Review Article

J Korean Soc Radiol 2024;85(5):834-847
https://doi.org/10.3348/jksr.2024.0118
eISSN 2951-0805

JOURNAL of THE KOREAN SOCIETY of RADIOLOGY

Explainable & Safe Artificial Intelligence in Radiology

의료 영상 분석을 위한 설명 가능하고 안전한 인공지능

Synho Do, PhD[1,2,3]*

[1]Laboratory of Medical Imaging and Computation, Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
[2]KU-KIST Graduate School of Converging Science and Technology at Korea University, Seoul, Korea
[3]Kempner Institute, Harvard University, Boston, MA, USA

Received September 8, 2024
Revised September 21, 2024
Accepted September 24, 2024

***Corresponding author**
Synho Do, PhD
Department of Radiology, Massachusetts General Hospital and Harvard Medical School, 125 Nashua Street, Suite 2210, Boston, MA 02114, USA.

Tel 1-339-222-4409
E-mail sdo@mgh.harvard.edu

Artificial intelligence (AI) is transforming radiology with improved diagnostic accuracy and efficiency, but prediction uncertainty remains a critical challenge. This review examines key sources of uncertainty—out-of-distribution, aleatoric, and model uncertainties—and highlights the importance of independent confidence metrics and explainable AI for safe integration. Independent confidence metrics assess the reliability of AI predictions, while explainable AI provides transparency, enhancing collaboration between AI and radiologists. The development of zero-error tolerance models, designed to minimize errors, sets new standards for safety. Addressing these challenges will enable AI to become a trusted partner in radiology, advancing care standards and patient outcomes.

Index terms Explainable AI; Safe AI; Zero-Error Tolerance Model

서론

인공지능(artificial intelligence; 이하 AI)은 의료 분야에 혁신을 불러일으키고 있으며, 영상의학은 이 변화의 최전선에 있다. 머신러닝(machine learning) 모델은 의료 영상의 진단, 해석, 관리 방식을 재정의하며, 이전에는 불가능했던 정확도와 효율성, 그리고 환자 결과의 개선을 제공한다. 그러나 이러한 AI 시스템의 뛰어난 성능 뒤에는 간과할 수 없는 도전 과제, 즉 예측 불확실성이 존재한다. 환자의 생명과 직결되는 영상의학 같은 고위험 의료 분야에서는 이 불확실성을 이해하고 관리하는 것이 기술적인 필요를 넘어 임상적으로 필수적인 과제이다.

본 리뷰는 영상의학에서 예측 불확실성의 핵심을 탐구하고 그 원인을 밝히며, 이를 해결하기 위한 독립적 신뢰성 지표의 필요성과 설명 가능한 AI의 중요성을 강조한다(1). 이를 통해 영상의학 전문의가 신뢰할 수 있는 AI 시스템을 더 안전하게 통합할 수 있는 방향을 제시하며, 진단 과정에서 잠재적인 위험을 감소시키고 AI를 신뢰할 수 있는 동반자

로 발전시키기 위한 포괄적이고 실용적인 지침을 제공한다.

영상의학 AI 모델에서 예측 불확실성의 원인

Fig. 1에서 요약된 바와 같이, AI 모델의 예측 불확실성은 임상 환경에서 AI 결과의 신뢰성을 저해할 수 있는 중요한 도전 과제이다. 이러한 불확실성은 여러 원인에서 발생하며, 각각은 예측 실패 가능성에 기여할 수 있다. 예측 불확실성의 주요 원인으로는 분포 외(out-of-distribution; 이하 OOD) 불확실성, 데이터 내재적 불확실성(aleatoric uncertainty), 그리고 모델 불확실성(model uncertainty)이 있다(1).

분포 외 불확실성(OOD Uncertainty)

분포 외(OOD) 불확실성, 또는 인식론적 불확실성(epistemic uncertainty)은 AI 모델이 훈련 세트와 크게 벗어난 데이터를 만날 때 발생한다. 이는 모델이 충분한 지식을 갖추지 못한 상황을 반영하며, 주로 훈련 데이터셋의 한계에서 비롯된다. 예를 들어, 다양한 인구 집단의 부족한 대표성, 서로 다른 영상 기법, 또는 시간이 지나며 변화하는 질병 패턴 등이 그 원인일 수 있다. 영상의학에서는 특정 장비나 인구 집단, 혹은 임상 프로토콜에 맞춰 훈련된 모델이 새로운 환경에서 사용될 때, OOD 불확실성이 발생할 수 있으며, 이는 예측 성능의 불안정성을 초래할 수 있다(1-4).

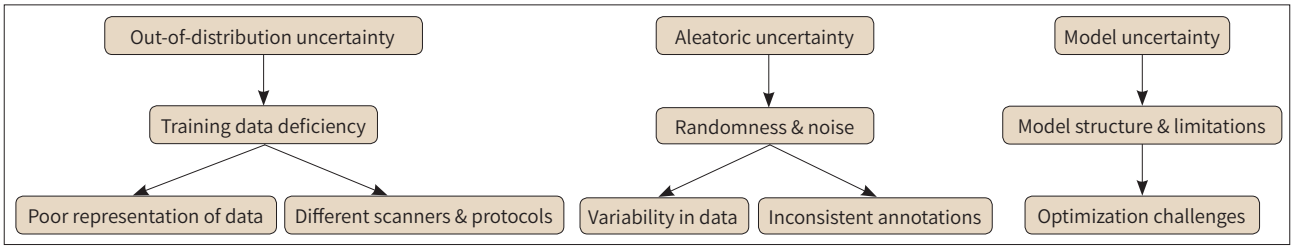
데이터 내재적 불확실성(Aleatoric Uncertainty)

데이터 내재적 불확실성은 데이터 자체에 내재된 무작위성, 노이즈 또는 불일치로부터 발생한다. 이는 낮은 이미지 품질, 인공물(artifact), 그리고 사람의 주석 변동성 등에 의해 발생할 수 있다. 데이터 내재적 불확실성은 데이터의 확률적 특성에서 비롯되므로, 고품질의 모델을 사용하더라도 종종 피할 수 없는 문제이다(5-7).

모델 불확실성(Model Uncertainty)

모델 불확실성은 모델 구조, 파라미터 선택, 그리고 모델의 고유한 한계에서 발생하는 불확실성이다. 이는 모델이 데이터를 얼마나 잘 일반화하고 해석할 수 있는지와 관련이 있으며, 주로 훈련 데이터의 양과 질, 선택된 모델 구조의 적합성에 영향을 받는다. 복잡한 신경망은 고품질의 충분한 데이터를 사용할 때 우수한 성능을 발휘할 수 있지만, 제한된 데이터나 노이즈가 많은 경우에

Fig. 1. Sources of prediction uncertainty in artificial intelligence models.



는 더 간단한 구조가 선호될 수 있다.

안전한 AI 통합을 위한 예측 불확실성 관리

AI 시스템이 의료, 특히 영상의학에서 확산되면서 운영 효율성 증대, 진단 시간 단축, 진단 정확도 개선의 기회가 열리고 있다. 그러나 이러한 시스템의 예측 불확실성을 관리하는 것은 환자에게 잠재적 위험을 미연에 방지하기 위해 필수적이다. 이를 해결하기 위해 예측 불확실성 지표 도입, 데이터 품질 개선, 작업별 안전 기준 도입 등의 전략이 요구된다(1).

영상의학에서의 분포 외(OOD) 불확실성

영상의학에서 OOD 불확실성은 특히 문제가 될 수 있다. 이는 환자의 인구통계적 특성, 영상 장비, 또는 임상 프로토콜의 작은 차이만으로도 데이터 분포에 큰 변화가 생길 수 있기 때문이다. AI 모델이 이와 같은 새로운 데이터를 접하게 되면 성능이 급격히 떨어져 진단 정확도가 손상될 수 있다.

분포 외(OOD) 불확실성의 구체적인 예시

영상 장비의 다양성

영상의학에서는 대형 병원의 최신 MRI와 CT 장비부터 시골 작은 병원의 오래된 휴대용 X-ray 기기에 이르기까지 다양한 장비가 사용된다. 고해상도 이미지를 기반으로 훈련된 AI 모델은 구식 장비나 정교하지 않은 기기의 데이터를 접하게 되면 OOD 불확실성에 직면하게 된다. 이러한 훈련 데이터와 실제 환경 간의 불일치는 예측 성능의 불안정을 야기하고, 기흉이나 흉막삼출증 같은 질환의 잘못된 분류로 이어질 수 있다.

기흉과 흉막삼출증은 특히 이미지 품질에 매우 민감한 미세한 방사선학적 단서를 바탕으로 진단되므로 중요한 사례들이다. 예를 들어, 오래된 X-ray 기기는 낮은 대비, 높은 노이즈, 그리고 불규칙한 해상도를 가진 이미지를 생성하여 기흉의 폐 가장자리나 흉막삼출증의 미세한 액체층 같은 중요한 소견을 가릴 수 있다. 고품질의 이미지로 훈련된 AI 모델이 이러한 낮은 품질의 분포 외 데이터를 접할 때, 미세한 이상을 인식하지 못해 진단이 누락되거나 잘못될 수 있다.

또한, 영상 장비 제조사 간의 하드웨어와 소프트웨어, 영상 프로토콜의 차이는 훈련 데이터와 실제 환경 간의 불일치를 심화시켜 OOD 불확실성을 더욱 악화시킨다. 이러한 불일치들은 AI 모델이 다양한 영상 조건에서 강하고 유연하게 작동할 수 있도록 훈련 데이터의 다양성을 높이고, 도메인 적응 기법을 도입하며, 독립적인 신뢰성 지표를 개선하는 등의 완화 전략이 필수적임을 보여준다.

인구 집단의 차이

대부분 성인 인구를 대상으로 훈련된 AI 모델은 소아나 노인 환자에게 적용될 때 OOD 불확실성에 노출된다. 이들 환자 그룹은 해부학적, 생리학적으로 성인과 크게 달라 모델 성능에 영향을

미칠 수 있다. 예를 들어, 소아 흉부 X-ray는 성인 이미지에서는 볼 수 없는 흉선 그림자, 덜 골화된 뼈, 그리고 독특한 폐 해부학적 특징을 자주 보인다.

이러한 해부학적 변이는 소아의 독특한 소견을 모르는 AI 모델이 흉선 그림자를 병리적 종양으로 오인해 잘못된 진단을 내리거나 불필요한 추가 검사와 치료로 이어질 수 있다. 이와 같은 오진은 OOD 불확실성의 본질적인 문제를 보여준다. 즉, AI 모델은 훈련된 환경에서 벗어난 데이터를 제대로 일반화하지 못할 수 있다는 것이다.

소아 및 노인 환자 데이터는 대부분 성인 데이터를 중심으로 한 훈련 데이터셋에서 부족한 경우가 많다. 이러한 인구 집단의 특이성은 정확한 진단에 매우 중요한 요소로, 다양한 환자 그룹에서 AI 모델의 성능을 높이려면 훈련 데이터셋 확장, 전이 학습 도입, 그리고 지속적인 모델 업데이트 같은 전략이 필요하다.

지역 및 인종적 변이

AI 모델은 지역 및 인종적 차이에 따라 질병의 방사선학적 소견이 달라질 수 있어 OOD 불확실성에 자주 직면한다. 유전적, 환경적, 의료 접근성 등의 요인으로 인해, 훈련 데이터와 다른 집단의 데이터를 접할 때 모델 성능이 불안정해질 수 있다. 예를 들어, 흉부 X-ray에서 결핵의 방사선 소견은 아시아와 서양 인구 간에 폐 병변의 분포, 중증도 등에서 유의미한 차이를 보일 수 있다.

주로 서양 인구 데이터를 기반으로 훈련된 AI 모델은 이러한 지역적 변이를 충분히 인지하지 못해 아시아 환자에서 결핵과 같은 질환을 과소 진단하거나 오진할 수 있다. 이는 OOD 불확실성의 핵심 과제를 보여준다. 즉, 제한적이거나 동질적인 데이터셋에서 훈련된 모델은 실제 임상 데이터의 다양성을 일반화하기 어려워질 수 있다는 것이다.

이를 해결하기 위해서는 다양한 지역과 인종적 배경을 반영한 훈련 데이터셋을 구축하고, 질병의 다양한 표현을 통합한 모델 업데이트가 필요하다. 이렇게 함으로써 AI 시스템은 일반화 능력을 개선하고 진단 오류를 줄이며, 다양한 인구 집단에 걸쳐 공평한 의료 서비스를 제공할 수 있다.

질병 패턴의 변화

COVID-19 팬데믹과 같이 질병 패턴이 급격하게 변화할 때 AI 모델의 OOD 불확실성은 크게 증가한다. 팬데믹 이전 데이터를 기반으로 훈련된 모델은 COVID-19 관련 폐렴이나 급성 호흡 곤란 증후군(acute respiratory distress syndrome; ARDS) 같은 새로운 질환을 제대로 분류하지 못했다. 이는 AI 모델의 훈련된 지식과 새로운 임상 환경 간의 불일치가 AI 성능에 미치는 영향을 보여준다(8).

COVID-19 팬데믹은 기존의 훈련 데이터가 구식이 되거나 불충분해질 수 있는 상황을 명확히 보여주었다. 예를 들어, COVID-19 폐렴의 특유한 방사선 소견인 유리음영과 말초 폐 침범은 기존 폐렴 데이터를 기반으로 훈련된 모델에는 없었기 때문에, 이러한 모델들은 종종 COVID 관련 소견을 오 분류하거나 놓쳐 중요한 시기에 진단 오류를 일으켰다.

이러한 사례는 AI 모델이 변화하는 임상 환경에 적응할 수 있도록 지속적으로 업데이트되고 재훈련되어야 할 필요성을 강조한다. 정기적인 업데이트 없이 모델은 빠르게 구식이 되어 OOD 불확실성에 취약해질 수 있으며, 이에 따라 임상적 유용성도 감소할 수 있다. 변화하는 질병 패턴을 실

시간으로 반영하고 신속한 재훈련이 가능한 강력한 메커니즘을 구축하는 것이 AI 모델의 신뢰성과 적응성을 유지하는 데 필수적이다.

영상의학에서의 데이터 내재적 불확실성

데이터 내재적 불확실성은 데이터 자체에 내재되어 있으며, 낮은 이미지 품질, 다양한 영상 획득 파라미터, 그리고 사람의 주석 간 불일치 등으로 발생한다. 이러한 불확실성은 영상 획득 과정에서 피할 수 없는 요소로 작용하며, 고도로 훈련된 AI 시스템에서도 예측 결과의 변동성을 유발할 수 있다. 인공물, 낮은 해상도, 또는 영상의학 전문의의 주관적인 레이블링 차이에서 발생하는 데이터 노이즈는 영상의학에서 일관되고 신뢰할 수 있는 AI 성능을 달성하는 데 지속적인 도전 과제를 제공한다.

데이터 내재적 불확실성의 구체적인 예시

이미지 인공물(Image Artifact)

이미지 인공물는 영상의학에서 흔히 발생하는 데이터 내재적 불확실성의 원인 중 하나로, 움직임으로 인한 흐림(motion blurring), 빔 하드닝(beam hardening), 금속 선상 인공물(metal streaks)과 같은 요소에서 발생한다. 이러한 인공물은 영상 데이터에 노이즈를 추가하여 중요한 소견을 가리고 해석을 복잡하게 만든다. 예를 들어, 복부 CT 스캔에서 금속 임플란트로 인해 발생하는 선형 인공물은 작은 종양을 가려 AI 모델이 이러한 미세한 이상을 정확히 감지하기 어렵게 만든다. 영상의학 전문의는 이러한 인공물을 고려해 평가를 조정할 수 있지만, AI 모델은 이와 같은 결함이 포함된 사례에 대한 특정 훈련이 부족하면 이를 처리하는 데 어려움을 겪는다. 이러한 인공물로 인해 발생하는 내재적 불확실성은 AI가 이미지 변동성을 일관되게 처리하지 못해 불확실하거나 잘못된 예측을 하게 될 수 있음을 보여주는 중요한 한계이다.

영상 획득 시 발생하는 노이즈

저선량 CT 스캔과 같은 저선량 영상 기법은 방사선 노출을 줄이기 위해 설계되었지만, 그 결과로 노이즈가 더 많은 이미지를 생성하게 된다. 이 높은 노이즈는 AI가 해부학적 구조를 명확하게 구분하는 데 혼란을 일으켜 데이터 내재적 불확실성을 유발한다. 특히 작은 폐 결절과 같은 미세한 소견을 다룰 때 이러한 문제가 두드러진다. 예를 들어, 폐암 스크리닝에서는 높은 노이즈가 중요한 세부 사항을 가려 양성과 악성 병변을 구분하는 데 어려움을 줄 수 있다. 이러한 추가적인 노이즈는 AI 모델이 저품질 이미지에서 신뢰할 수 있는 평가를 내리기 어렵게 하며, 결국 진단 정확도에 영향을 미친다.

불일치하는 주석(레이블링)

불일치하는 주석은 영상의학에서 데이터 내재적 불확실성을 야기하는 중요한 요인이다. 이는 각기 다른 영상의학 전문의가 의료 영상을 해석하고 레이블을 지정하는 방식의 자연스러운 변동

성에서 기인한다. 이러한 변동성은 AI 모델 훈련 시 주요 도전 과제로 작용하며, 일관성 없는 정답 레이블 또는 때로는 상충하는 레이블을 만들어낸다. 예를 들어, 흉부 X-ray에서 경계성 폐렴을 진단할 때 영상의학 전문의들 간의 임상적 판단과 경험, 미세한 방사선 소견의 해석 차이로 인해 진단이 달라질 수 있다. 일부 전문의는 경미한 혼탁을 폐렴으로 레이블 할 수 있지만, 다른 전문의는 이를 비특이적 소견이나 정상 변이로 간주할 수 있다.

이러한 레이블의 불일치는 AI 모델의 훈련 과정에서 레이블 노이즈로 작용하며, AI는 일관되고 표준화된 정답 대신 다양한 주관적 해석을 반영한 데이터로 학습하게 된다(9). 그 결과, 특히 훈련 데이터가 상충하는 신호를 제공하는 경우, 모델의 예측이 불확실해질 수 있다. 이는 임상 환경에서 일관된 예측을 제공하는 AI 모델의 능력에 직접적으로 영향을 미친다. AI 모델이 인간 전문가들 사이의 변동성을 반영할 수밖에 없기 때문이다.

이 문제는 실제 임상 환경에서 더욱 심화된다. 훈련 데이터의 일관성이 부족하면 모델의 성능이 변동하고, 진단의 불확실성도 커질 수 있다. 예를 들어, 불일치하는 주석으로 훈련된 모델은 미세하거나 경계성 상태를 신뢰성 있게 진단하기 어려워, 진단 누락이나 잘못된 진단으로 이어질 수 있다. 이러한 문제는 모델 개발 시 표준화된 합의 기반 레이블링 프로토콜을 적용하여 변동성을 최소화하는 것이 중요함을 보여준다. 주석의 일관성을 개선함으로써 레이블 불일치로 인한 데이터 내재적 불확실성을 크게 줄이고, 임상 환경에서 AI 모델의 신뢰성과 성능을 높일 수 있다(10-12).

영상의학에서의 모델 불확실성

모델 불확실성은 AI의 내부 구조와 훈련 데이터를 실제 임상 상황에 일반화하는 능력과 관련이 있다. 이는 모델의 아키텍처, 훈련 방법론, 그리고 시스템에 내재된 해석 가능성의 수준에 의해 영향을 받는다.

모델 불확실성의 구체적인 예시

모델 아키텍처 선택

모델 아키텍처의 선택은 AI 모델의 성능에 중요한 역할을 하지만, 특정 영상의학 작업에 가장 적합한 아키텍처를 알 수 없을 때 모델 불확실성의 주요 원인이 될 수 있다. 종종 실무자들은 쉽게 구현할 수 있는 잘 알려진 오픈 소스 모델에 의존하게 된다. 이러한 접근은 빠른 배포가 가능하다는 장점이 있지만, 선택된 아키텍처가 특정 임상 문제에 최적화되지 않았을 경우 불확실성이 발생할 수 있다.

예를 들어, 합성곱 신경망(convolutional neural network)은 구현이 용이하고 상대적으로 낮은 계산 자원으로 정상 및 비정상 흉부 X-ray를 분류하는 데 자주 사용된다. 그러나 이러한 모델은 폐 결절의 특정 아형을 식별하거나 중첩된 병변을 구분하는 더 복잡한 작업에서는 한계를 가질 수 있다.

트랜스포머, 앙상블 모델, 또는 최신 신경망과 같은 더 복잡한 아키텍처는 데이터를 더 깊이 처

리하고 풍부한 통찰을 제공할 수 있지만, 이들 모델은 더 높은 계산 자원 요구와 임상 워크플로에서의 통합 및 상호 운용성 문제와 같은 트레이드오프를 동반한다. 최적의 모델 아키텍처 선택 과정은 종종 시행착오와 벤치마킹, 그리고 적응에 기반하기 때문에 간단하지 않다.

결국 범용 모델에 의존하는 것은 특정 임상 문제의 세부 사항과 완벽하게 일치하지 않을 수 있어 모델 불확실성을 초래할 수 있다. 개발된 모델이 새로운 실제 데이터를 접하게 될 때, 성능이 예기치 않게 달라질 수 있다. 따라서 모델 아키텍처를 신중하게 평가하고, 적응시키며, 지속적으로 검토하는 것이 모델 불확실성을 완화하고 영상의학 AI 시스템의 신뢰성을 높이는 데 필수적이다.

훈련 데이터에 대한 과적합

과적합은 AI 모델이 훈련 데이터에 포함된 노이즈, 이상치, 특정 패턴에 지나치게 적응하여 발생하는 모델 불확실성의 흔한 원인이다. 이로 인해 모델이 새로운 데이터를 제대로 일반화하지 못하게 된다. 이 문제는 모델의 목표 함수가 지역 최소값이 아닌 전역 최소값으로 수렴할 것이라는 보장이 없기 때문에 발생한다.

예를 들어, 골절을 감지하도록 훈련된 모델은 유사한 데이터에서는 좋은 성능을 보일 수 있지만, 실제 환자에서 다양한 해부학적 차이와 같은 변동성에 직면했을 때 성능을 유지하지 못할 수 있다. 이는 모델이 훈련 데이터의 특성을 너무 깊이 학습한 나머지 실제 임상 환경의 변동성을 처리하지 못하는 문제를 반영한다.

과적합 문제를 해결하기 위해서는 다양한 시나리오를 포괄하는 훈련 데이터가 필요하며, 다양한 데이터 하위 집합에서 모델 성능을 검증하는 강력한 검증 방법이 요구된다. 또한, 과적합으로 인한 모델 불확실성을 줄이기 위해 조기 종료, 교차 검증, 정규화 기법, 드롭아웃 등의 전략을 사용하는 것이 중요하다. 이로써 모델의 일반화 능력을 향상시키고, 영상의학 AI의 신뢰성을 높일 수 있다.

모델의 블랙박스 특성

영상의학에서 흔히 사용되는 복잡한 딥러닝 아키텍처의 블랙박스 특성은 모델 불확실성의 중요한 원인이자, 임상 적용의 주요 장애물이다(13-16). 이러한 모델은 투명성이 부족하여, 영상의학 전문의가 AI의 예측 근거를 이해하기 어렵게 만든다. 이 해석 불가능성은 고위험 임상 환경에서 설명 가능한 의사 결정 경로를 제공하는 데 익숙한 영상의학 전문의들에게 특히 문제가 된다.

예를 들어, AI 모델이 간 병변을 악성으로 분류했을 때 그 판단 근거를 제시하지 않으면, 모호한 사례에서 신뢰를 잃을 수 있다. 영상의학 전문의는 임상 경험과 맞지 않거나 쉽게 설명할 수 없는 AI 출력에 의존하는 것을 주저하게 되며, 이는 AI에 대한 신뢰를 떨어뜨린다.

최근 FDA 승인을 받은 AI 응용 프로그램이 증가하고 있음에도 불구하고, 여전히 많은 모델이 블랙박스 상태에 머물러 있어 임상 환경에서의 활용을 제한한다. 이러한 모델의 불투명성은 의료 제공자와 규제 기관이 성능뿐 아니라 이해 가능하고 신뢰할 수 있는 설명을 요구하게 만들고 있으며, 이는 AI의 넓은 채택을 가로막는 큰 장벽이 되고 있다. 따라서, AI 모델의 의사 결정 과정을 투명하게 제공할 수 있는 설명 가능한 AI 기술 개발이 시급하며, 이를 통해 모델 불확실성을 줄이고 AI의 수용과 통합을 촉진할 수 있다(Table 1).

Table 1. Key Sources of Uncertainty in Radiological AI Systems

Uncertainty Type	Specific Context	Description
Out-of-distribution uncertainty	Variability in imaging equipment	Models trained on high-quality images struggle with older or less sophisticated devices, leading to unpredictable performance and potential misclassification
	Population differences	Models trained on adult populations may misinterpret pediatric or geriatric images due to anatomical and physiological differences, resulting in incorrect diagnoses
	Regional and ethnic variability	Models trained on Western populations may misdiagnose conditions in diverse ethnic groups due to regional variations in disease presentation
	Changes in disease patterns	AI models often fail to recognize emerging conditions, like COVID-19, if they were trained on outdated data, emphasizing the need for continuous updates
Aleatoric uncertainty	Image artifacts	Artifacts such as motion blur, metal streaks, and beam hardening introduce noise, complicating the detection of subtle findings like small tumors
	Noise in image acquisition	Low-dose imaging techniques increase noise, which obscures critical anatomical details, complicating AI's ability to make accurate distinctions
	Inconsistent annotations	Variability in human annotations introduces inconsistencies in training data, leading to fluctuating and unreliable AI model predictions, especially in ambiguous cases
Model uncertainty	Model architecture choices	The selection of model architecture can significantly impact performance, with simpler models often struggling with complex tasks and advanced models requiring careful integration
	Overfitting to training data	Models that are overfitted to training data may perform well in development but fail in clinical practice due to poor generalization to new, unseen scenarios
	Black-box nature of models	The opaque nature of many AI models limits their interpretability, leading to skepticism among clinicians and barriers to clinical adoption

AI = artificial intelligence

영상의학 AI 모델에서 독립적 신뢰성 지표의 필요성

영상의학에서 AI 모델은 일반적으로 예측 확률을 출력하여 모델이 결정에 대해 얼마나 확신하는지를 나타내는 내부 신뢰성 점수를 제공한다. 그러나 이 점수는 훈련에 사용된 데이터, 레이블, 모델 아키텍처에 전적으로 의존한다. 모델 자체는 자신의 한계나 지식의 경계를 인식하지 못하며, 학습한 범위 내에서만 작동한다. 이는 AI가 자신이 모르는 것을 인지하지 못하는 철학적 문제를 일으킨다. 즉, AI는 자신이 알지 못하는 것을 알지 못하며, 학습한 패턴이 실제 현실을 반영하는지, 아니면 특정 데이터셋의 특수한 인공물인지를 인식하지 못한다(17).

이러한 근본적인 한계로 인해 AI 모델은 분포 외 데이터나 예상치 못한 변동, 예를 들어 새로운 환자 인구나 다른 영상 장비, 혹은 진화하는 질병 패턴을 만났을 때 지나치게 자신감을 가질 수 있다. 이러한 상황은 AI 모델이 경험해 보지 못한 것들이며, 그 결함을 인지하지 못한 채 잘못된 확신으로 예측을 내놓을 수 있다. 이는 임상 환경에서 특히 위험할 수 있으며, 부정확한 AI 출력이 환자 진료 결정에 직접적인 영향을 미쳐 오진이나 부적절한 치료로 이어질 수 있다.

이러한 문제를 해결하기 위해서는 독립적인 신뢰성 지표가 필요하다. 독립적 지표는 내부 신뢰성 점수와 달리, 새로운 테스트 데이터가 모델의 훈련 경험과 얼마나 잘 일치하는지를 외부에서 객관적으로 평가한다. 이러한 지표는 입력 데이터가 모델의 익숙한 범위를 벗어날 때 모델의 예측

신뢰성을 의심해야 할 가능성을 신호함으로써 불확실한 예측을 식별하는 데 도움을 줄 수 있다.

독립적인 신뢰성 지표를 도입하면 AI 시스템은 자신의 성능을 보다 정확하게 평가하고 불확실한 예측을 경고하여 영상의학 전문의가 보다 정보에 근거한 결정을 내릴 수 있도록 지원한다. 이러한 접근은 영상의학에서 AI의 안전성과 견고성을 높일 뿐만 아니라, 임상 AI 응용에서 설명 가능성과 신뢰성에 대한 요구를 충족시키는 데 기여한다. 궁극적으로 AI는 유용한 통찰력을 제공하면서도 고유한 맹점을 보완하는 안전장치를 필요로 하며, 이를 통해 AI는 환자 진료에서 신뢰할 수 있는 동반자로 자리 잡아, 무분별한 위험의 원천이 되지 않도록 보장할 수 있다.

영상의학에서의 예시 시나리오: 독립적 신뢰성 지표의 역할

AI 모델이 흔한 질환의 흉부 X-ray를 주로 훈련받았지만, 드문 간질성 폐질환이나 덜 흔한 폐 감염 같은 희귀 질환에는 충분히 노출되지 않았다고 가정해 본다. 임상 환경에서 AI 모델이 이러한 희귀 질환을 접했을 때도 여전히 높은 확신을 가진 예측을 생성할 수 있으며, 잠재적으로 이미지를 정상으로 진단하거나 질환을 더 익숙한 상태로 잘못 분류할 수 있다. 이러한 높은 확신은 실제 이해에 기반한 것이 아니라 훈련 중 학습했던 패턴에 근거한 것이기 때문에 매우 위험할 수 있다.

임상적으로 볼 때, 이 시나리오는 중요한 위험성을 강조한다. AI 모델은 자신의 한계를 인지하지 못하기 때문에 잘 알지 못하는 상황에 처해 있을 때 이를 깨닫지 못한다. 이때 독립적 신뢰성 지표의 가치가 빛을 발한다. 이 지표는 새로운 사례가 훈련 데이터와 얼마나 잘 일치하는지를 평가하는 외부적이고 객관적인 평가자로 작용한다. 만약 독립적 신뢰성 지표가 X-ray의 특징이 모델이 이전에 본 적이 없는 것과 큰 차이를 감지하면, 해당 예측을 불확실한 것으로 표시하거나, 결과를 주의 깊게 해석하고 영상의학 전문의가 추가 평가를 할 수 있도록 권장할 수 있다.

이러한 접근법은 임상 워크플로에서 중요한 안전장치로 작용된다. 모델의 내부 확신이 잘못된 경우를 식별함으로써, 독립적 신뢰성 지표는 영상의학 전문의가 잠재적으로 고위험이거나 특이한 사례에 더욱 집중할 수 있도록 돕고, 복잡한 질환이 간과되지 않도록 보장한다. 이 전략은 환자 안전을 강화할 뿐만 아니라, 인간의 감독이 중요한 상황에서 AI 지원 진단의 신뢰성을 유지하는 데 도움을 준다. 따라서 독립적 신뢰성 지표를 통합하는 것은 AI 예측과 임상 판단 사이의 격차를 메워, 영상의학에서 더 신뢰할 수 있는 AI 응용 프로그램을 실현하는 데 기여한다.

영상의학에서 설명 가능한 AI와 안전 장치의 중요성

영상의학에서 AI의 신뢰성과 안전성을 높이기 위해서는 설명 가능성과 추가적인 안전장치를 AI 모델에 통합하는 것이 필수적이다. 설명 가능한 AI는 모델이 어떻게 예측에 도달했는지를 설명함으로써, 임상의가 AI의 결정 논리를 이해하고 평가할 수 있도록 한다. 독립적 신뢰성 지표와 결합하면, 이러한 설명 가능성은 AI의 출력에 대한 명확한 근거를 제공하며, 잠재적 오류를 감지하는 중요한 안전망 역할을 한다.

임상 실무에서 설명 가능성은 영상의학 전문의가 익숙하지 않은 영상 조건이나 드문 질병을 마주했을 때 자신감을 조정하는 과정과 유사하다. 마찬가지로, AI가 예측에 대한 명확한 설명을 제공하면, 영상의학 전문의는 그 설명이 임상 기대와 일치하는지 평가할 수 있다. 만약 AI의 설명이

실제 영상 소견과 맞지 않는다면, 이는 AI의 예측이 신뢰할 수 없거나 잘못되었음을 경고하는 신호가 될 수 있다. 이 과정은 AI 출력이 신중히 다루어져야 할 때를 파악하고 영상의학 전문의가 세심하게 검토하도록 유도한다.

또한, AI의 설명과 예측 결과를 체계적으로 수집하고 분석하면 AI의 강점과 한계를 지속적으로 평가할 수 있다. AI가 잘 수행하는 영역과 반복적으로 어려움을 겪는 패턴을 분석함으로써, 임상의는 AI의 성능에서 개선이 필요한 특정 약점을 파악할 수 있다. 이 귀중한 피드백은 AI 모델의 업데이트, 재훈련 및 미세 조정을 위한 중요한 정보로 사용될 수 있으며, 특히 개선이 필요한 영역에 집중할 수 있다. 예를 들어, AI가 특정 유형의 폐 결절을 자주 잘못 분류하는 경우, 추가 데이터를 활용해 해당 영역에서 AI 모델을 재훈련함으로써 성능을 향상시킬 수 있다.

설명 가능한 AI는 신뢰성과 투명성을 높일 뿐만 아니라, 지속적인 학습과 개선의 메커니즘을 제공한다. 독립적 신뢰성 지표와 설명 가능성 도구를 결합해 AI 성능을 지속적으로 향상시키는 피드백 루프를 구축할 수 있으며, 시간이 지남에 따라 AI 모델은 더욱 신뢰할 수 있고 견고해질 것이다. 이러한 접근은 AI 기반 영상의학의 안전성을 높이고, AI 예측을 임상 전문성과 더욱 밀접하게 조정하여 환자 결과를 개선하는 데 기여한다.

영상의학 전문의와 AI의 협업: 독립적 신뢰성 지표를 통한 안전성과 정확성 향상

임상 실무에서 영상의학 전문의와 AI 모델 모두 익숙하지 않거나 복잡한 데이터를 다룰 때 어려움을 겪을 수 있다. 예를 들어, 영상의학 전문의는 새로운 영상 장비나 익숙하지 않은 MRI 제조사의 기기를 사용해 스캔을 해석할 때 자신감이 떨어질 수 있다. 마찬가지로, AI 모델도 훈련 데이터와 다른 데이터—예를 들어, 다른 장비로 촬영된 이미지, 새로운 질병 표현, 혹은 다양한 환자 인구—에 적응하는 데 어려움을 겪을 수 있다. 이러한 문제는 AI 시스템이 익숙하지 않은 데이터를 인식하고 이에 적절히 대응할 필요성을 강조한다.

AI 모델에 독립적 신뢰성 지표를 도입하면 이러한 불확실성을 해결하는 중요한 안전 메커니즘이 마련된다. 이 지표는 AI 시스템이 내부 신뢰성 점수와는 별개로 예측의 신뢰성을 평가할 수 있도록 한다. 예측이 불확실한 사례를 식별함으로써, AI 모델은 그러한 경우 신중한 검토를 표시할 수 있으며, 이는 영상의학 전문의가 익숙하지 않은 소견을 마주했을 때와 유사한 방식이다. 이러한 접근은 인간과 기계의 협력적 역학을 강화하며, AI의 계산 능력과 영상의학 전문의의 임상 전문성을 활용해 진단 정확도와 환자 안전을 향상시킨다.

독립적 신뢰성 지표는 AI가 보수적으로 작동하도록 하여, 높은 신뢰도를 가진 예측에서는 적극적으로 작동하고, 불확실한 경우에는 주의를 촉구하는 역할을 한다. 이 전략은 임상 워크플로에서 매우 중요하며, 잘못된 AI 출력이 심각한 결과를 초래할 수 있는 상황에서 매우 유용하다. 이러한 지표를 통합함으로써 AI 모델은 실제 데이터의 복잡성을 더 잘 처리할 수 있으며, 필요할 때 추가 평가를 유도하고, 성능을 임상 표준에 맞추도록 조정할 수 있다.

독립적 신뢰성 지표는 AI와 영상의학 전문의 간의 협업을 촉진하며, AI 시스템을 지속적으로 개선하는 피드백 루프를 형성한다. 이를 통해 AI 모델은 성능이 부족하거나 불확실한 영역에 대해 목표 재훈련과 미세 조정을 거쳐 발전할 수 있으며, 더 나아가 AI는 신뢰할 수 있는 진단 도구로 자

리 잡는다. 이러한 통합된 접근은 AI 지원 영상의학을 더욱 안전하고 효과적으로 만들어, 임상의와 환자 모두에게 혜택을 제공할 것이다.

실제 사례: 두개내 출혈 감지 및 분류

독립적 신뢰성 지표가 적용된 대표적인 사례는 두개내 출혈을 감지하고 이를 다섯 가지 아형으로 분류하는 AI 시스템이다. 이 시스템은 MGH (Boston, MA, USA)의 임상 환경에서 광범위하게 테스트되었으며, 안전성과 설명 가능성을 중시하는 방식으로 개발되었다(18, 19).

높은 신뢰도를 가진 사례에서의 선택적 작동

AI 시스템은 예측이 매우 신뢰할 수 있는 경우에만 감지를 활성화하여, 해당 사례를 표시하도록 프로그래밍되었다. 신뢰도가 낮은 경우에는 AI의 출력이 보류되며, 기존 임상 워크플로가 AI의 개입 없이 그대로 진행되도록 하여 환자 진료에 있어서 가장 신뢰할 수 있는 AI 통찰만 영향을 미치게 한다.

독립적 신뢰성 플래그 표시

정상으로 확신하는 사례에 대한 녹색 플래그

정상으로 식별된 사례에서는 시스템이 녹색 플래그를 발행하여 응급실(emergency department)에서 신경두경부 영상의학 전문의의 확인 후 조기 퇴원을 고려할 수 있도록 한다. 이 과정은 환자 흐름을 신속히 처리해 응급실 혼잡을 줄이고 대기 시간을 단축한다.

출혈로 확신하는 사례에 대한 적색 플래그

높은 신뢰도를 가진 출혈 사례에서는 적색 플래그가 발행되어 응급실 의사와 신경두경부 영상의학 전문의에게 경고를 보낸다. 이를 통해 신속한 진단 및 치료 조정이 이루어지도록 한다.

향상된 임상 워크플로 효율성

AI 시스템은 높은 신뢰도의 예측만 선택적으로 플래그를 표시해 임상 워크플로를 최적화하며, '환자 고속도로'를 만들어 의사 결정 과정을 가속화한다. 이는 안전성을 손상시키지 않으면서도 운영 효율성을 크게 높인다. 이 모델은 설명 가능한 AI가 어떻게 임상 실무에 원활하게 통합되어 운영 효율성과 환자 결과를 향상시킬 수 있는지를 보여주는 좋은 사례이다.

설명 가능성을 갖춘 오류 무관용(Zero Error Tolerance) AI로의 전환

오류 무관용(zero error tolerance) 접근 방식을 적용한 AI 모델의 개발은 임상 환경에서 AI 통합의 새로운 패러다임을 의미하며, 신뢰할 수 있는 상황에서만 작동하는 AI 시스템의 배포에 중점을 둔다. 이 접근 방식은 AI가 매우 확실할 때만 예측을 하도록 보장하며, 테스트에서 오류가 전혀 발생하지 않은 경우에만 작동해 불확실하거나 애매한 출력에 의존하지 않도록 한다. 이러한 모델

은 오류가 제거된 시나리오에만 사용되도록 하여, 특정하고 명확한 작업에서 완벽하게 신뢰할 수 있는 임상 의사 결정 도구로 작동한다.

오류 무관용(zero error tolerance) 전략은 설명 가능성과 투명성을 강조하여 임상의가 AI의 의사 결정 과정을 이해하고 신뢰할 수 있게 한다. 이 전략은 철저한 모델 검증과 지속적인 평가를 통해, AI가 정의된 범위 내에서 일관되게 신뢰할 수 있는 결과를 제공하는지 확인한다. AI가 오류 없는 결과를 보장할 수 있는 시나리오에서만 작동하도록 함으로써 이 모델은 임상 실무에서 안전성, 신뢰성, 책임성의 새로운 기준을 설정한다.

이 방법은 현재 실제 임상 환경에서 테스트 중이며, AI가 환자 안전을 저해하지 않으면서도 영상의학과 같은 의료 분야에 안전하게 통합될 수 있음을 입증하는 것을 목표로 하고 있다. 오류 무관용(zero error tolerance) AI 시스템을 사용함으로써, 임상의는 AI의 성능이 오류를 최소화하도록 설계되었음을 알고, 특정 작업에서 이러한 시스템을 자신 있게 신뢰할 수 있다. 이러한 접근 방식은 의료에서 AI의 안전성과 효율성을 높일 뿐만 아니라, 임상적 탁월성의 최고 기준을 준수하는 AI 기술의 광범위한 수용과 활용의 길을 열어준다.

토의 및 결론

AI는 진단의 정확성과 효율성, 의사 결정을 향상시키며 영상의학을 변화시키고 있다. 그러나 분포 외 데이터, 데이터 내재적 불확실성, 모델 불확실성에서 비롯되는 예측 불확실성은 여전히 중요한 과제로 남아 있다. 독립적 신뢰성 지표와 설명 가능한 AI 모델을 도입하는 것은 안전성과 신뢰성을 보장하는 데 필수적이며, 이를 통해 영상의학 전문의는 AI의 출력을 이해하고 신뢰할 수 있는 투명성을 확보할 수 있다. 이러한 접근 방식은 AI와 임상의 간의 협력을 강화하여 복잡한 사례에서도 계산적 통찰력과 임상 전문성이 함께 작용하도록 한다.

오류 무관용(zero error tolerance) AI 모델의 개발은 높은 신뢰도 시나리오에서만 작동하도록 함으로써, 안전성과 신뢰성의 새로운 기준을 세우는 중요한 진전을 보여준다. 이러한 모델은 실제 임상 테스트에서도 유망하며, 환자 진료를 저해하지 않으면서 임상 의사 결정을 크게 향상시킬 수 있는 잠재력을 보여준다. AI가 계속해서 발전함에 따라, 훈련 데이터셋을 확장하고 모델 아키텍처를 개선하며 설명 가능성을 높임으로써 불확실성을 줄이고 복잡한 임상 시나리오에 적응할 수 있는 모델을 개발하는 데 중점을 두어야 한다.

AI가 진단 과정에서 신뢰할 수 있는 동반자로 자리 잡을 것으로 기대되는 가운데, 영상의학의 미래는 매우 밝다. 예측 불확실성을 해결하고 오류 무관용(zero error tolerance) 전략을 도입함으로써, AI는 영상의학의 정밀성과 안전성을 강화할 수 있다. 견고하고 설명 가능한 AI 시스템의 원활한 통합을 향한 이 여정은 환자 진료의 기준을 새롭게 정의하며, 의료의 새로운 탁월성 기준을 설정할 것이다.

Supplementary Materials

English translation of this article is available with the Online-only Data Supplement at https://doi.org/10.3348/jksr.2024.0118.

Conflicts of Interest

Synho Do has been an Editorial Board Member of the Journal of the Korean Society of Radiology since 2021; however, he was not involved in the peer reviewer selection, evaluation, or decision process for this article. Otherwise, no other potential conflicts of interest relevant to this article were reported.

ORCID iD

Synho Do https://orcid.org/0000-0001-6211-7050

Funding

None

REFERENCES

1. Chua M, Kim D, Choi J, Lee NG, Deshpande V, Schwab J, et al. Tackling prediction uncertainty in machine learning for healthcare. *Nat Biomed Eng* 2023;7:711-718
2. Candemir S, Nguyen XV, Folio LR, Prevedello LM. Training strategies for radiology deep learning models in data-limited scenarios. *Radiol Artif Intell* 2021;3:e210014
3. Lambert B, Forbes F, Doyle S, Tucholka A, Dojat M. Improving uncertainty-based out-of-distribution detection for medical image segmentation. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.2211.05421. Accessed September 25, 2024
4. Onder O, Yarasir Y, Azizova A, Durhan G, Onur MR, Ariyurek OM. Errors, discrepancies and underlying bias in radiology with case examples: a pictorial review. *Insights Imaging* 2021;12:51
5. Sambyal AS, Krishnan NC, Bathula DR. Towards reducing aleatoric uncertainty for medical imaging tasks. Available at. https://doi.org/10.1109/ISBI52829.2022.9761638. Published 2022. Accessed September 25, 2024
6. Monteiro M, Le Folgoc L, Coelho de Castro D, Pawlowski N, Marques B, Kamnitsas K, et al. Stochastic segmentation networks: modelling spatially correlated aleatoric uncertainty. Available at. https://proceedings.neurips.cc/paper/2020/hash/95f8d9901ca8878e291552f001f67692-Abstract.html. Published 2020. Accessed September 25, 2024
7. Wang G, Li W, Aertsen M, Deprest J, Ourselin S, Vercauteren T. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing (Amst)* 2019;335:34-45
8. Chung J, Kim D, Choi J, Yune S, Song KD, Kim S, et al. Prediction of oxygen requirement in patients with COVID-19 using a pre-trained chest radiograph xAI model: efficient development of auditable risk prediction models via a fine-tuning approach. *Sci Rep* 2022;12:21164
9. Frénay B, Verleysen M. Classification in the presence of label noise: a survey. *IEEE Trans Neural Netw Learn Syst* 2014;25:845-869
10. Rolnick D, Veit A, Belongie S, Shavit N. Deep learning is robust to massive label noise. arXiv [Preprint]. Available at. https://doi.org/10.48550/arXiv.1705.10694. Published 2017. Accessed September 25, 2024
11. Jang R, Kim N, Jang M, Lee KH, Lee SM, Lee KH, et al. Assessment of the robustness of convolutional neural networks in labeling noise by using chest X-ray images from multiple centers. *JMIR Med Inform* 2020;8:e18089
12. Ju L, Wang X, Wang L, Mahapatra D, Zhao X, Zhou Q, et al. Improving medical images classification with label noise using dual-uncertainty estimation. *IEEE Trans Med Imaging* 2022;41:1533-1546
13. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. *ACM Comput Surv* 2018;51:1-42
14. Castelvecchi D. Can we open the black box of AI? *Nature* 2016;538:20-23
15. Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *J Med Ethics* 2021;47:329-335
16. Yang G, Ye Q, Xia J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond. *Inf Fusion* 2022;77:29-52
17. Kim D, Chung J, Choi J, Succi MD, Conklin J, Longo MGF, et al. Accurate auto-labeling of chest X-ray images based on quantitative similarity to an explainable AI model. *Nat Commun* 2022;13:1867
18. Lee H, Yune S, Mansouri M, Kim M, Tajmir SH, Guerrier CE, et al. An explainable deep-learning algorithm for

the detection of acute intracranial haemorrhage from small datasets. *Nat Biomed Eng* 2019;3:173-182
19. Yoon BC, Pomerantz SR, Mercaldo ND, Goyal S, L'Italien EM, Lev MH, et al. Incorporating algorithmic uncertainty into a clinical machine deep learning algorithm for urgent head CTs. *PLoS One* 2023;18:e0281900

의료 영상 분석을 위한 설명 가능하고 안전한 인공지능

도신호[1,2,3]*

인공지능(artificial intelligence; 이하 AI)은 진단 정확도와 효율성을 높여 영상의학 분야에 변화를 가져오고 있지만, 예측 불확실성은 여전히 중요한 과제로 남아 있다. 본 리뷰에서는 주요 불확실성의 원인인 분포 외(out-of-distribution) 불확실성, 데이터 내재적 불확실성(aleatoric uncertainty), 모델 불확실성을 다루며, 안전한 AI 통합을 위해 독립적인 신뢰성 지표와 설명 가능한 AI의 중요성을 강조한다. 독립적인 신뢰성 지표는 AI 예측의 신뢰성을 평가하는 데 기여하며, 설명 가능한 AI는 투명성을 제공하여 AI와 영상의학 전문의 간의 협업을 강화한다. 오류 무관용(zero error tolerance) 모델의 개발은 오류를 최소화하도록 설계되어, 안전성의 새로운 기준을 제시하였다. 이러한 문제를 해결함으로써 AI는 영상의학에서 신뢰할 수 있는 동반자로 자리 잡아, 환자 진료 수준과 결과를 개선하는 데 기여할 것이다.

[1]하버드 의과대학 매사추세츠종합병원 영상의학과,
[2]고려대학교 KU-KIST 융합대학원,
[3]하버드대학교 Kempner 연구소