어·어휘는 없는가? • 불필요한 설명이 있지는 않은가? • 해당 설명에 사용자가 기대하고 얻으려는 정보가 모두 들어있는가? • 해당 설명을 통해 진단이나 처방의 이유를 쉽게 파악할 수 있는가?
구체성	• 사용자의 구체적 행동을 이끌 수 있도록 명확한 주어·목적어·동사를 활용해 설명하는가?
적절성	• 주어진 설명이 사용자의 특정 지식수준을 요구하지는 않는가? • 의학적 배경지식 혹은 특정 의료 서비스 경험이 필요하지는 않은가? • 설명이 사용자에게 유용한가? • 독자를 고려한 전문 용어, 약어에 대한 설명을 제공하는가? • 설명이 제공되는 시점이 적절한가?
정확성	• 설명이 함께 제공되는 자료의 그림과 모두 일치하는가? • 사전에 제공된 예상 결과의 설명과 실제 결과가 모두 일치하는가? • 내부 알고리즘과 정확히 일치하는 설명인가?

13-2b 사용자 특성에 맞는 적합한 용어를 사용하였는가?

Yes ☐ No ☐ N/A ☐

- 텍스트를 통해 설명하는 경우, 다양한 독자를 배려해 전문 용어를 최대한 지양하고 필요한 경우, 용어에 대한 설명을 추가로 작성해 주는 것이 바람직하다. 그 예로 자연어 처리 기술 중 문장 내 특정 단어를 사용자 수준에 맞춘 적절한 단어로 변환하는 기술을 인터페이스에 적용할 수 있다.
- 그러나 사용자가 전문 의료진일 때, 전문가로서 충분히 이해할 수 있는 수준의 전문 용어라면 오히려 사용을 권장하는 것이 바람직하다. 용어의 변경으로 진단 결과가 조금이라도 달라질 수 있는 위험을 줄여야 하기 때문이다. 이때 특정 용어의 번역어가 혼재하여 문제가 발생할 수 있으므로, 정확한 학술적 명칭을 원어 그대로 병기하여야 한다.

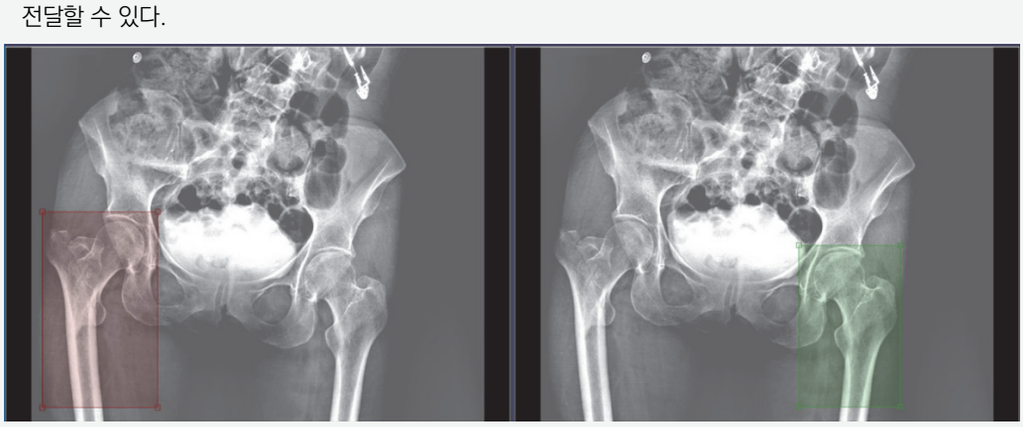
13-2c 사용자의 구체적인 행동과 이해를 이끌어낼 수 있도록 명확한 표현을 사용하였는가?

Yes ☐ No ☐ N/A ☐

- 좋은 설명은 사용자로부터 구체적인 행동과 이해를 이끌어낼 수 있어야 한다. 따라서 설명을 간결하고 명확하게 함으로써 모호하게 해석되지 않도록 작성하는 것이 중요하다.
- 시각적으로는 아래 참고와 같이 진단 결과에 따른 발병 위험의 정도에 대한 성공·실패·경고·위험 등 결과에 따른 색상을 일관성 있게 유지해 줌으로써 사용자가 한눈에 시스템 결과를 이해하도록 도울 수 있다. 그리고 텍스트나 음성으로 제공되는 설명에서는 지시 대명사를 되도록 사용하지 않고 대상을 명확하게 말해줄 필요가 있다. 또한, 비슷한 발음이 연이어 정확한 청취가 어려우면 다른 단어로 대체하는 것이 바람직하다.
- 진단이나 처방을 단순 보조하는 의료기기 및 소프트웨어는 환자가 진단이나 처방 결과를 그대로 전부 수용하지 않도록 전문 의료진의 최종 판단에 따라야 함을 안내하여야 한다.
- FDA에서는 직관적인 인공지능 장치를 만들어 의료 시스템에 대한 수용성과 통합성을 증대할 수 있음을 안내하고 있고, 이때 최종적으로 사용자의 안전을 확보할 수 있도록 IEC TR 62366-2:2016 – Guidance on the application of usability engineering to medical devices의 사용 적합성 엔지니어링Usability Engineering을 적용할 것을 권고한다[61].

참고 사용자를 위한 시각화 방법 예시[62]

- 골절과 비골절을 표현할 때 사용자 인터페이스 그래픽graphical user interface을 활용하여 골절은 붉은색, 비골절은 초록색 세그멘테이션 라벨링을 수행하는 등 색을 구분해 명확한 표현을 제시함으로써 더욱 효과적으로 정보를 전달할 수 있다.



참고 FDA의 직관적인 AI 장치 인터페이스 가이드[61]

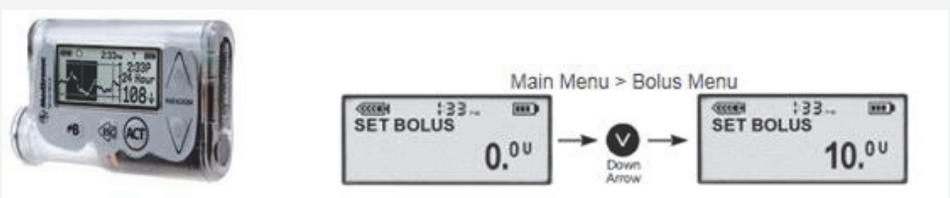
- 의료 시스템에 대한 수용성과 통합이 증가하도록 직관적인 AI/기계학습 장치를 만든다.
- 전문가들은 직관적인 디자인을 전달하는 데 필요한 사용자 인터페이스의 다음과 같은 몇 가지 기능을 식별하여 적용을 권고하고 있다.

- ✓ 발견 가능성: 필요할 때 찾을 수 있다.
- ✓ 행동 유동성: 설계가 주어진 기능에 대해 특정 작업을 수행하는 방법을 적절하게 제안한다.
- ✓ 반응형 피드백: 특정 행동의 의미와 예상 결과는 사용자에게 명확하고, 결과를 즉각적으로 전달한다.
- ✓ 예측 가능성: 기대에 부응한다.
- ✓ 이해 가능성: 잘 이해된다.
- ✓ 효율성: 사용과정에서 불필요한 상호작용과 반복을 피한다.
- ✓ 탐색 가능성: 사용자가 실수를 두려워하지 않고 의료기기를 사용할 수 있게 한다.
- ✓ 용서성: 실수했을 경우 쉽게 복구할 수 있다.

• 의료기기의 설계에 인적 요소 원칙을 통합하여 최종 사용자에게 다음과 같은 많은 이점을 제공할 수 있다.

- ✓ 기술의 권장 사용법을 더 쉽게 준수하게 할 수 있다.
- ✓ 기기 출력 및 기능을 더 쉽게 이해할 수 있다.
- ✓ 직관적인 설계로 고객 지원에 대한 요구를 줄일 수 있다.

참고 FDA에서 사용 적합성 관련 리콜조치 명령을 내린 대표적 의료기기 사례[63]



• 인슐린 펌프

- ✓ 사용 목적: 당뇨병 환자의 체내에 개별적으로 정확한 용량의 인슐린을 주입
- ✓ 기기 결함: 최솟값 0에서 아래쪽 화살표를 누르면 최댓값이 설정됨
- ✓ 리콜 사유: 인슐린 주입량을 낮추려는 환자들이 원하는 의도와는 달리 최댓값이 설정되어 약물이 과다 주입됨(저혈당 쇼크 사망)

13-2d 설명이 필요한 위치와 타이밍은 적절한가?

Yes No N/A
☐ ☐ ☐

• 잘 작성된 설명이 적절한 위치 및 타이밍에 나타나 이해를 돕는 것도 중요하다. 이를 위해 설명이 단발성이어야 하는지, 여러 번 반복하여 강조시켜야 할지 숙고하고, 어느 위치에 놓여야 사용자가 잘 읽을 수 있을지 고려하는 것이 필요하다.

• 또한, 의료기기에서 알람 또는 알림 등을 통한 경보는 종종 환자, 사용자 및 제삼자의 피해를 방지할 수 있다. 경보는 전문 의료진의 주의 또는 응답이 필요한 잠재적 또는 실제 위험 상황(환자, 운영자, 장치 등)을 나타낸다. 환자의 손상 범위는 일반적으로 시간이 지남에 따라 증가하기 때문에 심각한 손상을 방지하는 데 필요한 여유를 제공할 수 있다. 따라서, 이러한 진단 결과에 대해 알림 제공 기준 및 타이밍을 표준 등에 기반하여 반영하여야 한다.

- 이와 더불어 작성된 설명의 위치와 타이밍이 적절한지를 조사하기 위해서는 13-2e 의 웹로그 분석, A/B 테스트 A/B testing 등 사용자 조사user research 기법을 활용할 수 있다.

참고 의료기기의 경보 시스템과 해당 문서화 및 테스트 표준 적용 사례[64]

- IEC 60601-1-8 – General requirements, tests and guidance for alarm systems in medical electrical equipment and medical electrical systems: 경보 시스템과 해당 문서화 및 테스트에 대한 요구사항을 제시하고 있다. 제조업체에 사양 및 의료기기 설계에 대한 구체적인 지침을 제공한다.
  - ✓ 표준 요구사항 개요
    - 경보의 긴급성 처리(우선순위)
    - 알람 신호 전달 방법(시각적, 청각적)
    - 경보 상태에 대한 경보 동작(지연, 억제 등)
    - 라벨링 및 첨부 문서
    - 경보 시스템 테스트 및 검증
  - ✓ 표준에서 다루지 않는 사항
    - 경보 시스템의 필요성(위험 분석 또는 개별 표준에서 발생)
    - 경보 상태를 트리거하는 상황
    - 우선순위에 대한 경보 할당(때로는 개별 표준에서 발생)
    - 알람 발생 기술(예: 피에조 또는 라우드 스피커)
- IEC 60601-1-8 의 요구사항
  - ✓ 경보 발생을 포함한 제품 요구사항
    - 음량
    - 음조
    - 기간
    - 반복
    - 우선순위
  - ✓ 문서 요구사항
    - 경보가 위험을 관리하는 데 사용해야 하는 측정값인 위험 목록
    - 알람의 기능에 대한 기술적 설명
    - 경보의 우선순위 지정
    - 경보의 민감도에 대한 요소(민감도, 특이성)
    - 경보 시스템의 알고리즘 및 기능 공개
  - ✓ 테스트 요구사항
    - 사용 지침의 완전성
    - 경보 시스템의 기능
      - * 경보 신호의 가시성
      - * 알람 신호의 가청도(볼륨 레벨, 시간 관리 기능: 톤 시퀀스, 주파수, 일시 중지 등)
      - * 알람 동작의 올바른 기능(알람 끄기, 알람 억제 등)
      - * 특정 상황에서의 기능(EMC, 작동 등)
    - 위험관리 파일(경보 기준, 경보 시스템의 오작동)
    - 작동 요소의 올바른 기능 및 라벨링

- 경보 시스템의 기술적 설명
- 경보 시스템 검증(사용성 파일)

• IEC 60601-1-8 표준의 2020년 주요 수정내용 (일부 참고)

✓ 경보 피로 및 경보 감지 가능성

- 경보 피로: 알려진 경보 피로 문제에 대한 제조업체의 인식을 높이고, 제조업체가 경보 사용자 인터페이스를 설계할 때 효과를 고려하도록 요청
- 경보 피로의 증가 이유: 과도한 경보, 성가시거나 짜증을 유발하는 알람 또는 거짓 긍정 경보
- "임상적으로 실행가능"하다고 할 수 있도록 위험 활동의 효율성이 너무 많거나, 잘못된 경보로 인해 위험 인지성이 감소해서는 안 됨. 이를 위해 표준에서 다음과 같은 용어를 도입
  * 알람 피로
  * 알람 홍수
  * 성가신 경보 신호
  * 임상적으로 실행 불가능
  * 임상적으로 실행가능

✓ 음향 신호의 불량한 경보 감지성

- 효과적이지 않은 음향 신호: 구별하기 어렵고 혼란스럽거나, 중요한 상황에 연결하기 어려움
- 음향 신호의 효과적이고 사용 가능한 음향 인터페이스 설계를 위해 "청각 아이콘", "청각 포인터" 기술이 개발 및 적용됨
  * 청각 아이콘: 음향이 나타내는 범주에 대한 강력한 의미론적 연결을 만드는 소리
  * 청각 포인터: 주의를 끄는 소리는 우선순위를 나타내며 커뮤니케이터의 현지화에 도움이 됨

13-2e 사용자 경험을 평가할 수 있는 다양한 사용자 조사 기법을 활용하였는가?

Yes ☐ No ☐ N/A ☐

- 사용자 경험UX, User eXperience은 한 개인이 특정한 제품, 시스템, 또는 서비스를 사용하며 느끼는 모든 것을 의미한다. 또한, 그 개인이 인지하는 유용성, 사용 편의성, 효율성 등의 시스템 특성을 포함한다. 설명을 평가하기 위해 사용자 조사user research 기법을 활용할 수 있다.
- 사용자 조사 기법은 크게 접근 방식과 자료 획득 방식으로 구분할 수 있다. 우선, 사용자 조사 기법의 접근 방식에 따라 정량적(간접적) 조사quantitative user research와 정성적(직접적) 조사qualitative user research로 구분되며, 사용자 조사를 위해 자료를 얻는 방식에 따라 사용자 행동을 통한 조사와 태도를 통한 조사로 구분한다. 접근 및 자료 획득 방식을 고려해 적합한 사용자 조사 기법을 선정하고, 사용자 경험을 평가하는 것이 바람직하다.
- 접근 방식에 따른 구분 및 방법
  - 정량적(간접적) 조사: 사용자의 행동이나 태도에 대한 데이터를 도구 등을 통해 간접적으로 수집하는 방법
    (예: 웹로그 분석, A/B 테스트, 설문 조사, 고객 지원 자료 분석)
  - 정성적(직접적) 조사: 사용자의 행동이나 태도를 직접 관찰하는 방법
    (예: 인터뷰, 표적 집단 인터뷰focus group interview, 프로토타입 테스팅prototype testing)

✓ 자료 획득 방식에 따른 구분 및 방법

- 사용자 행동 기반 조사behavioral user research: 사용자가 무슨 행동을 하는지를 조사하는 방법
  (예: 웹로그 분석, A/B 테스트, 아이 트래킹eye tracking)
- 사용자 태도 기반 조사attitudinal user research: 사용자가 무엇을 말하는지를 조사하는 방법
  (예: 카드 소팅card sorting, 심층 인터뷰, 요구사항 조사)

참고 사용자 경험에 대한 정성적 평가 시 고려되어야 할 사항의 예시[51]

- 의료 인공지능의 신뢰성을 향상하기 위해 시스템의 임상적 유용성을 정성적으로 조사하여 평가할 수 있다. 의료진 등 실사용자들에게 아래와 같은 내용을 인터뷰하면 인공지능 시스템의 임상적 유용성과 도입 가능성을 높이는 데 이바지할 수 있다.
  - ✓ 인공지능 모델이 제공한 진단 가능 결과가 지나치게 낙관적이거나 비관적이었는가?
  - ✓ 인공지능 시스템이 환자 진료 과정에서 어느 정도로 개입하거나 방해하였는가?
  - ✓ 인공지능 시스템이 환자 진료 과정에서 소통 증진에 도움이 되었는가?
  - ✓ 해당 인공지능 시스템을 어떤 상황 또는 종류의 환자에게 사용할 의향이 있는가?
  - ✓ 인공지능 시스템이 제공한 진단 결과에 대한 이해 수준은 어느 정도였는가?
  - ✓ 인공지능 시스템이 제공한 치료 대안에 대한 이해 수준은 어느 정도였는가?
  - ✓ 인공지능 시스템 및 사용자 인터페이스에 대한 만족도는 어떠하였는가?
  - ✓ 인공지능 시스템 사용 중 발생한 오류 메시지 또는 알림의 유용성은 어느 정도였는가?
  - ✓ 해당 인공지능 시스템을 사용함으로써 생산성이 얼마나 증가하였는가?

05 운영 및 모니터링

책임성 투명성

요구사항 14 인공지능 시스템의 추적가능성 및 변경이력 확보

대표행위자 | 시스템 엔지니어 협력 대상 | 인공지능 모델 개발자 데이터 과학자

- 인공지능 시스템 운영 단계에서 문제 원인 추적을 위한 시스템 로그, 데이터 모니터링, 인공지능 모델과 사람 간의 의사결정 기여도 추적, 변경이력 관리 등의 방안을 확보한다.

14-1 인공지능 시스템의 의사결정에 대한 추적 방안을 수립하였는가?

Yes No N/A
☐ ☐ ☐

해당여부 판단

의료 분야 인공지능 서비스 중 예측 또는 판단 기능에 인공지능이 적용되는 서비스의 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 의료 인공지능 시스템의 출력(추론, 분류 결과 등)을 의료진이 의사결정에 활용할 때 사회·윤리·법적 파급효과를 분석하고, 그에 따라 발생할 수 있는 문제에 대한 추적 방안을 고려하여야 한다. 또한, 운영 중에도 학습이 이루어지도록 설계·개발된 의료 인공지능 시스템이라면 학습 데이터와 모델에 대해 지속적인 모니터링이 필요하다.
- 인공지능 시스템의 경우, 전통적인 소프트웨어와 다르게 생명주기의 프로세스가 반복되는 특성이 있어 서비스 운영 단계에서도 전체 생명주기를 고려한 추적 방안을 확보해야 한다.
- 인공지능 모델의 구축, 데이터셋, 시스템 자체 등 기능적 측면과 인공지능 시스템 운영자 및 사용자 등 인적 요인으로 인해 발생 가능한 인공지능 시스템 출력 결과의 영향을 추적하기 위해서 시스템 단계별로 로그 수집 대상 정보를 정의하고 모니터링을 지속해야 한다.

14-1a 인공지능 시스템의 의사결정에 대한 기여도 추적 방안은 확보하였는가?

Yes No N/A
☐ ☐ ☐

- 현재 인공지능 기반 의료기기는 의료진의 결정 지원의 용도로 활용되고 있다. 따라서 인공지능 시스템의 결정에 대한 의료진의 인지 편향에 대비하여 시스템 결정에 대한 세부화된 기여도 기준을 내부적으로 확립하고, 시스템 운용 과정에서 의사결정의 영향 정도를 추적할 방안(예: 로그 수집)을 확보해야 한다[65].
- 의료 인공지능을 통한 의사결정의 조합은 다음과 같은 조합으로 이루어질 수 있다.

✓ 인공지능 기반 의료기기 시스템의 전적인 의사결정
✓ 인공지능 기반 의료기기 시스템의 의사결정 결과를 의료진이 검토 및 반영 후 의사결정
✓ 주로 의료진이 의사결정을 내리지만, 주요 환부 시각화 등 보조적으로 인공지능 시스템의 출력 활용

- 또한, 인공지능 시스템 내 다수 인공지능 모델의 출력 결과를 조합(예: 앙상블 모델)하여 의사결정에 활용할 때, 모델별 출력이 최종 결과에 미치는 기여도 기준을 세분화하여 사용자(의료진)에게 제공하여야 한다.
- EU의 의료기기 규제는 참고와 같이 요구사항을 제시하고 있으며, '심각한 사고에 관한 정보', '바람직하지 않은 부작용에 대한 데이터', '지속적인 재평가에 사용되어야 하는 적절한 지표 및 임곗값' 등에 의사결정 기여도가 활용될 수 있다.

참고 EU 의료기기 규제 – 시장 출시 후 감시 계획에 따른 정보 수집 및 활용 요구사항(부록 III.)[66]

- 제조자는 제83조에서 86조에 따라 시장 출시 후 감시에 관한 기술 문서를 작성할 때 명확하고 조직적이며 쉽게 검색 가능하고 모호하지 않은 방식으로 제시하여야 하고, 특히 아래의 설명 요소를 포함해야 함
  - ✓ 이용 가능한 정보의 수집 및 활용
    - PSUR(Periodic Safety Update Report) 및 현장 안전 교정의 정보를 포함하여 **심각한 사고에 관한 정보**
    - 심각하지 않은 사건에 대한 기록 및 **바람직하지 않은 부작용에 대한 데이터**
    - 추세 보고의 정보
    - 관련 전문가 또는 기술 문헌, 데이터베이스 및/또는 레지스터
    - 사용자, 유통업체 또는 수입업체가 제공한 피드백 및 불만 사항을 포함한 정보
    - 유사한 의료기기에 대한 공개적으로 이용 가능한 정보
  - ✓ 감시 계획에 포함되어야 하는 내용
    - 위 정보를 수집하기 위한 사전 예방적이고 체계적인 프로세스. 이 프로세스는 기기의 성능에 대한 올바른 특성화를 허용하고 기기와 시장에서 판매되는 유사한 제품을 비교할 수 있도록 해야 함
    - 수집된 데이터를 평가하기 위한 효과적이고 적절한 방법 및 프로세스
    - **지속적인 재평가에 사용되어야 하는 적절한 지표 및 임곗값**(규정 부록 I.의 3절에 언급된 위험 분석 및 위험관리)
    - 불만을 조사하고 현장에서 수집된 시장 관련 경험을 분석하기 위한 효과적이고 적절한 방법과 도구
    - 제88조에 규정된 경향 보고 대상 이벤트를 관리하는 방법 및 프로토콜(사건의 빈도 또는 심각도 및 관찰 기간의 통계적으로 유의한 증가를 설정하는 데 사용되는 방법 및 프로토콜 포함)
    - 권한 있는 당국, 인증 기관, 경제 기관과 효과적으로 의사소통하는 방법 및 프로토콜
    - 제83, 84, 86조에 규정된 제조자 의무를 이행하기 위한 절차에 대한 참조
    - 시정 조치를 포함한 적절한 조치를 식별하고 시작하기 위한 체계적인 절차
    - 시정 조치가 필요할 수 있는 장치를 추적하고 식별하는 효과적인 도구
    - 부록 XIV의 B부에 언급된 PMCF(Post-Market Clinical Follow-up) 계획 또는 PMCF가 적용되지 않은 이유에 대한 정당화

14-1b 인공지능 시스템의 의사결정 추적을 위한 로그 수집 기능을 구현하였는가?

Yes ☐ No ☐ N/A ☐

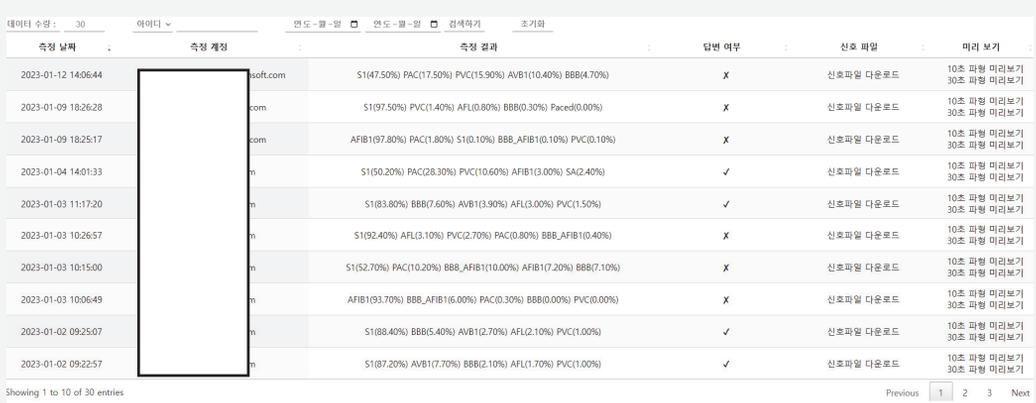
- 인공지능 시스템의 전 생명주기를 고려한 추적가능성 확보를 위해서는 모델의 학습 과정, 운용 시 의사결정 결과, 사용자 입력 데이터 등의 정보에 대한 지속적인 수집이 필요하다. 이를 위해 시스템 프로세스별 로그를 수집할 정보를 선정하고, 정보 간의 중요도를 정의한 뒤 로그 레코드 형식을 결정하여 로그를 수집해야 한다.
- 특히 인공지능 시스템 운영 과정에서의 오류 원인 추적을 위해서는 모델 구축 방법과 데이터셋 측면을 포함한 오류 원인의 분석이 필요하므로, 두 가지 측면을 고려하여 로그를 수집하여야 한다.

인공지능 시스템 운영 과정에서 발생 가능한 오류 원인 예시

오류 구분	오류 원인 예시
모델 구축 방법 측면의 오류	• 모델·데이터의 대상 선정, 수집, 정제, 라벨링 등의 통제 미흡으로 인해 구축 절차, 구조, 학습 모델 측면의 다양한 오류 데이터 생성
데이터셋 측면의 오류	• 데이터셋 설계의 부족, 구문 정확성 위배, 데이터 구축 중복 등으로 인한 학습 데이터 품질 저하

Use case M사의 후선 지원 업무back office 프로그램을 통한 심전도 진단 의사결정 로그 수집 사례

- 심전도를 통한 인공지능 모델의 진단 결과의 추적을 위해 후선 지원 업무 프로그램을 자체적으로 개발 및 운영
- 해당 프로그램에서는 측정 날짜, 사용자 계정, 질환별 인공지능 모델의 진단 확률 및 심전도 측정 결과를 확인할 수 있으며, 전문 의료진의 답변 여부를 기록해 최종 의사결정 주체에 대한 명확한 추적이 가능케 함


데이터 수량 : 30 아이디 연도-월-일 연도-월-일 검색하기 초기화

측정 날짜	측정 계정	측정 결과	답변 여부	신호 파일	미리 보기
2023-01-12 14:06:44	[redacted]soft.com	S1(47.50%) PAC(17.50%) PVC(15.90%) AVB1(10.40%) BBB(4.70%)	✗	신호파일 다운로드	10초 파형 미리보기 30초 파형 미리보기
2023-01-09 18:26:28	[redacted]com	S1(97.50%) PVC(1.40%) AFL(0.80%) BBB(0.30%) Paced(0.00%)	✗	신호파일 다운로드	10초 파형 미리보기 30초 파형 미리보기
2023-01-09 18:25:17	[redacted]com	AFIB1(97.80%) PAC(1.80%) S1(0.10%) BBB_AFIB1(0.10%) PVC(0.10%)	✗	신호파일 다운로드	10초 파형 미리보기 30초 파형 미리보기
2023-01-04 14:01:33	[redacted]m	S1(50.20%) PAC(28.30%) PVC(10.60%) AFIB1(3.00%) SA(2.40%)	✓	신호파일 다운로드	10초 파형 미리보기 30초 파형 미리보기
2023-01-03 11:17:20	[redacted]m	S1(83.80%) BBB(7.60%) AVB1(3.90%) AFL(3.00%) PVC(1.50%)	✓	신호파일 다운로드	10초 파형 미리보기 30초 파형 미리보기
2023-01-03 10:26:57	[redacted]m	S1(92.40%) AFL(3.10%) PVC(2.70%) PAC(0.80%) BBB_AFIB1(0.40%)	✗	신호파일 다운로드	10초 파형 미리보기 30초 파형 미리보기
2023-01-03 10:15:00	[redacted]m	S1(52.70%) PAC(10.20%) BBB_AFIB1(10.00%) AFIB1(7.20%) BBB(7.10%)	✗	신호파일 다운로드	10초 파형 미리보기 30초 파형 미리보기
2023-01-03 10:06:49	[redacted]m	AFIB1(93.70%) BBB_AFIB1(6.00%) PAC(0.30%) BBB(0.00%) PVC(0.00%)	✗	신호파일 다운로드	10초 파형 미리보기 30초 파형 미리보기
2023-01-02 09:25:07	[redacted]m	S1(88.40%) BBB(5.40%) AVB1(2.70%) AFL(2.10%) PVC(1.00%)	✓	신호파일 다운로드	10초 파형 미리보기 30초 파형 미리보기
2023-01-02 09:22:57	[redacted]m	S1(87.20%) AVB1(7.70%) BBB(2.10%) AFL(1.70%) PVC(1.00%)	✓	신호파일 다운로드	10초 파형 미리보기 30초 파형 미리보기

Showing 1 to 10 of 30 entries Previous 1 2 3 Next

후선 지원 업무 프로그램을 통한 의사결정 결과 수집 사례

14-1c 지속적인 사용자 경험 모니터링을 위해 사용자 로그를 수집 및 관리하고 있는가?

Yes ☐ No ☐ N/A ☐

- 서비스 이용 로그 분석은 서비스 운영 상태에 관한 확인뿐만 아니라, 사용자가 겪는 문제가 무엇인지 확인할 수 있는 가장 기본적인 방법이 될 수 있다. 서비스 로그는 서비스가 운영되는 동안 지속해서 수집되며 서비스 고도화에 따라 다양한 형태로 누적될 수 있다.
- 서버 인프라에 대한 로그를 통해 서비스 운영 상태에 대한 모니터링을 수행할 수 있으며, 사용자 상호작용 로그는 사용자가 어떤 서비스를 많이 이용하고 어떤 서비스에서 오류를 겪는지 분석할 수 있다. 이를 위해 인프라 관점에서는 로그 분석 소프트웨어를 활용할 수 있으며, 사용자 관점에서는 기업이 자체적으로 인터페이스 또는 상호작용의 호출에 따른 로그를 수집하거나 로그 분석 도구를 활용할 수 있다.

참고 사용자 경험 로그 수집 및 관리 시 주의해야 할 디지털 의료 데이터 보호 규정 준수 사례[67]

- 의료기기 소프트웨어 개발 시 HIPAA(Health Insurance Portability and Accountability Act, 건강 보험 이동성 및 책임법) 규정 및 이의 준수 필요성
  - ✓ 소프트웨어 회사가 환자의 개인 식별자를 수집하고 처리하는 솔루션과 상호 작용하는 경우, HIPAA 표준이 소프트웨어 제공 업체에 적용됨
  - ✓ 즉, 환자의 식별 가능한 건강 정보를 유지, 공유 또는 단순히 액세스할 수 있는 의료 산업의 모든 소프트웨어 회사는 HIPAA를 준수해야 함
- HIPAA 규정 준수 의료기기 소프트웨어 체크리스트
  - ✓ 어떤 종류의 엔티티가 소프트웨어에서 사용되는가?
  - ✓ 어떤 데이터가 앱에서 사용, 공개 및 저장되는가?
  - ✓ 소프트웨어가 암호화되어 사용되는가?
- 의료기기 소프트웨어 시스템이 HIPAA를 준수하도록 만드는 방법
  - ✓ 개인 정보 보호 규칙
    - 사기 및 도난을 줄이면서 건강 데이터의 흐름을 개선하고 환자에게 데이터 검사, 사본 수신 및 데이터 조정 요청 등 건강 정보에 대한 일부 권리를 부여하기 위한 것
  - ✓ 보안 규칙
    - 해당 조직에서 생성, 수신, 사용 또는 유지 관리하는 ePHI(electronic Protected Health Information)의 보호 기준을 설정, 이 규칙에 따라 "기밀성, 무결성 및 보안을 유지하도록 적절한 관리적·물리적 및 기술적 보호"를 구현해야 함
  - ✓ 규정 준수 규칙
    - 미국의 보건복지부(HHS, United States Department of Health and Human Services)가 HIPAA를 시행하는 방법을 명시하고 있으며, 규제 기관은 과실을 평가하고 규정 미준수에 대해 벌금을 부과함
  - ✓ 신고 규정
    - HIPAA 적용 조직과 비즈니스 파트너는 HIPAA 적용 기관 및 비즈니스 관련자에게 종이 기반 및 전자 PHI를 포함한 보안이 되지 않은 PHI 위반에 대해 알려야 함
  - ✓ 일반 규칙
    - 의료 솔루션의 상호운용성에 관한 규칙을 설정하고 수많은 HIPAA 개인 정보 보호, 보안 및 시행규칙

을 수정하여 위반 보고를 피하고, PHI 사용에 대한 추가 개인 정보 제한을 부과함

- 안전한 솔루션 개발에 도움이 되는 소프트웨어 개발 시 HIPAA 규정 준수 체크리스트
  - ✓ 사용자 권한 부여: 미국 정보는 소프트웨어 애플리케이션의 ID 보증을 4단계로 구성. 가장 기본적인 수준에서는 단일 인증 요소만 사용됨
    - 지식: 합법적인 사용자만 액세스할 수 있는 고유한 데이터 집합을 PIN 또는 암호로 입력해야 함
    - 고유성: 생체 인식 스캔의 사용을 예측하여 복제하거나 수정할 수 없는 사용자의 고유 특성을 확인함
    - 장소: 사용자가 액세스할 때, 특정 위치에 있을 때만 액세스 권한을 부여함
    - 소유: 보안 코드 등 추가 데이터를 사용자에게 제공함. 결과적으로 정보의 법적 보관을 확보하기 위해서 방문자는 이러한 데이터를 입력해야 함
  - ✓ 비상 모드: 응급 상황에서 환자 기록을 안전하게 유지하기 위한 절차, 의무 및 관행을 간략하게 설명함
    - 직업, 연락처 정보 및 의무를 포함한 모든 팀 구성원의 포괄적인 목록
    - 전략 수행을 위한 단계별 접근 방식 수행(어떻게, 언제, 누구에 의해 절차를 수행할 것인지 등)
  - ✓ 수정 계획: 비즈니스 동료가 환자 데이터를 보호하기 위해 수행하는 단계를 지정하는 보안 전략임
    - 동일한 팀 구성원의 의무가 식별되어야 함
    - 미래 문제를 극복하기 위한 행동 계획이 있어야 함
    - 데이터 보안을 유지하기 위해 완료되는 모든 작업의 목록이 필요함
  - ✓ 데이터 백업: HIPAA 규정에 따라 전자적으로 보호되는 모든 건강 정보를 신뢰할 수 있는 다른 데이터 저장 시스템에 복사해야 함
    - 암호화는 더 간단하고 빠른 데이터 보안 기술임. 최상의 데이터 보호를 위해 소프트웨어는 256비트 AES 프로토콜과 이중 인증을 사용해야 함
    - 모니터링: 백업 시스템이 실패하면 시스템은 조직의 구성원에게 즉시 알릴 수 있어야 함
  - ✓ 모니터링: 시스템 개발자와 소유자는 액세스 알고리즘의 효율성과 보안을 자주 테스트해야 함
    - 감사 제어 및 활동 로그: 자동화된 위험 감지 시스템을 사용하여 시스템에 쉽게 진입하려는 의심스러운 시도를 식별함
    - 자동 로그아웃: 모든 의료기기 소프트웨어는 사용자가 교대 근무가 끝난 후 시스템에서 자동으로 로그아웃하도록 개발되어야 함

14-2 학습 데이터의 변경 이력을 확보하고, 데이터 변경이 미치는 영향을 관리하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

인공지능 시스템의 성능 개선을 위해 주기적으로 학습 데이터의 변경을 필요로 하는 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 인공지능 모델은 사용한 데이터에 따라 학습 모델도 함께 달라진다. 이에 따라 모델의 설계나 주요 파라미터들의 변경이 함께 이루어질 수 있다. 따라서 모델의 개발 과정에서 의료용 학습 데이터가 변경될 경우, 학습 데이터 버전관리 및 변경이 발생한 원인을 추적할 수 있어야 한다.
- 또한, 신규 의료 데이터를 포함하여 인공지능 모델의 추가 학습이 필요한 경우, 학습 데이터 변경으로 인한 모델의 성능 영향을 평가하기 위해 기존 학습 데이터에 추가된 신규 데이터 비율에 따른 모델 성능 변화 추적이 가능하도록 기록 및 관리하는 것이 바람직하다.
- 이러한 학습 데이터 변경 이력 관리를 위해 학습 데이터 버전관리를 위한 오픈소스 도구 활용, 자체 시스템 구축 등을 고려할 수 있으며, 학습 데이터를 사용 또는 운용하는 이해관계자가 데이터 변경으로 인한 영향을 확인할 수 있도록 학습 데이터 변경 원인, 변경된 학습 데이터의 구조, 학습 모델의 추론 결과 및 모델 변경으로 인한 성능평가 결과 등에 대한 정보를 제공해야 한다.

14-2a 데이터 흐름 및 계보lineage를 추적하기 위한 조치를 마련하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템의 경우, 데이터의 변경으로 인해 모델의 확장이나 재설계 등의 시스템 변경이 발생할 수 있다. 따라서 시스템의 변경을 유도하는 데이터의 흐름 및 계보를 계속해서 추적해야 한다.
- 데이터 흐름 및 계보는 데이터 변경에 대해 역방향, 순방향, 종단간end-to-end 관점으로 나누어 추적할 수 있으며, 추적을 위한 고려사항은 다음과 같다.
  - ✓ 데이터 흐름 및 계보 추적을 관리하기 위한 데이터 정책팀을 구성하는 것이 유용한가?
  - ✓ 데이터 흐름 및 계보 추적을 위해 메타데이터를 기록하고 유지보수할 것인가?
  - ✓ 데이터 흐름 및 계보 추적을 위한 데이터 적재, 매핑, 관리, 시각화 리포팅 기능을 구현하는 것이 유용한가?
  - ✓ 인공지능 개발과정에서 모델의 특성값을 저장 및 공유하는 특성 저장소feature repository 기능을 구현하는 것이 유용한가?
  - ✓ 데이터는 출처까지 역추적될 수 있는가?

14-2b 데이터 소스 변경에 대한 모니터링 방안을 확보하였는가?

Yes ☐ No ☐ N/A ☐

- 질병의 진단이나 예후 예측 등을 위해 학습 데이터를 실시간 수집하고 인공지능 모델을 실시간으로 학습시키는 등 온라인 학습 방법을 적용하는 시스템이 아닌 경우 본 검증항목은 현재 고려사항이 아닐 수 있다.
- 의료 인공지능 알고리즘 개발을 위해 오픈소스 데이터셋을 활용하는 경우 데이터셋의 변경이나 업데이트가 빈번할 수 있으므로, 모델의 성능 개선을 위해서는 주기적인 모니터링을 통해 최신의 데이터셋을 반영하여야 한다.

14-2c 데이터 변경 시, 버전관리를 수행하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 모델 개발과정에서 학습 데이터의 업데이트, 오류로 인한 라벨링 재수행 등 데이터 변경이 이루어지면 학습 결과인 모델도 변경된다. 또한 이전에 학습에 사용한 데이터셋과 특성이 완전히 다르거나 데이터셋 전체를 교체할 경우 성능이 크게 저하될 수 있으며, 이 경우에는 추가 학습이 필요할 수 있다.
- 따라서 학습 데이터의 변경이 수행될 경우, 단순히 사용된 학습 데이터의 버전뿐만 아니라 해당 버전으로 학습한 인공지능 모델을 함께 관리하여야 한다. 특히, 신규 데이터의 추가로 인한 학습 데이터 변경이 필요한 경우, 학습 혹은 테스트에 사용된 신규 데이터 비율을 기록하고, 그에 따른 모델의 성능 변화가 함께 추적 가능하여야 한다.
- 이를 위해 기계학습 프로젝트를 위한 오픈소스 기반의 데이터 버전관리 도구(예: DVC[Data Version Control][68])의 도입을 고려하거나, 학습 데이터 버전 관리 시스템을 자체적으로 구축하여 학습 데이터의 버전과 모델의 버전관리를 수행해야 한다.

14-2d 데이터 변경 시, 이해관계자를 위한 정보를 제공하는가?

Yes ☐ No ☐ N/A ☐

- 다수의 이해관계자가 참여하는 인공지능 시스템 개발 과정에서 데이터 변경으로 인한 인공지능 모델의 설계, 주요 초매개변수 변경 및 재학습 등의 조치를 이해하려면 이해관계자의 역할을 고려한 정보의 제공이 필요하다.
- 데이터 변경에 따라 이해관계자별로 제공되어야 하는 정보는 다음과 같다.

데이터 변경 시 이해관계자에게 제공해야 할 정보 예시

이해관계자	제공 정보
비즈니스 결정권자	• 데이터 변경에 따른 모델의 세세한 변경점보다 기존 시스템의 목적, 서비스 의도 등의 변경점이나 시스템 전체의 방향성 등에 초점을 맞춘 정보
데이터 과학자	• 기존 데이터와 변경된 데이터의 특징, 포맷, 규모 등의 차이점 등의 정보
시스템 개발자	• 변경된 데이터 설명을 참고하여 기존 모델과의 호환성, 모델 구조 재설계, 모델 재학습 세부 전략 (예: 목적함수, 학습 시간, 학습 알고리즘), 예상 출력 결과 변경점 등에 대한 정보
모델 검증자	• 변경된 테스트 데이터셋 구성, 재설계 및 재학습된 모델에 대한 주요 성능평가 결과, 기존 모델과의 성능 비교 결과 등의 정보
모델 운영자	• 검증을 마친 변경 모델에 대한 운영 및 사용자 모니터링 결과 등을 수집 및 분석한 정보

14-2e 신규 데이터 확보 시, 인공지능 모델의 성능평가를 재수행하였는가?

Yes ☐ No ☐ N/A ☐

- 신규 데이터를 확보한 뒤, 인공지능 시스템에 사용하기 위해서는 기존 운영 중인 인공지능 모델과의 성능 비교가 필요하다. 사람이 판단하기에 신규 데이터가 기존 학습 데이터와 유사하여도 학습된 인공지능 모델이 기존 학습 데이터에서 학습한 데이터 특성과 다를 수 있다.
- 따라서 신규 데이터를 대상으로 의료 분야의 대표적인 인공지능 알고리즘을 사용하여 성능평가를 진행하고 분석하는 과정이 필요하다. 신규 데이터 확보에 따른 성능평가를 위해서는 다음과 같은 과정을 참고한다. 이 과정에서 전문 의료진과의 협업은 필수적이다.
  - ✓ 성능평가 및 비교 분석을 위한 기존 학습 모델 및 관련 대표 인공지능 모델 확보
  - ✓ 의료 인공지능 분야 및 모델에 적절한 성능평가 지표 선정
  - ✓ 성능평가를 위한 실험 설계(정량적·정성적 실험 방법 선정, 실험 모델들의 파라미터 설정, 세부 실험 계획 등)
  - ✓ 실험 진행 및 결과 분석(결과에 따라 신규 데이터 평가 또는 필요 시 모델 재설계, 확장, 재학습 등 결정)

책임성 투명성

요구사항 15 서비스 제공 범위 및 상호작용 대상에 대한 설명 제공

대표행위자 | 시스템 엔지니어 협력 대상 | 시스템 기획자 시스템 운영자 인공지능 모델 개발자 비즈니스 결정권자 전문 의료진

- 사용자가 인공지능 시스템이 제공하는 서비스를 올바르게 사용하고, 제공된 서비스를 오·남용하지 않도록 서비스의 목적, 범위, 제한 사항, 면책 조항disclaimer, 상호작용 대상을 포함한 내용을 설명한다.

15-1 인공지능 서비스의 올바른 사용을 유도하기 위한 설명을 제공하는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

전문 의료진 또는 환자가 활용하는 의료 인공지능 시스템의 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 인공지능의 활용 범위가 넓어지면서 사용자가 서비스 기능에 대한 기대를 실제 서비스 제공 범위보다 더 넓게 오해하는 경우가 발생한다. 따라서 서비스 목적, 범위, 제한사항, 면책조항에 대한 설명을 제공함으로써 인공지능 기술의 오·남용을 방지하고 서비스에 대한 사용자의 기대치를 조정하는 것이 중요하다.
- 특히 환자에게 대면이 아닌 원격으로 진단 결과를 제공하는 서비스는 사용자의 초기 기대치가 잘못 형성될 확률이 높다. 따라서 서비스 제공자는 서비스 제공 목적과 의도가 무엇인지, 서비스의 제공 범위는 어디까지이며 이와 관련된 한계는 무엇인지 설명함으로써 서비스에 대한 사용자의 기대치를 설정해야 한다.
- 서비스 제공자는 인공지능이 제공하는 결과가 사용자에게 미치는 영향을 설명하고, 필요시 해당 결과를 되돌릴 수 있는지 또한 제공하여 해당 서비스의 올바른 사용을 유도해야 한다.

15-1a 서비스의 목적과 목표에 대한 설명을 제공하는가?

Yes ☐ No ☐ N/A ☐

- 서비스 목적goal은 서비스 제공사가 인공지능 시스템을 어떤 목적으로 제공하는지에 대한 방향성을 담은 것이며, 목표objective는 사용자가 해당 기능을 사용함으로써 구체적으로 무엇을 어떻게 얻을 수 있는지를 의미한다. 제공하는 의료 서비스의 목적과 목표를 설명함으로써 사용자는 사용 맥락에 맞는 적합한 기능을 선택하여 활용할 수 있다.

- 의료진은 의료 인공지능 시스템을 참고하여 직접 진단을 내리므로, 사용자가 의료진이면 서비스의 목적과 목표에 대한 설명이 상대적으로 쉽고 명확할 수 있다. 그러나 사용자가 일반인이거나 환자면 서비스의 오·남용 가능성이 있으므로, 사전에 서비스 전반에 대한 자세한 설명을 알기 쉽게 제공하여야 한다. 그리고 향후 개발되어 의료진 대신 의사결정 및 치료 등을 수행할 수 있는 의료 인공지능 시스템의 경우에는 모든 사용자를 대상으로 서비스의 목적과 목표를 설명하여야 한다.
- 인공지능 서비스가 오용 또는 남용될 경우, 인공지능 모델이나 시스템상의 새로운 취약점을 생성하거나 예상치 못한 사회적 이슈를 발생시킬 수 있다. 따라서 서비스가 의도한 목적을 벗어나 잘못 사용되는 것을 방지하기 위해, 이해관계자는 잠재적 오·남용 영역을 식별한 후 사용자가 이를 인식할 수 있도록 관련 사례와 처벌 내용 등을 알려야 한다.

참고 fitbit 불규칙한 리듬 알림 의료기기 서비스 목적 및 기술적 특징[69]

5. Device Description

Intended Use

The Fitbit Irregular Rhythm Notifications feature is a software as a medical device (SaMD). The Fitbit Irregular Rhythm Notifications feature analyzes pulse rate data to identify heart rhythms that are consistent with atrial fibrillation (AFib) and if identified, provide a notification to the user.

Technological Characteristics

The Fitbit Irregular Rhythm Notifications consists of an algorithm that classifies pulse rate data, and a mobile application run within the Fitbit app that serves as the user interface (UI) and device display.

The Fitbit Irregular Rhythm Notifications leverages pulse rate data collected from compatible commercially available, general purpose wrist-worn products (e.g., smartwatch or fitness tracker). Photoplethysmograph (PPG) sensors consist of light-emitting diodes (LED) and photodiodes that detect changes in blood flow of a user's vasculature at any given moment. When the heart beats, it sends a pressure wave through the vasculature causing a blood flow increase. By monitoring the fluctuations the consumer wrist-worn products can measure pulse rate data. When the user is still the sensor detects when individual pulses reach the periphery (i.e., wrist) and measures beat-to-beat intervals.

If the analyzed data are consistent with signs of atrial fibrillation, a notification indicating that a heart rhythm showing signs suggestive of AFib will be displayed to the user. The Fitbit Irregular Rhythm Notifications will only surface a notification of a heart rhythm showing signs of AFib once in a 24-hour period.

The Fitbit Irregular Rhythm Notifications mobile app functions within the Fitbit consumer application and is run on a compatible, user-provided general purpose mobile computing product (e.g., smartphone or tablet). The Fitbit Irregular Rhythm Notifications mobile app serves as the display/user interface for the Fitbit Irregular Rhythm Notifications.

fitbit사의 불규칙한 리듬 탐지 및 알림 서비스의 소프트웨어로서의 의료기기 서비스 목적 및 범위 설명

- 사용 목적
  - ✓ fitbit 불규칙한 리듬 알림 기능은 SaMD이다.
  - ✓ fitbit 불규칙한 리듬 알림 기능은 맥박수 데이터를 분석하여 심방세동AFiB, Atrial fibrillation과 일치하는 심장 리듬을 식별하고, 식별된 경우 사용자에게 알림을 제공한다.

- 기술적 특성
  - ✓ fitbit 불규칙한 리듬 알림은 맥박수 데이터를 분류하는 알고리즘과 사용자 인터페이스 및 기기 디스플레이 역할을 하는 fitbit 앱 내에서 실행되는 모바일 애플리케이션으로 구성된다.
  - ✓ fitbit 불규칙한 리듬 알림은 시중에서 판매되는 호환 가능한 범용 손목 착용 제품(예: 스마트워치 또는 피트니스 트래커)에서 수집한 맥박수 데이터를 활용한다. PPG 센서는 특정 순간에 사용자 맥관 구조의 혈류 변화를 감지하는 발광 다이오드와 포토다이오드로 구성된다.
  - ✓ 심장이 뛰면 맥관 구조를 통해 압력파를 보내 혈류를 증가시킨다. 이러한 변동을 모니터링함으로써 소비자 손목 착용 제품은 맥박수 데이터를 측정할 수 있다. 사용자가 가만히 있을 때 센서는 개별 맥박이 주변(즉, 손목)에 도달하는 시기를 감지하고 비트간 간격을 측정한다.
  - ✓ 분석된 데이터가 심방세동의 징후와 일치하면 AFiB를 암시하는 징후를 나타내는 알림이 사용자에게 표시된다. fitbit 불규칙한 리듬 알림은 24시간에 한 번만 AFiB의 징후를 나타내는 심장 리듬 알림을 표시한다.

15-1b 서비스의 한계와 범위에 대한 설명을 제공하는가?

Yes ☐ No ☐ N/A ☐

- 의료 서비스의 제공 범위와 한계를 설명함으로써 사용자 기대치를 조정할 수 있다. 서비스 결과에 대한 품질은 사용자 그룹 특성, 사용 환경, 사용 데이터 등 다양한 요인에 의해 영향받아 결과가 도출될 수 있으므로 사용자에게 서비스 한계와 제공 범위에 대해 말하는 것이 중요하다.
- 예를 들어, 현재 대다수 의료 인공지능 시스템이 제공하는 진단 및 처방 서비스는 전문 의료진을 완벽히 대체하지는 못한다. 따라서 해당 의료기기 및 소프트웨어의 진단 및 처방 결과는 단순 참고용이며, 최종 판단은 전문 의료진의 결정에 따라야 함을 명시하여야 한다.

참고 fitbit 불규칙한 리듬 알림 의료기기 서비스 한계 및 범위[69]

VII. 510(K) Summary

1. **Submitter Information:**
Fitbit LLC
199 Fremont Street, 14th Floor
San Francisco, CA 94105

Contact Person: Randy Parry
Phone: (415) 985-4778
Email: parran@google.com
Date Prepared: March 8, 2022

2. Subject Device Information
Name of Device: Fitbit Irregular Rhythm Notifications
Common or Usual Name: Irregular Rhythm Analysis Software
Classification Name: Photoplethysmograph Analysis Software For Over-The-Counter Use
Regulatory Class: Class II
Product Code: QDB - 21 CFR 870.2790

3. Predicate Device
Apple Inc., Irregular Rhythm Notification Feature
(DEN180042)

4. Indications for Use

The Fitbit Irregular Rhythm Notifications is a software-only mobile medical application that is intended to be used with compatible consumer wrist-worn products to analyze pulse rate data to identify episodes of irregular heart rhythms suggestive of atrial fibrillation (AFib) and provide a notification to the user.

The Fitbit Irregular Rhythm Notifications is intended for over-the-counter (OTC) use. It is not intended to provide a notification on every episode of irregular rhythm suggestive of AFib and the absence of a notification is not intended to indicate no disease process is present; rather the Fitbit Irregular Rhythm Notifications is intended to opportunistically surface a notification of possible AFib when sufficient data are available for analysis.

These data are only captured when the user is still. Along with the user's risk factors, the Fitbit Irregular Rhythm Notifications can be used to supplement the decision for AFib screening. The Fitbit Irregular Rhythm Notifications is not intended to replace traditional methods of diagnosis or treatment.

The Fitbit Irregular Rhythm Notifications has not been tested for and is not intended for use in people under 22 years of age. It is also not intended for use in individuals previously diagnosed with AFib.

fitbit사의 불규칙한 리듬 탐지 및 알림 서비스의 소프트웨어로서의 의료기기 서비스 사용 및 한계 설명

- 사용 표시Indications for Use
  - ✓ fitbit 불규칙한 리듬 알림은 사용자의 맥박수 데이터를 분석하여 심방세동AFiB을 암시하는 불규칙한 심장 박동 사례를 식별하여 사용자에게 알린다.
  - ✓ fitbit 불규칙한 리듬 알림은 일반의약품OTC, Over the counter용이다. AFiB를 암시하는 불규칙한 박동의 모든 사례에 대해 알림을 제공하기 위한 것이 아니며, 알림이 없다고 하여 질병 진행이 없음을 나타내는 것이 아니다. 오히려 fitbit 불규칙한 리듬 알림은 분석에 사용할 수 있는 충분한 데이터가 있을 때, 가능한 AFiB에 대한 알림을 기회적으로 표면화한다.
  - ✓ 이러한 데이터는 사용자가 가만히 있을 때만 캡처된다. 사용자의 위험 요인과 함께 fitbit 불규칙한 리듬 알림을 사용하여 AFiB 검사 결정을 보완할 수 있다. fitbit 불규칙한 리듬 알림은 기존의 진단 또는 치료 방법을 대체하기 위한 것이 아니다.
  - ✓ fitbit 불규칙한 리듬 알림은 22세 미만의 사용자를 의도하지 않았고, 또한 테스트를 거치지 않았으므로 해당 사용자에게는 사용할 수 없다. 또한 이전에 AFiB 진단을 받은 개인에게 사용하기 위한 용도가 아니다.

15-2 상호작용의 대상을 명확히 설명하는가?

Yes No N/A
☐ ☐ ☐

해당여부 판단

인공지능 서비스가 사용자와 직접적으로 상호작용을 하는 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 최근 인공지능 시스템을 의인화함으로써 사용자의 친밀감을 향상하고 사용성을 높이려는 서비스가 많아지고 있다. 그러나 인공지능 기술이 고도화되며 인간과 구분이 어려워져 사용자는 상호작용의 대상이 사람인지, 시스템인지 혼란을 겪을 수 있다. 따라서 서비스 제공자는 사용자가 상호작용하는 대상을 명확히 알림으로써 사용자가 겪을 혼란을 줄여야 한다.
- 또한, 일반인과 환자는 전문 의료진의 도움 없이 자가 진단을 위해 인공지능 시스템을 사용할 때 여러 혼란을 겪을 수 있다. 의료 인공지능 시스템 사용 시작 시에 본 서비스가 인공지능에 의해 제공됨을 알리고, 본 서비스가 제공하는 진단 결과를 신중하게 수용할 수 있도록 안내하는 것이 바람직하다.

15-2a 사용자가 인공지능과 상호작용하고 있다는 사실을 사용자에게 명확히 설명하였는가?

Yes No N/A
☐ ☐ ☐

1단계: 사용자와 인공지능이 상호작용하는 서비스 범위 명시하기

- 사용자에게 자동화된 응답을 제공하는 등 서비스 내에서 인공지능이 사용자와 상호작용하는 범위를 명시해야 한다.
- 특히, 의인화는 인공지능을 인간과 유사한 상호작용의 대상으로 만들므로, 사용자의 혼선을 예방하고 사용자의 기대치를 조정하기 위해, 사용자에게 상호작용의 대상이 인공지능임을 명시해야 한다.
  - ✓ 의인화된 시스템이 의료진을 위해 환자에 대한 진단 또는 처방 서비스를 보조하는 경우
  - ✓ 의인화된 시스템이 직접 환자 또는 일반인에게 진단 또는 처방 서비스를 제공하는 경우
  - ✓ 의인화된 시스템이 직접 환자 또는 일반인에게 운동 권유 등 헬스케어 서비스를 제공하는 경우

2단계: 서비스 내의 최종 의사결정을 인공지능이 수행하는지 명시하기

- 인공지능이 서비스의 최종 의사결정에 어떻게 직·간접적으로 개입하는지 명시하여야 한다.
- 일례로 인공지능을 활용하는 진단 및 처방 서비스의 경우, 진단이나 처방과 같은 추론 결과가 인공지능에 의해 결정되는지 명시하여야 한다.
- 특히, 의료진이 아닌 환자를 대상으로 하는 의료 인공지능 시스템의 경우에는 아래와 같은 가이드라인 예시를 참고하여 상호작용 방식을 설계할 수 있다.

참고 환자를 위한 의료 인공지능 시스템 설계 시 상호작용 고려사항 예시[70]

분류		고려사항
포함 정보	질환 정보	질환에 대한 일반적인 정보(병명, 증상, 원인 등)
	질환 치료	치료 방법 및 정보
	다음 계획/단계	진단 결과에 따라 취할 수 있는 다음 계획/단계
	시스템 정보	시스템에 대한 일반적인 정보(사용된 데이터, 시스템 인증 여부 등)
	시스템 입력	사용자가 입력한 데이터
	시스템 처리	시스템 알고리즘 또는 진단에 사용된 기술적 절차에 대한 정보
	시스템 출력	시스템이 출력한 데이터(사전 진단, 권고 등)
정보 전달 방식	공감(안심 유도)	환자를 안심시킬 수 있도록 신중하게 선택된 표현을 통한 전달
	명료성, 일반성	다양한 교육 수준을 가진 환자를 위한 쉬운 용어
사용자 상호작용	입력값 확인	데이터 입력값 확인 가능 여부(확인, 취소 등)
	의료진과의 면담	의료진과의 면담 신청 가능 여부
	추가 질문 가능성	추가 질문 가능 여부
	입력값 비교(시각화)	여러 입력값을 비교할 수 있는 시각화 정보 제공 여부
	정보 요청 가능성	추가 세부 정보 요청 가능 여부

PART 3

부록

01 약어표

약어표

API	Application Programming interface
ETSI	European Telecommunications Standards Institute
GEN	Graph Extrapolation Network
LDA	Linear Discriminant Analysis
MLOps	Machine Learning model Operationalization management
LRP	Layerwise Relevance Propagation
NIST	National Institute of Standards and Technology
LSTM	LongShort Term Memory
SimCLR	Simple framework for Contrastive Learning of visual Representations
SVM	Support Vector Machine
ADASYN	Adaptive Synthetic Sampling
AI	Artificial Intelligence
AIMI	Artificial Intelligence in Medicine and Imaging
ALTAI	Assessment List for Trustworthy Artificial Intelligence
AUC	Area Under the Curve
AWS	Amazon Web Services
BDPL	Boundary Differentially Private Layer
CAD	ComputerAided Detection
CAM	Class Activation Map
CDS	Clinical Decision Supporting
CNN	Convolutional Neural Network
CSV	Comma Separated Values
CT	Computed Tomography
CUSUM	Cumulative Sum
CVE	Common Vulnerabilities and Exposures
DICOM	Digital Imaging and Communications in Medicine
DVC	Data Version Control
EC	European Commission
ECG	Electrocardiogram
EMR	Electronic Medical Record
EPRS	European Parliamentary Research Service
ETRI	Electronics and Telecommunications Research Institute
EU	European Union
EWMA	Exponentially Weighted Moving Average

01 약어표

EWS	Early Warning Scores
FDA	Food and Drug Administration
GAN	Generative Adversarial Networks
IBM	International Business Machines Corporation
ICT	Information and Communication Technology
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
iForest	Isolation Forest
IMDRF	International Medical Device Regulators Forum
IoU	Intersection over Union
IP	Intellectual Property
IRB	Institutional Review Board
ISO	International Organization for Standardization
ITK	Insight Toolkit
JPG	Joint Photographic Experts Group
JSON	JavaScript Object Notation
KDI	Korea Development Institute
LIME	Local Interpretable Modelagnostic Explanation
LOF	Local Outlier Factor
LOS	Length Of Stay
MAD	median absolute deviation
mAP	mean Average Precision
MC	Medcouple
MIT	Massachusetts Institute of Technology
ML	Machine Learning
MLMD	Machine Learning-enabled Medical Devices
MONAI	Medical Open Network AI
MP	Meaningful Perturbation
MRI	Magnetic Resonance Imaging
OECD	Organization for Economic Cooperation and Development
OpenCV	Open Source Computer Vision
OSI	Open Source Initiative
PACS	Picture Archiving and Communication System
PMCF	PostMarket Clinical Followup
PNG	Portable Network Graphics
PPV	Positive Predictive Value

01 약어표

PSUR	Periodic Safety Update Report
ROC	Receiver Operating characteristic Curve
ROS	Random Over Sampling
SaMD	Software as Medical Device
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling TEchnique
SOP	Service-Object Pair
TAI	Trustworthy AI
TEE	Trusted Execution Environment
TR	Technical Reports
TRL	Technology Readiness Level
TTA	Telecommunications Technology Association
UCI	University of California Irvine
UNESCO	United Nations Educational, Scientific and Cultural Organization
UW Health	University of Wisconsin Hospitals and Clinics
UX	User eXperience
WAV	Waveform Audio File Format
WEF	World Economic Forum
WHO	World Health Organization
WIT	WhatIf Tool
XAI	eXplainable AI
XML	eXtensible Markup Language

02 용어표

용어표

• 본 용어표에 정의된 용어 외, 인공지능 기술 용어에 대한 정의는 《2023 신뢰할 수 있는 인공지능 개발 안내서 – 일반 분야》를 참고하시기 바랍니다.

용어명	정의
DICOM	의료용 디지털 영상 및 통신 표준은 의료용 기기에서 디지털 영상표현과 통신에 사용되는 여러 가지 표준을 총칭하는 말
거짓 양성	통계상 실제로는 음성인데 검사 결과는 양성이라고 나오는 것 유의어) 1종 오류(Type Ⅰ Error), 위(僞)양성, 거짓 경보(False Alarm)
거짓 음성	통계상 실제로는 양성인데 검사 결과는 음성이라고 나오는 것 유의어) 2종 오류(type Ⅱ error), 위(僞)음성
기계학습 가능 의료기기(MLMD)	의도된 의료목적을 달성하기 위해 부분 혹은 전체에서 기계학습을 사용하는 의료기기
디지털 헬스케어	의료 영역에 정보통신기술(ICT)을 융합해 개인 건강과 질병에 맞춰 필요한 의료 서비스나 건강 관리 서비스를 제공하는 산업 또는 기술
민감도(Sensitivity)	질병이 있는 대상자를 검사했을 때 양성으로 나온 비율
보호변수	AI의 판단에 영향을 주지 않도록 설정한 변수(예: 인종, 성별, 지역)
블랙박스	기능은 알지만, 작동 원리를 이해할 수 없는 복잡한 기계 장치나 시스템 또는 물체
양성 예측도(PPV)	검사 결과 양성인 대상자가 실제로 질병이 있는 비율
유병률(Prevalence)	특정 집단에서 검사 당시 질병을 앓고 있는 대상자의 비율
음성 예측도(NPV)	검사 결과 음성인 대상자가 실제로 질병이 없는 비율을 의미한다.
의료 인공지능	의료 데이터를 학습하고 특정 패턴을 인식해 진단 또는 예측하거나 환자에게 적합한 맞춤 치료 방법을 제공하는 기술 ※ 환자의 상태를 실시간 모니터링하는 등 의료 정보에 대한 접근성 향상에 기여하고 있으며, 의료 인공지능에 의해 표준적 치료 중심에서 진단·예방 등 맞춤형 치료로, 의사의 지식 경험 기반 진료에서 데이터 기반의 정밀도 높은 진료로 의료 패러다임 전환 촉진을 가져오고 있음
의료기기	기구, 장치, 도구, 기계, 기기, 임플란트, 체외진단시약, 소프트웨어 재료 또는 유사한(관련된) 물품으로 사람에게 단독 또는 조합되어 사용되도록 제조자가 의도한 특정 의료목적에 해당하는 제품
의료영상저장 전송시스템(PACS)	디지털 의료 영상 이미지를 DICOM이라는 국제표준규약에 맞게 저장, 가공, 전송하는 시스템
의료영상진단보조(CAD)	초음파 영상 등 의료 영상 데이터를 AI에 학습시켜 정량적으로 분석하여 자동으로 병변을 검출하거나 보조 진단을 수행함으로써, 의료진이 진단을 내릴 때 보조하는 소프트웨어

02 용어표

용어명	정의
의료용 소프트웨어(SaMD)	진단 또는 치료 목적으로 사용되는 독립형 건강 및 의료 관련 소프트웨어
의료윤리	의사 개인의 윤리가 아니라 의사로서 의업을 실천하는 의료 현장에서, 의사의 모든 활동에 적용되는, 자세히 살펴보면 결국 의사 본인을 위한 자율적인 실천 규범
임상생명심사위원회(IRB)	임상 연구에 참여하는 연구대상자의 권리 · 안전 · 복지를 위하여 인간을 대상으로 하는 모든 생명의과학연구의 윤리적, 과학적 측면을 심의하여, 연구계획을 승인할 수 있는 독립된 합의제 의결기구
인공지능 신뢰성 프레임워크	신뢰성 확보를 위해 실무적으로 고려해야 할 3가지(생명 주기별 요구사항 분류, 인공지능 윤리기준 준용, 신뢰성 확보 대상 정의) 설계 요소를 도출하여, 요구사항과 검증항목 마련 설계 요소 모두를 반영할 수 있도록 매트릭스 형태로 체계화한 것
임상결정지원(CDS)	환자로부터 얻어진 임상 정보를 바탕으로 의료인이 질병을 진단하고 치료할 때 의사결정을 도와주는 시스템
임상시험	사람을 직접 대상으로 사람에게서 추출(또는 적출)된 검체나 사람에 대한 정보를 이용하여 이루어지는 모든 시험 또는 연구이자 개발 중인 신약/의료기기 사용 허가 전에 유효성과 안전성을 검증하는 과정 유의어) 임상 연구
임상의(사)	환자 진료와 교육, 연구에 종사하는 임상의학 분야의 전문가
전문의(사)	특정 분야에 대해 추가적인 수련을 받고 전문의 자격을 취득한 의사
전자의무기록(EMR)	병원에 방문한 환자에 대한 진료 기록을 디지털 형태로 체계적으로 수집되어 전자적으로 저장된 환자 의무기록
전향적 연구	연구하고자 하는 요인(위험 요소)을 미리 설정한 후 일정 기간 변화를 추적하는 연구법
최신안전성정보보고(PSUR)	허가 후 정해 진 시점에 품목허가권자가 규제 기관에 제출 목적으로 의약품의 유익성-위해성 균형에 대한 평가 결과를 제공하기 위한 약물감시 문서
특이도(Specificity)	질병이 없는 대상자를 검사했을 때 음성으로 나온 비율
판매 후 임상 사후관리(PMCF)	장치의 CE 마킹 후 그리고 허가된 라벨링에 따라 기기를 사용할 때의 임상 안전성 및 성능에 대한 특정 질문에 대한 답의 제공을 위하여 실시된 조사
피험자 동의	의료기기 임상시험 실시기준 등 관련 규정(의료기기법 시행규칙 제24조 제1항 제4호)에 따라 임상시험을 시작하기 전, 피험자로부터 받아야 하는 동의 ※ 피험자의 동의를 받고 이를 문서화해야 하며, 피험자 동의서 서식, 피험자 설명서 및 그 밖의 문서화 정보는 임상시험심사위원회(IRB)의 승인을 얻어야 함
후향적 연구(임상)	연구대상자와 직접적으로 접촉하지 않으면서 의무기록을 조사하여 특정 데이터를 수집·통계 처리하여 결과를 산출하는 연구 ※ 피험자의 의무기록, 의료 영상, 생체 신호, 병리 검사 등의 데이터, 임상시험 결과 등을 사용할 수 있고, 짧은 시간에 적은 비용으로 특별한 윤리적 고려 없이 쉽게 진행할 수 있으나 데이터의 부정확성·부재, 교란변수, 기타 편향을 고려해야 함

03 참고문헌

참고문헌

- [1] KAIST, 한국4차산업혁명정책센터, "**인공지능(AI)의 의료활용과 주요 이슈,**" 이슈페이퍼, no. 7, 2019. 12.
- [2] 식품의약품안전처, "**의료기기 GMP 국제 품질관리 민원인 안내서,**" 식품의약품안전처, 2017. 12.
- [3] U.S. Food and Drug Administration, "**Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback,**" U.S. Food and Drug Administration, 2019. 4.
- [4] Greenlight grue, **ISO 14971 Risk Management for Medical Devices: The Definitive Guide,** [Online], Available: https://www.greenlight.guru/blog/iso-14971-risk-management#risk-assessment
- [5] 식품의약품안전처, "**인공지능 의료기기의 허가·심사 가이드라인,**" 식품의약품안전처, 2022. 5.
- [6] 식품의약품안전처, "**의료기기의 사이버보안 허가·심사 가이드라인,**" 식품의약품안전처, 2022. 1.
- [7] International Medical Device Regulators Forum, "**Software as a Medical Device: Possible Framework for Risk Categorization and Corresponding Considerations,**" IMDRF/SaMD WG/N12FINAL:2014, 2014. 9.
- [8] 식품의약품안전처, "**인공지능 의료기기 임상시험방법 설계 가이드라인,**" 식품의약품안전처, 2022. 7.
- [9] Frank Liao, Sabrina Adelaine, Majid Afshar, Brian W. Patterson, "**Governance of Clinical AI applications to facilitate safe and equitable deployment in a large health system: Key elements and early successes,**" Frontiers in Digital Health, 2022. 8.
- [10] FDA, **Medtronic Recalls Remote Controllers Used with Paradigm and 508 MiniMed Insulin Pumps for Potential Cybersecurity Risks,** [Online], Available: https://www.fda.gov/medical-devices/medical-device-recalls/medtronic-recalls-remote-controllers-used-paradigm-and-508-minimed-insulin-pumps-potential
- [11] FDA LISTING, **U.S. FDA Medical Devices Registration and FDA Device Listing,** [Online], Available: https://www.fdalisting.com/page/102/us-fda-medical-devices-registration-and-fda-device-listing
- [12] Christopher J. Kelly, corresponding, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, Dominic King, "**Key challenges for delivering clinical impact with artificial intelligence,**" NIH, 2019. 10.
- [13] MIT News, **Artificial intelligence predicts patients' race from their medical images,** [Online], Available: https://news.mit.edu/2022/artificial-intelligence-predicts-patients-race-from-medical-images-0520
- [14] dataDx, **Can You Trust Your Data?,** [Online], Available: https://datadx.com/can-you-trust-your-data/
- [15] Nawaf Alharbe, Mohamed Ali Rakrouki, Abeer Aljohani, "**A Healthcare Quality Assessment Model Based on Outlier Detection Algorithm,**" MDPI processes, vol. 10, no. 6, 2022. 6.
- [16] AKM I. Newaz, N. I. Haque, A. K. Sikder, M. A. Rahman, A. S. Uluagac, "**Adversarial Attacks to Machine Learning-based Smart Healthcare Systems,**" GLOBECOM 2020 - 2020 IEEE Global Communications Conference, 2020. 12.
- [17] Yisroel Mirsky, Tom Mahler, Ilan Shelef, Yuval Elovici, "**CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning,**" 28th USENIX Security Symposium, 2019. 8.

03 참고문헌

- [18] Christoph Baur, et al, "**Generating Highly Realistic Images of Skin Lesions with GANs,**" MICCAI 2018 ISIC Skin Lesion Workshop, 2018.
- [19] Junqiao Chen, et al, "**The Validity of synthetic clinical data:a validation study of a leading synthetic data generator (Synthea) using clinical quality measures,**" BMC Medical Informatics and Decision Making, 2019.
- [20] Gichoya J. W., et. al., "**AI recognition of patient race in medical imaging: a modelling study,**" The Lancet Digital Health, 2022.
- [21] Bavil, I., and Jones, D. S., "**Race Correction and the X-Ray Machine — The Controversy over Increased Radiation Doses for Black Americans in 1968,**" New England Journal of Medicine, vol. 387, no. 10, 2022.
- [22] Advisory Board, **These medical devices produce racially biased results. Here's why that matters,** [Online], Available: https://www.advisory.com/daily-briefing/2022/09/28/medical-devices
- [23] N. Tomašev et al., "**A Clinically Applicable Approach to Continuous Prediction of Future Acute Kidney Injury,**" Nature, vol. 572, pp. 116-119, 2019. 8.
- [24] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, E. Ferrante, "**Gender Imbalance in Medical Imaging Datasets Produces Biased Classifiers for Computer-aided Diagnosis,**" Proceedings of the National Academy of Sciences, vol. 117, no. 23, pp. 12592-12594, 2020. 5.
- [25] M. W. Sjoding et al., "**Racial Bias in Pulse Oximetry Measurement,**" New England Journal of Medicine, vol. 383, no. 25, pp. 2477-2478, 2020. 12.
- [26] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, "**Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations,**" Science, vol. 366, no. 6464, pp. 447-453, 2019. 10.
- [27] Y. Park et al., "**Comparison of Methods to Reduce Bias From Clinical Prediction Models of Postpartum Depression,**" JAMA Network Open, vol. 4, no. 4, pp. 1-11, 2021. 4.
- [28] Campbell, Claudia M.a, Edwards, Robert R.b, Fillingim, Roger B.c. "**Ethnic differences in responses to multiple experimental pain stimuli,**" Pain(National Library of Medicine), vol. 113 no. 1, p 20-26, 2005. 6
- [29] Hoffman KM, Trawalter S, Axt JR, Oliver MN., "**Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites,**" Proc Natl. Acad. Sci. U.S.A. vol. 113, no. 16, 2016. 4.
- [30] NA LIU, XIAOMEI LI, ERSHI QI, MAN XU, LING LI, BO GAO, "**A Novel Ensemble Learning Paradigm for Medical Diagnosis With Imbalanced Data,**" IEEE Access, Vol. 8, 2020. 6.
- [31] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "**SMOTE: Synthetic Minority Over-sampling Technique,**" arXiv:1106.1813v1, 2011. 6.
- [32] H. Han, W. Y. Wang, B. H. Mao, "**Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning,**" ICIC 2005, Part I, LNCS 3644, pp. 878 – 887, 2005.
- [33] H. He, Y. Bai, E. A. Garcia, S. Li, "**ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning,**" 2008 IEEE International Joint Conference on Neural Networks, 2008. 6.

03 참고문헌

- [34] Matloob Khushi, Kamran Shaukat, Talha Mahboob Alam, Ibrahim A. Hameed, Shahadat Uddin, Suhuai Luo, Xiaoyan Yang, Maranatha Consuelo Reyes, "**A Comparative Performance Analysis of Data Resampling M ethods on Imbalance Medical Data,**" IEEE Access, Vol. 9, 2021. 8.
- [35] Casey Haber, Robert Gove, "**A Visualization Tool for Analyzing the Suitability of Software Libraries via Their Code Repositories,**" OSF, pp. 387~187, 2020. 6.
- [36] Open Source Initiative, **The Open Source Definition,** [Online], Available: https://opensource.org/osd
- [37] Cookson, Richard, et al. "**Using cost-effectiveness analysis to address health equity concerns,**" Value in Health, vol. 20, no. 2, pp. 206-212, 2017.
- [38] Canning, Ashley J., et al. "**Parallel functional annotation of cancer-associated missense mutations in hist one methyltransferases,**" Scientific reports, vol. 12, no. 1, 2022. 11.
- [39] Christopher Schmidt, **Approaching Unbalanced Datasets Using Data Augmentation,** [Online], Available: https://medium.com/@cjc.schmidt/approaching-unbalanced-datasets-using-data-augmentation-8b4978e1cf2e
- [40] Suzuki, Etsuji, et al. "**Causal inference in medicine part I-counterfactual models-an approach to clarifying discussions in research and applied public health,**" Nihon eiseigaku zasshi, vol. 64, no. 4, pp. 786-798, 2009. 09.
- [41] Amaral-Silva, H. T., et al. "**Medical image registration using tsallis entropy in statistical parametric mapping (SPM),**" 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, IEEE, 2010.
- [42] Zheng, Huadi, et al. "**BDPL: A boundary differentially private layer against machine learning model extraction attacks,**" Computer Security-ESORICS 2019: 24th European Symposium on Research in Computer Security, pp. 66-83, 2019. 09.
- [43] 김휘영, 정대철, 최병욱, "**딥러닝 기반 의료 영상 인공지능 모델의 취약성: 적대적 공격,**" 대한영상의학회지, vol. 80, no. 2, pp. 259-273, 2019.
- [44] Finlayson, Samuel G., et al. "**Adversarial attacks on medical machine learning,**" Science, vol. 363, no. 6433, pp. 1287-1289, 2019. 03.
- [45] Dongyu Meng, Hao Chen, "**MagNet: a Two-Pronged Defense against Adversarial Examples,**" arXiv:1705.09064, 2017. 9.
- [46] S. Dey et al., "**Human-centered Explainability for Life Sciences, Healthcare, and Medical Informatics,**" Patterns, vol. 3, no. 5, pp. 1-16, 2022. 5.
- [47] S. M. Lauritsen et al., "**Explainable Artificial Intelligence Model to Predict Acute Critical Illness from Electronic Health Records,**" Nature Communications, vol. 11, no. 3852, 2020. 7.
- [48] A. Raza et al., "**Designing ECG Monitoring Healthcare System with Federated Transfer Learning and Explainable AI,**" Knowledge-Based Systems, vol. 236, issue C, pp. 1-19, 2022. 1.

03 참고문헌

- [49] 최병욱, **의료 AI의 현재와 미래,** [Online], 한국보건의료연구원 보건의료이슈, vol. 60, Available: https://hineca.kr/1868.
- [50] SCIENCE|BUSINESS, **AI is not a bad doctor – if you trust it,** [Online], Available: https://sciencebusiness.net/news/ai-not-bad-doctor-if-you-trust-it
- [51] European Parliament, "**Artificial Intelligence in Healthcare: Applications, Risks, and Ethical and Societal Impacts,**" European Parliamentary Research Service, 2022. 6.
- [52] World Health Organization, "**Ethics and Governance of Artificial Intelligence for Health: WHO Guidance,**" WHO, 2021. 6.
- [53] Sendak, Mark, et. al. "**"The human body is a black box" supporting clinical decision-making with deep learning,**" Proceeding of the 2020 conference on fairness, accountability, and transparency, 2020. 1.
- [54] Kompa, Benjamin, et. al. "**Second opinion needed: communicating uncertainty in medical machine learning,**" NPJ Digital Medicine, vol. 4, no. 1, pp. 1-6, 2021. 5.
- [55] 한국인터넷진흥원, **의료기기 보안시험 해설서,** 2022. 7.
- [56] 최상태, "**왓슨을 중심으로 본 의료 인공지능의 유용성 및 검증의 필요성,**" 미래의료인문사회과학, vol. 1, no. 2, pp. 35-56, 2018. 12.
- [57] Bruckert, Sebastian, et. al. "**The next generation of medical decision support: A roadmap toward transparent expert companions,**" Frontiers in artificial intelligence, 3, 507973.
- [58] P. Solanki, J. Grundy, W. Hussain, "**Operationalising Ethics in Artificial Intelligence for Healthcare: A Framework for AI Developers,**" AI and Ethics, 2022. 7.
- [59] S. H. Park, K. Han, "**Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction,**" Radiology, vol. 286, no. 3, pp. 800-809, 2018. 3.
- [60] D. B. Larson, H. Harvey, D. L. Rubin, N. Irani, J. R. Tse, C. P. Langlotz, "**Regulatory Frameworks for Development and Evaluation of Artificial Intelligence-Based Diagnostic Imaging Algorithms: Summary and Recommendations,**" Journal of the American College of Radiology, vol. 18, no. 3, pp. 413-424, 2021. 3.
- [61] FDA, "**Artificial Intelligence(AI) and Machine Learning (ML) in Medical Devices,**" 2020. 10.
- [62] M. Demirer et al., "**A User Interface for Optimizing Radiologist Engagement in Image Data Curation for Artificial Intelligence,**" Radiology: Artificial Intelligence, vol. 1, no. 6, pp. 1-7, 2019. 11.
- [63] 식품의약품안전처, "**의료기기의 전기·기계적 안전에 관한 공통기준규격 실무안내서 – 사용적합성,**" 2017. 12.
- [64] Johner Institute, **IEC 60601-1-8: 12 steps back to conformity,** [Online], Available: https://www.johner-institute.com/articles/product-development/and-more/iec-60601-1-8/
- [65] S. Ameen, M. C. Wong, K. C. Yee, P. Turner, "**AI and Clinical Decision Making: The Limitations and Risks of Computational Reductionism in Bowel Cancer Screening,**" Applied Sciences, 2022. 3.
- [66] The European Parliament and the Council of the European Union, "**REGULATION (EU) 2017/745 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on Medical Devices,**" Official Journal of the European Union, 2017. 4.

03 참고문헌

- [67] HHS, **Health Information Privacy,** [Online], Available: https://www.hhs.gov/hipaa/index.html
- [68] DVC, **Open-source Version Control System for Machine Learning Projects,** [Online], Available: https://dvc.org/
- [69] U.S. FOOD & DRUG, **Devices@FDA,** [Online], Available: https://www.accessdata.fda.gov/scripts/cdrh/devicesatfda/index.cfm
- [70] R. Larasati, A. De Liddo, E. Motta, "**AI Healthcare System Interface: Explanation Design for Non-Expert User Trust,**" In: ACMIUI-WS 2021: Joint Proceedings of the ACM IUI 2021 Workshops (Glowacka, Dorota and Krishnamurthy, Vinayak eds.), CEUR Workshop Proceedings, 2903, 2021. 4.

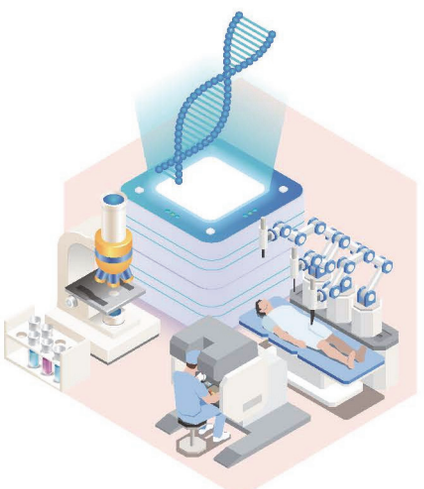
▌한국정보통신기술협회

이강해 단장　곽준호 팀장
조경우 책임　채희문 책임
황재영 책임　변은영 선임
신예진 선임　박경은 전임
오상훈 전임　강상연 연구원

2023
신뢰할 수 있는 인공지능 개발 안내서 의료 분야

초판 인쇄 2023년 06월 26일
초판 발행 2023년 07월 06일
저　　자 한국정보통신기술협회
발 행 인 최영해 · 김갑용
발 행 처 진한엠앤비
주　　소 서울시 서대문구 독립문로 14길 66 205호(냉천동 260)
전　　화 02) 364-8491(代) / 팩스 02) 319-3537
홈페이지 http://www.jinhanbook.co.kr
편집·제작 ㈜디자인여백플러스
등록번호 제25100-2016-000019호 (등록일자: 1993년 05월 25일)

ISBN 979-11-290-4926-1 (93550)　[정가 18,000원]



2023

신뢰할 수 있는 인공지능 개발 안내서 의료 분야

Jinhan M&B

정가 18,000원

93550

9 791129 049261

ISBN 979-11-290-4926-1

과학기술정보통신부
Ministry of Science and ICT

TTA 한국정보통신기술협회
Telecommunications Technology Association