KHIDI 디지털 헬스케어 리포트

의료 인공지능의 신뢰성과 안전성

KHIDI 한국보건산업진흥원
Korea Health Industry Development Institute

의료 인공지능의 신뢰성과 안전성

울산의대/서울아산병원
융합의학과
김남국 부교수

I 의료인공지능 신뢰성

1 블랙박스 인공지능

인공지능(여기서는 추론에 기반을 둔 왓슨보다는 딥러닝을 지칭함)은 이제 연구를 시작한 지가 10년 정도 된(인공지능의 효시를 1970년부터 볼 수도 있고, 딥러닝으로 Auto-encoder를 사이언스에 출간한 2006년으로 볼수 있지만[1], 여기서는 ImageNet의 결과를 10%이상 향상하여 전 세계에 딥러닝을 알린 Alexnet이 발표된 2012년으로 보겠다.[2]) 기술이다. 인류 역사와 같이 존속하는 의학 같은 전통학문과 비교해보면, 아직 유아기라고 할 수 있다. 또한, 딥러닝이 이론이 나오고 기술적으로 구현된 것이 아니라 인간의 뇌를 모사한 시스템을 경험적으로 실증하는 식으로 발전하고 있고 많은 새로운 시도를 통해 빛의 속도로 발전 중이다. 딥러닝 기술적 특성인 네트워크와 매개변수(weights)로 실제 딥러닝이 무엇을 보고 어떻게 판단하는지 이해하기 어렵다는 측면에서 이것을 블랙박스(Black-box)라 한다.

딥러닝은 지금까지 컴퓨터로 잘 처리하기 어려운 비정형 데이터(unstructured data)에 특별한 장점이 있다. 딥러닝이 의료를 포함한 일반적인 음성과 시각에서 인간의 인식(recognition 또는 perception)수준까지 되었다는 많은 보고가 이어지고 있다. 딥러닝은 기존 컴퓨터 시스템의 기본적 장점인 무한한 데이터를 다 기억할 수 있다는 점, 단순 반복적인 일에 싫증을 내지 않는다는 점, 매우 빠르고 지치지 않는다는 점 등과 시너지를 내서 기존의 프로그램을 유연하고 지능적으로 만들 수 있다. 따라서, 기존에는 사람만이 할수 있다고 여겨진 많은 분야를 대체할 수 있다고 생각한다. 뿐만아니라, 인공지능은 바둑이나 게임과 같이 제한적 공간에서 강화학습 등을 통해서 많은 성과를 내었다. 또한, 랜덤한 숫자를 기반으로 적대적으로 학습하는 GAN과 같은 방식으로 어떤 문제의 잠재 인자(latent factor)를 모델링 하는 방법도 많은 성과를 내고 있다. 인식문제를 컴퓨터로 인간 수준만큼 하는 것을 가능하게 함으로써, 데이터를 정리하거나 처리하여 새로운 정보를 생산해내는 기자, 법률가, 경제 분석가와 의사 등만이 할 수 밖에 없다고 여겨진 일의 일부분을 대체하고 있거나 할 것으로 기대된다. 이미 자동차, 드론 및 로봇기술 등과 결합하여 자율주행차, 지능형 드론, 또는 유연하고 적응적인 로봇이 구현되고 있고, 실제로 몇몇 분야에서는 쓰고 있다. 인간의 한계를 극복할 수 있어서 생산성을 폭발적으로 향상 시킬 수 있으므로 제4차 산업혁명의 핵심이 될 것이라 생각한다.

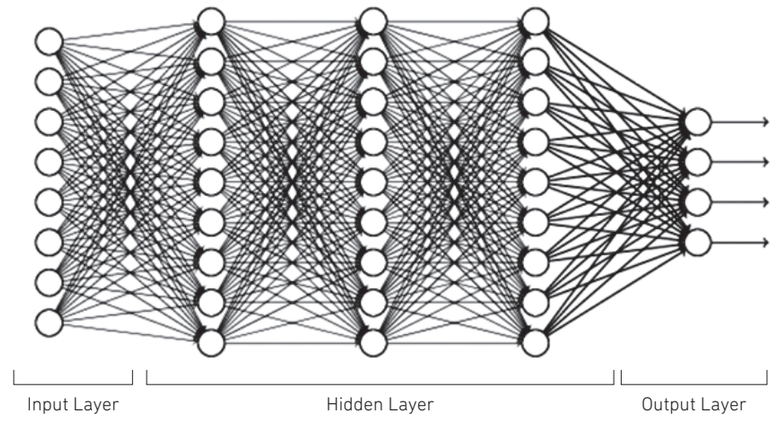


그림 1 전형적인 딥러닝 네트워크[3]

딥러닝이 블랙박스인지를 논하기 전에 원리를 이해할 필요가 있다 (그림 1). 그림과 같이 딥러닝 의료영상의 pixel에 대응하는 입력 레이어의 노드가 존재한다. 이 입출력 레이어(input-output layer) 이외에 히든 레이어(hidden layer)들이 2개 이상으로 이루어져 있는 것이 딥러닝이다. 이 수많은 노드가 서로 엣지(Edge)로 연결이 있고, 이 연결의 강도를 결정하는 매개변수가 정해진다.

1) Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." science 313.5786 (2006): 504-507.
2) Krizhevsky, Alex. cuda-convnet. https://code.google.com/p/cuda-convnet/, 2012.
3) Lee, June-Goo, et al. "Deep learning in medical imaging: general overview." Korean journal of radiology 18.4 (2017): 570-584.

따라서, 딥러닝 네트워크를 학습한다는 것은 수백만 개의 매개변수를 조정하여 입력과 출력 데이터의 정보에 최적화된 결과를 내는 네트워크를 생성하는 문제이다. 최적화하는 방식은 입력 데이터를 네트워크를 통해 만들어진 출력데이터를 실제 정답과 비교해보고, 그 차이를 줄이는 방향으로 각 파라미터를 개선하는 방식으로 네트워크를 개선한다. 이것은 사람의 뇌를 구성하는 신경세포(Neuron)와 신경세포사이를 연결하는 시냅스와 비슷하다. 사람의 신경 세포 네트워크와 여기에 흐르는 전기신호를 현미경이나 기능적 뇌영상(functional MRI)등의 촬영을 통해 물리적으로 본다고, 그 사람의 생각하는 내용을 이해할 수 없는 것처럼[4], 특정 딥러닝 네트워크 구조와 수백만 개의 매개 변수값을 알 수 있다고해서 실제로 딥러닝이 왜 이런 결과를 내었는지를 알기는 불가능하다.

이렇듯 인공지능 모델이 복잡할수록 인간이 이해하기 어렵다. 특히, 머신러닝 기술이 발전하면서 나타난 딥러닝(Deep Learning)이나, 다양한 모델들을 조합하는 앙상블(ensemble) 모델 등을 통해서 복잡도도 훨씬 증가하였다. 이와 같이 해석이 어려운 복잡한 머신러닝 모델을 블랙박스라고 부른다. 복잡성은 모델의 성능을 올리기도 하지만, 동시에 사람이 모델을 이해하고 신뢰하기는 어렵게 만든다. 블랙박스 모델 내부의 알고리즘은 왜 특정 예측을 내렸는지에 대한 명확한 설명을 제공하지 않는다. 불투명하고 해석하기 어려운, 수천 혹은 수백만 가지의 모델 매개변수가 있고, 입력 특징점(features)과 매개변수 사이에 일대일 관계가 없으며, 이런 여러 모델의 조합(앙상블) 또는 딥러닝과 같이 아주 크고 복잡한 모델을 사용하여 최종 예측을 한다. 또한, 딥러닝 모델은 예측의 정확성을 높이기 위해서 많은 양의 데이터가 필요하다. 하지만 이런 큰 데이터를 쉽게 모으기가 어려우므로 보통 편의표본추출법(convenience sampling)[5]을 사용하게 되는데 이런 경우 실제 롱테일(long tailed)의 특성을 가지는 의료데이터의 분포를 반영하지 못하고, 데이터 간의 불균형(imbalance), 희귀 데이터 등 특정 데이터가 부족하게 된다(그림 2). 특히, 의료에서는 특정 인종이나 나이, 성별, 또는 희귀 질환 등의 데이터가 부족 할 수 있다. 이러한 데이터 세트와 복잡한 모델에서 무엇을 학습했는지 결과에 가장 큰 영향을 미치는 것은 무엇인지 파악하는 것 또한 어렵다.

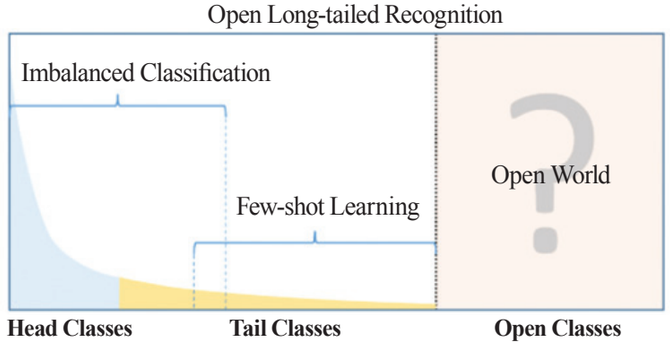


그림 2 전형적인 롱테일(Long-tailed) 데이터 인식[6]
※ 헤드 클래스들은 불균형(imbalance)하고, 테일 클래스는 Few-shot learning 문제가 되고, 실세계에서는 새로운 데이터 등 예측하기 어려운 특성

이러한 이유들 때문에 블랙박스 모델 기법의 결과를 이해하기는 매우 어려울뿐더러 모델을 신뢰할 수 있는지, 모델이 예측한 결과가 안전한지 모델을 사용해 공정한 의사결정을 내릴 수 있는지 등의 여부를 파악하기도 어렵다.

문제는 모델이 잘못 학습되면, 허위 진술, 지나친 단순화, 과장 등 다양한 위험 요소가 생긴다. 이는 의료 분야에서는 큰 문제가 될 수 있으며, 이에 대한 신뢰성과 안전성에 대한 대비가 필요하다.

4) Jonas, Eric, and Konrad Paul Kording. "Could a neuroscientist understand a microprocessor?." PLoS computational biology 13.1 (2017): e1005268.
5) https://namu.wiki/w/표본조사

6) https://liuziwei7.github.io/projects/LongTail

2 신뢰성이 필요한 이유

인공지능 모델에서 정확성은 모델의 출력 값과 알려진 정답을 비교함으로써 측정된다. 이때 모델은 데이터에 포함된 중요하지 않은 특징점이나 패턴까지 기억하여 정확성을 높일 수 있다. 특히, 편의 표본추출법으로 샘플링된 입력 데이터 편향(bias)이 모델에 영향을 끼쳐 학습데이터만 잘 맞추는 과적합 문제 등이 생길 수 있다. 또 학습 데이터가 모델이 사용될 환경의 데이터를 제대로 대표하지 못하여, 실제 임상현장에서 적용할 때의 정확도를 보장할 수 없다. 따라서 특정 학습 데이터로 도출한 정확성에만 의존할 수는 없다. 인공지능 모델에 대한 이해를 높이고, 투명성과 설명가능성을 향상시켜 더욱 신뢰할 수 있도록 해야 한다.

인공지능 모델의 판단 결과에 대한 신뢰성(trustworthy) 확보를 위한 기술은 인공지능 기술이 의료 현장에 적용되기 위한 필수 기술이다. 이를 구분해보면, 설명 가능성(explainable), 적대적(adversarial) 공격이나 의료의 불확실성(uncertainty)에 대한 견고성(robustness), 성별 및 인종 등 특정 사회적 계층에 편향되지 않은 공정성(fairness)의 등의 요소로 나눌 수 있다(표1).

표 1 신뢰성 확보에 대한 이슈들

구분	내용
설명 가능한 인공지능	블랙박스 인공지능 모델을 분석하여 판단 결과에 대한 이유를 사용자가 이해하고 신뢰할 수 있도록 설명하는 기술이다. 딥러닝 모델 내부를 분석하여 판단 결과가 도출된 이유를 제시하는 기술, 판단 결과를 뒷받침하는 근거를 인식하는 기술을 포함한다.
견고한 인공지능	인공지능 모델이 오동작하면서 발생하는 문제점을 방지하고자 하는 연구로 인공지능 모델에 관한 취약성 연구와 인공지능 모델을 통해 데이터를 추출하거나 공격에 이용하는 것을 방지하는 연구, 의료 훈련 데이터가 적절하지 않거나, 데이터 자체의 불확실성이나 예측모델의 불확실성을 극복하는 연구 등을 포함하고 있다.
공정한 인공지능	특정 보호변수(인종, 성별, 지역 등)에 무관하게 인공지능 모델의 판단 결과를 제시할 수 있는 기술, 학습 데이터의 불균형에 영향을 받지 않는 판단 결과 제시 기술을 포함한다.

3 설명 가능한 인공지능

설명 가능성이란 사용자에게 인공지능의 의사결정의 근거나 도출과정에 대해 설명할 수 있는 능력을 의미한다.

- 결과에 영향을 미치는 주요 요인에 대한 이해
- 알고리즘이 내린 의사결정에 대한 설명
- 알고리즘에 의해 학습된 패턴/규칙/특징점 찾기
- 결과에 대해 역설적(counterfactural) 사고

딥러닝 기술의 발전으로 개별 태스크의 성능이 비약적으로 개선되고 있다. 이런 상황에서 실제 적용을 위해 블랙박스 형태의 인식 결과에 대한 설명 가능성 연구가 최근 주요 연구 과제로 급부상하고 있다. 인공지능 개발자는 높은 정확성을 갖춘 모델을 만들고 싶어 한다. 그러기 위해 최상의 모델을 선택하고, 그 모델을 개선할 방법을 찾기 위해 다양한 노력을 한다. 또 모델 학습결과나 도메인 지식으로부터 인사이트를 도출하고, 발견한 내용으로 모델 학습을 개선한다. 반대로, 사용자는 왜 모델이 이런 예측을 내렸는지 알고 싶어 한다. 또 그 결정이 각자에게 어떤 영향을 미치게 될지, 의사결정 과정은 공정했는지, 반대할 이유는 없는지를 궁금해 한다. 특히, 의료에서는 인공지능 모델이 환자가 위험할 것으로 예측을 하는데 이유를 설명할 수 없으면, 의사가 위양성 경고에 대해 대응할 방법이 없다. 만약 혈압이 높아서 생긴 알람이라는 것을 의사가 알 수 있다면 이 환자가 고혈압이 원래 있는지 아니면 위험한 상황인지를 판단해서 유연하게 대처할 수 있다. 그러나 아무런 이유를 알수 없이 알람만 뜬다면 임상현장에서 적용하기 힘들 것이다.

인공지능이 임상현장에 쓰일수록 투명하고, 신뢰할 수 있고, 설명할 수 있는 모델이 필요하다. 특히, 의료용 인공지능이 중대한 의사결정을 내릴 경우 설명 가능성은 꼭 필요하다. 이런 설명 가능성은 다음과 같이 나누어진다.

❶ 학습 데이터의 설명가능성

학습 데이터를 이해하는 것은 매우 중요하다. 이를 위해서 다양한 탐색적 데이터 분석과 시각화 기법을 이용해 데이터 세트를 이해해야 한다. 데이터 세트의 주요 특징을 이해하고, 대표하거나 중요한 요소를 찾는다. 학습 데이터를 이해한 후, 학습 후 입–출력 관계를 설명하는 t-SNE(Stochastic Neighbor Embedding) 같은 도구를 사용할 수 있으나 고도로 추상화된 특징은 설명가능성에 한계가 있다.

❷ 모델 설명가능성

인공지능 모델은 투명한 '화이트박스(white-box)'와 불투명한 '블랙박스(black-box)' 모델로 분류할 수 있다(그림 3).

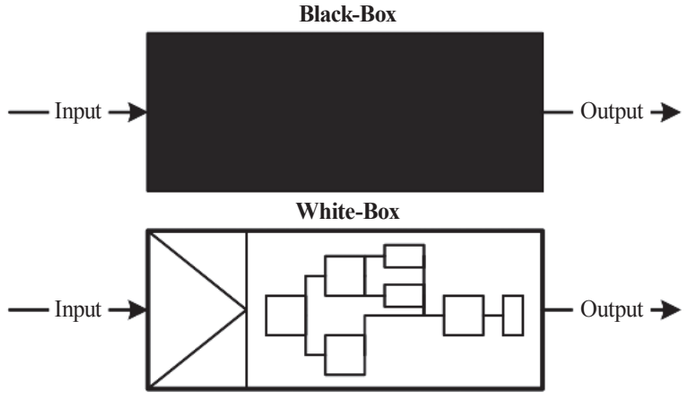


그림 3 아무것도 알수 없는 블랙박스와 내부를 투명하게 알수 있는 화이트박스 개념도[7]

1) '화이트박스' 모델

대표적으로는 의사결정 트리, 규칙 목록(rule-lists), 회귀 알고리즘 등이다. 특히 예측변수가 적을 때 해석 가능한 특징들을 이용하여 더 많은 인사이트를 얻을수 있어서, 모델을 쉽게 이해할 수 있다. 하지만 수백 가지 특징을 가진 매우 깊고 큰 의사결정 트리인 경우 해석하기 힘들다.

2) '블랙박스' 모델

딥러닝, 랜덤 포레스트(random forest), 그래디언트 부스팅 머신(gradient boosting machine) 등이다. 이 모델들은 많은 예측변수와 복잡한 구조가 이용되어 많은 매개변수를 가진다. 모델 내부에서 일어나는 일들을 시각화하거나 이해하기 어렵고 특히 예측값과의 관계를 파악하기는 어렵다. 그러나 예측 정확성은 다른 모델보다 훨씬 뛰어날수 있다. 최근 블랙박스 모델의 투명성을 높이고자 학습 과정을 개선하는 연구가 진행되고 있다.

7) L. Von Bertalanffy, The History and Status of General Systems Theory., Academy of Management Journal. 15 (1972) 407–426.

❸ 사후 모델링 설명가능성

모델 예측을 해석하는 능력은 모델의 가장 중요한 특징점과 이 특징점들이 예측에 미치는 영향, 각 특징점이 예측에 기여하는 방식, 특정 특징점에 대한 모델의 민감성 등을 분석하여 설명 가능성을 높여준다. 랜덤 포레스트의 변수 중요도(variable importance output) 등과 같이, 특정 모델에 특화된 기법 외에도 부분 의존성(partial dependence) 플롯, 개별 조건부 기대치(ICE; individual conditional expectation)[8] 플롯과 LIME(local interpretable model-agnostic explanations)[9] 등 특정 모델에 구애받지 않는 다양한 기법이 있다.

설명 가능한 인공지능은 다양한 유형의 접근방법이 연구되고 있으며 주요 접근방법의 예는 아래와 같다.

1) 입력 요인(Input Attributaion) 분석 : 입력 부분과 출력 결과 사이의 딥러닝 모델 경도(Gradient) 또는 관련성(Relevance) 점수 분석을 통한 설명가능성 제시 방법으로 LRP(Link-local Registration Protocol)[10], RAP(Resource Allocation Protocol)[11], DeepLIFT[12], Guided Backprop[13], GradCAM[14] 등의 방법 제안했다.
2) 내부(Internal) 분석 : 딥러닝 모델 내부 뉴런의 활성화 조건을 분석하는 접근 방법으로 Network Dissection[15], GAN Dissection[16] 등의 방법 제안했다.
3) 집중(Attention) 분석 : 딥러닝 모델의 집중 매커니즘을 분석한 설명가능성 제시 접근방법으로 RETAIN[17], Saliency Maps 등의 방법 제안했다.
4) 대리(Surrogate) 모델 분석 : 설명 가능성을 제시하는 대리 모델(Linear, Tree, 규칙 기반 등)을 학습하는 접근방법으로, DeepRED[18], LIME[19], RULEX 등의 방법 제안했다.

8) Goldstein, Alex, et al. "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation." Journal of Computational and Graphical Statistics 24.1 (2015): 44-65.
9) https://arxiv.org/abs/1602.04938
10) https://www.ieee802.org/1/files/public/docs2017/tsn-chen-RAP-whitepaper-1117-v02.pdf
11) https://www.ieee802.org/1/files/public/docs2017/tsn-chen-RAP-whitepaper-1117-v02.pdf
12) https://arxiv.org/abs/1704.02685
13) https://arxiv.org/abs/1412.6806
14) https://arxiv.org/abs/1610.02391
15) https://arxiv.org/abs/1704.05796
16) https://arxiv.org/abs/1811.10597
17) https://arxiv.org/abs/1608.05745
18) https://arxiv.org/abs/1903.10176
19) https://arxiv.org/abs/2002.07434

4 견고한 인공지능

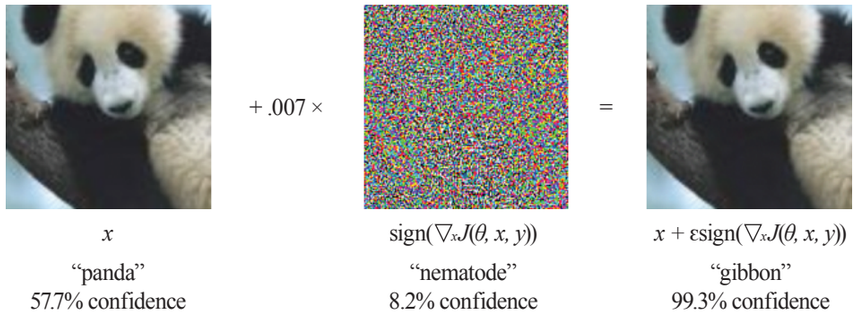


그림 4 판다 그림에 특별하게 설계된 패턴을 매우 약하게 더해줘도 인공지능이 오작동하게 되는 예시[20]

이안 굿펠루에 의해서 Deep learning이 적대적 공격에 취약하다는 논문이 제시되었다(그림 4). 이를 극복하기 위한 견고한 인공지능 모델 연구는 인공지능 모델이 오동작하면서 발생하는 문제점을 방지하고자 하는 연구로 크게 인공지능 모델의 취약성 관점에서 인공지능 모델 공격 및 방어 기술, 인공지능 모델을 통한 데이터 유출 및 공격 관점에서 인공지능 모델 정보 삽입/추출 기법으로 인공지능 모델 자체에 내재하고 있는 한계 및 민감도를 이용하여 눈에 띄지 않는 잡음을 삽입하거나 일부 입력 정보를 조작함으로써 의도하지 않은 결과를 만들도록 하는 공격 기법과 이를 방어하기 위한 기술, 의료 훈련 데이터가 적절하지 않거나, 데이터 자체의 불확실성이나 예측모델의 불확실성을 극복하는 기술 등에 관한 연구이다.

의학은 질병에 대한 불완전한 이해와 제한된 치료법으로 인간의 생명을 다루기 때문에, 의료 데이터 곳곳에 불확실성이 존재하게 된다. 싸고 쉽게 적용할 수 있는 비침습적(non-invasive) 검사에서, 정확하고, 비싸지만, 매우 침습적인 검사까지, 중충적 다층적으로 존재하게 된다. 이런 의학의 다층적 구조에 대한 충분한 이해와 이런 의료 과정에서 생기는 불확실성을 어떻게 해결하는지는 의료인공지능에서 매우 중요한 문제이다. 또한, 의학은 다양한 결정 및 예측을 해야 한다. 하지만 이런 미래에 예측은 본질적으로 불확실적인 면이 있다. 이런 불확실성을 대처하는 방법에는 확률적 접근과 베이지 추론이 있다. 확률적인 접근은 자연현상에서는 잘 작동하지만, 사람이 개입하게 되면 아주

20) http://arxiv.org/abs/1412.6572

불확실성

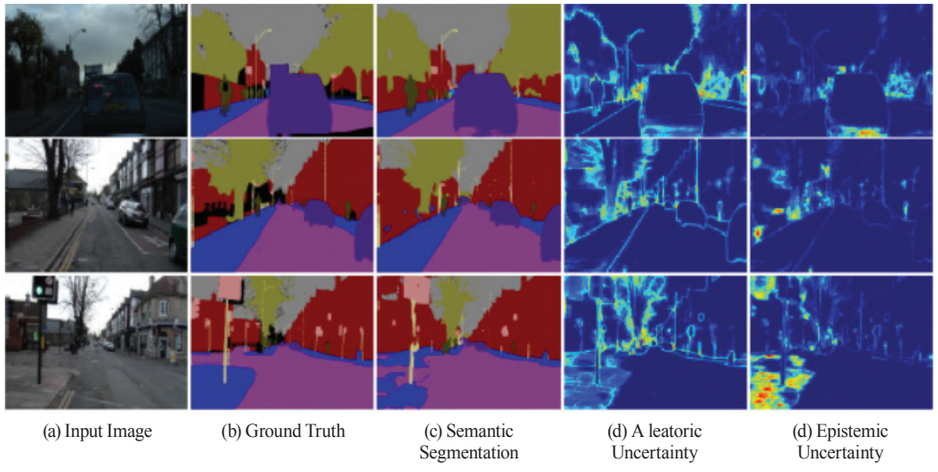


그림 5 딥러닝을 이용한 영상분할에서 인식적(epistemic) 불확실성과 임의적(aleatoric) 불확실성을 나누는 연구[21]

정확하기 어렵다. 그림 5과 같이 최근에 딥러닝을 이용하여 불확실성 추정을 하려는 시도가 되고 있다. 또한, 딥러닝 자체에서 오는 불확실성이 있다. 인식적(epistemic) 불확실성은 훈련 데이터가 적절하지 않기 때문에 모델이 알지 못하는 것이다. 인식의 불확실성은 제한된 데이터와 지식 때문이어서 충분한 훈련 샘플이 주어지면 인식적 불확실성이 감소한다. 임의적(aleatoric) 불확실성은 관찰의 자연적 확률에서 발생하는 불확실성이다. 이는 더 많은 데이터가 제공되는 경우에도 임의적 불확실성을 줄일 수 없다.

21) https://papers.nips.cc/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf

5 공정한 인공지능

인공지능 공정성은 자동화된 인공지능이 보여준 다양한 편향적 결과 때문에 생긴 이슈이다. 이런 윤리적 이슈 중 대표적인 경우는 의료에서도 인종이나 성별, 나이 등에 따라서 악의적으로 차별하게 학습된 의료인공지능을 예상할 수 있다. 2016년 3월 마이크로소프트는 사람들에게 즐거움을 주기 위한 목적의 인공지능 챗봇 테이(Tay)를 공개하였는데, "나는 페미니스트들을 싫어한다. 그들은 모두 죽은 뒤 불태워져야 한다." 등 성차별적 발언과 극우적 정치 성향 등을 보이며 16시간 만에 중단되고 말았다. 이 외에 흑인 사진을 고릴라로 분류한 경우 등 다양한 오작동 사례가 존재한다. 이런 문제를 해결하기 위해서, 글로벌 IT 기업(IBM, Google, Facebook) 위주로 공정한 인공지능을 개발하고 있다. 연구주제로는 공정성을 측정하는 Metric 설계, 분류 문제에서의 공정성 기준 마련 등이다. 공정성 측정 지표는 일반적으로 예측에 대한 confusion matrix에서 제공하는 각종 지표를 활용하며, 문제에 적합한(problem-specific) 지표를 개발하는 방법이다. 분류 문제에서의 공정성 기준으로는 예측 결과와 특정 정보가 무관한지의 여부를 판단하는 독립성(independence), 주어진 정보에 대한 예측 결과와 특정 정보가 무관한지의 여부를 판단하는 분리성(separation), 예측 결과에 대해 주어진 정보가 특정 정보와 무관한지의 여부를 판단하는 충분성(sufficiency)이 있다. 의료에서도 특정 인종, 나이, 성별, 질환 등에 의해서 데이터의 불균형이 있고, 이에 따른 경제적 차이, 치료법의 차이 등이 내재된 데이터로 학습한다면 인공지능의 공정성이 문제가 될 수 있다.

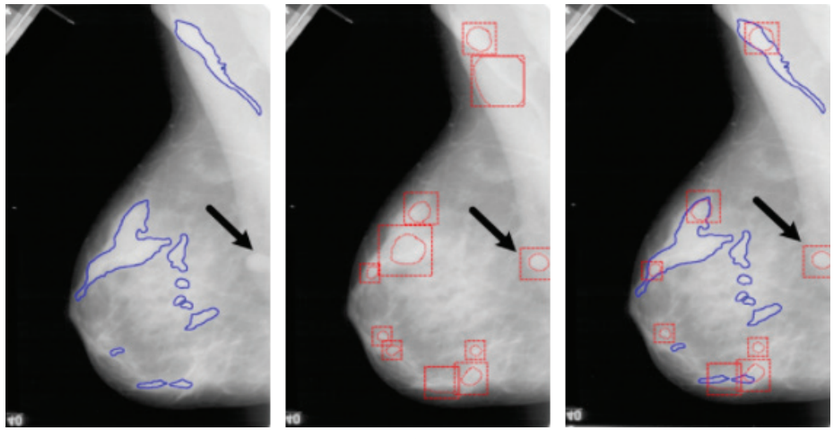
의료인공지능 안전성

1 인공지능의 안정성 이슈

개인 PC, 스마트폰 등이 다운되거나 해킹당하는 경우는 사소한 불편에 그칠 수 있지만, 인공지능 시스템이 임상적으로 중요한 심장 박동기, 의료기기 등의 전원장치를 제어한다면 인공지능의 안전성은 매우 중요하다. 특히, 인공지능이 사람의 생명을 다루는 치명적인 작업을 하도록 학습되어있는 경우, 의도적이든 의도적이지 않던 치명적인 오류를 포함하여 학습된 인공지능 시스템은 대량의 환자에게 문제를 유발할 수 있다.

최근 이슈가 되는 딥러닝은 그 성능의 우월성으로 인하여, 기존 기계학습의 재도약을 끌어냈고 기존 알고리즘들에 비해 뛰어난 성능을 보여 많은 기업이 딥러닝 알고리즘을 이용한 제품과 서비스 개발에 뛰어들고 있다. 문제는 딥러닝의 구조가 복잡하다는 것이다. 또한, 딥러닝의 블랙박스 문제가 해결되지 못해 오작동 원인을 파악하기 어렵다는 한계도 있다. 인공지능 기반 제품 및 서비스 이용 시 발생한 사건, 사고와 관련한 책임소재 문제도 의료기기로 사용되는 경우 규제와 관련해 가장 큰 이슈이다. 의료와 같이 인간의 생명을 다루고 의사와 협업하는 분야에서는 심각한 문제이며 인공지능의 현실적 활용 범위에 가장 큰 영향을 줄 수 있는 이슈이기 때문에 면밀한 법 제도와 규제 검토가 필요하다.

2 안정성 평가방법



(a) (b) (c)

그림 6 유방촬영 영상을 이용하여 질환 부분을 컴퓨터보조진단 예시

인공지능이 안전하게 작동하는지를 판단하는 가장 간단한 방법은 여러 가지 상황에서 테스트해보는 것이다. 예를 들어 컴퓨터보조진단(computer aided detection)을 보자면, 다양한 병원에서, 다양한 질환을 가지는 환자를 다양한 영상장비로 촬영한 영상을 이용하여 수많은 상황에서 사고 없이 잘 동작하는지를 확인하는 것이다 (그림 6). 이처럼 다양한 환경에서 작동하여 봄으로써 주어진 소프트웨어의 오류를 파악하는 것을 '소프트웨어 테스팅(Software Testing)'이라고 합니다. 이 테스팅은 다양한 시나리오 하에서 소프트웨어를 작동시켜 보기만 하면 되어 비교적 간단하다는 장점이 있지만, 테스팅을 통해 시험해보지 못한 시나리오에서 오류가 발생할 가능성이 있다는 단점이 있다. 가능한 모든 상황을 고려하지 않는다면 '오류가 절대로 발생할 수 없다'와 같은 보장을 얻기는 불가능한데, 모든 상황을 고려한다는 것은 현실적으로 불가능하다. 이와 같은 문제점을 해결하고 오류가 존재하지 않는다는 '절대적인 보장'을 얻기 위하여 소프트웨어를 논리적으로 분석함으로써 오류가 존재할 수 없음을 엄밀하게 증명하는 방법을 '정형 검증(Formal Verification)'이라고 한다. 정형 검증은 크게 세 단계를 통해 진행됩니다. 먼저 주어진 소프트웨어에 대해 그 작동 기작을 나타내는 수학적 모델을 설립하고, 다음으로는 검증하고자 하는 속성을 논리적 언어로서 명세한다. 마지막으로 소프트웨어의 수학적 모델이 검증하고자 하는 속성을 지닌다는 것을 증명하거나 속성을 위반하는 예시인 반례를 찾아 속성을 지니지 않는다는 것을 보인다. 하지만, 이와 같은 정형 검증 기법은 인공지능 모델에 대해 적용하기는 여러가지 어려움이 있다. 이는 인공지능 모델과 전통적인 소프트웨어의 구조 및 기능적 차이 때문으로, 먼저 인공지능 소프트웨어에 주어지는 입력이 전통적인 소프트웨어에 주어지는 입력에 반해 매우 복잡하다. 실제 인공지능 소프트웨어에서는 음성, 이미지, 텍스트 등 고차원의 입력이 주어진다. 고차원의 입력에 따라 소프트웨어에 무수히 많은 상태가 생겨날 수 있다. 또 다른 어려움은 인공지능 소프트웨어가 작동하는 정확한 원리를 알기 어렵다는 점에서 기인한다. 정형 검증 방법론을 적용하기 위해서는 주어진 소프트웨어를 수학적인 모델로서 표현해야 하는데, 여기에는 그 작동 원리를 파악하고 이에 따라 '압축'해 수학적인 모델을 만들어야 한다. 전통적인 소프트웨어는 한 줄 한 줄 실행되는 프로그램 형태를 띠고 있어 그 압축이 용이한 반면, 인공지능 모델은 복잡한 함수의 집합체로 표현되어 압축하는 것이 매우 어렵다.

안전성을 올바르게 수학적으로 명시하는 것 또한 매우 어렵다. 덧붙여, 여러 가지 서로 다른 조건에서 파생된 안전성을 동시에 만족하게끔 명시하는 것은 더욱 어렵다. 이 모든 어려움을 해결하고 인공지능 소프트웨어의 안전성을 완벽하게 검증했다고 하더라도 실질적인 상황에서 활용 시 문제가 생길 수 있다. '직진 주행 중에는 차선을 벗어나지 않는다'는 안전성이 검증된 자율주행 자동차가 응급 환자를 싣고 급히 달려가는 응급차를 위해 잠시 가던 길을 멈추고 살짝 틀어 길을 양보해주는 것은 쉽지 않다. 인공지능의 안전성 검증은 안전성 그 자체뿐만 아니라 실제 인공지능이 활용되는 상황을 면밀히 고려해야 한다.

Ⅲ 신뢰성/안전성 주요국 대응 동향

1 미국

미국은 국가 차원의 인공지능 육성전략을 담은 '국가 인공지능 연구개발 전략 계획(The national artificial intelligence research and development strategic plan)을 2016년 10월에 발표했다. 이 보고서에는 인공지능 기술의 연착륙을 위한 정부의 역할을 사회적 이익 측면, 기술 표준 측면, 장/단기적 연구지원 측면, 업계와의 협력 측면 등에서 규정하고 총 7개의 전략을 제시했다. 그중 아래와 같이 3개 전략이 인공지능의 안전성 및 인간과의 공생에 관한 것이다.[22]

전략 #2) 인간과의 협력을 위한 효과적인 상호작용방법 연구
전략 #3) 윤리적, 법적, 사회적 함의를 이해하고 이에 맞게 인공지능 시스템을 디자인할 수 있는 방법론 개발
전략 #4) 인공지능이 안전하고 보안문제 없이 운영되도록 보장

트럼프 대통령은 2019년 2월 '인공지능 분야에서 미국의 리더십 유지'라는 행정명령을 발표하고, 여기서 5대 원칙을 제시했다. 이 5대 원칙 중 하나로 인공지능 시스템에 적합한 '기술 및 안전 표준'을 개발하도록 요청하였다. 이에 맞추어, 식품의약국(FDA)의 '인공지능/기계 학습(인공지능/ML)기반 의료기기 소프트웨어(SaMD) 수정을 위한 규제 프레임워크'를 소개하고 있다. FDA 보고서를 따르면, 질병의 치료, 진단, 완화 또는 예방하기 위한 목적의 인공지능/ML 기반 소프트웨어를 의료기기로 정의하고,

22) The White House Fact Sheets(2019.2.11.), President Donald J. Trump Is Accelerating America's Leadership in Artificial Intelligence

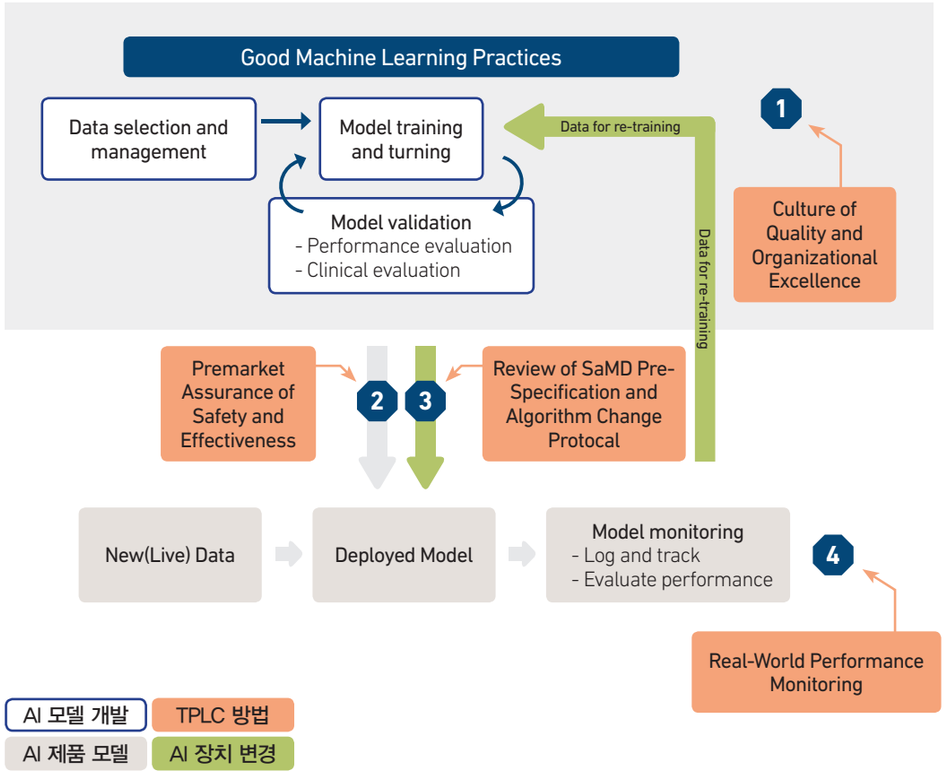


그림 7 인공지능/머신러닝 개발에 관한 FDA의 전체 생명주기 방법 적용[23] (1. 제품 품질에 대한 목표 설정, 2. 안전과 유효성에 대한 시판 전 검증, 3. SaMD와 알고리즘 변경에 대한 검증, 4. 현실 세계에서 검증)

이를 '인공지능/ML기반 SaMD(이하 SaMD)'로 명명했다. 적절한 규제 감독을 통해 안전하고 효과적인 소프트웨어 기능을 제공할 때 SaMD가 의료기기로서 인정한 것이다. 안전한 SaMD 구현을 위해서 소프트웨어의 위험 수준을 정하고, 위험이 일정 수준 이상인 소프트웨어에 대한 사전검증을 시행한다. 이 사전검증은 소프트웨어 개발 전체 주기(Total Product Life Cycle)에서 시행된다. 이런 접근법은 시판 전 개발에서 시판 후 성과에 이르기까지 소프트웨어 제품의 평가 및 모니터링을 하게 하여, 해당 SaMD제품에 대한 지속적인 검증을 하게 한다. 그림 7은 FDA의 적절한 기계학습 개발절차(Good Machine Learning Practices)로 기계학습 기반의 의료 소프트웨어 의료기기를 개발하는 방법을

23) FDA(2019), Proposed Regulatory Framework for Modification to Artificial Intelligence/Machine Learning (AI/ML) based Software as a Medical Device(SaMD)

도식화한 것이다. 이러한 모델은 시판 전 검증과 제품 변경 시 재검증을 통하여 안전성과 유효성을 확보한다. 또한, 시판 후에도 검증을 통하여 안전과 제품 성능을 확보한다.

미국 스탠퍼드 대학교는 인공지능안전센터(Center for AI Safety)를 운영하고 있다. 이 센터는 인공지능 견고성 검증, 안전 중요 자동 시스템에 대한 검증 등의 연구를 수행하고, 인공지능을 적용한 시스템의 안전 확보에 대한 다양한 연구보고서를 발간하고 있다. 또한 정형 검증, 안전한 로봇, 인공지능 안전 세미나 등 인공지능안전 관련 과목들을 이용한 다양한 교육을 하고있다.

2 유럽 연합

유럽연합은 2012년부터 2014년까지 인공지능과 로봇의 사회적 영향을 고려한 로봇법(Robolaw) 프로젝트를 수행하였다. 구체적으로 로봇공학 및 인공지능 기술의 급격한 발전에 비추어 현존하는 법적 기본 틀이 작동 가능한지 여부와 로봇공학 분야의 발전이 규범, 가치 및 사회적 과정에 어떤 방식으로 영향을 미치는지 등을 수술로봇, 자율주행차, 로봇 인공기관 등의 사례를 통해 연구하였다. 이 주요 결과는 2014년 6월 'D6.2 로보틱스 가이드라인(D6.2 Guidelines on Regulating Robotics)으로 공표되었다.[24] 주요 내용을 살펴보면, 인공지능의 안전성을 보장하기 위해 로봇에 전자 인간(electronic personhood)이라는 지위를 부여하고, 인간에게 도움을 주는 목적으로 개발되어야 함을 규정하였다. 또한, 해킹 등 사회적 악용 가능성을 최소화하고, 인간에게 위협을 가하지 않으며, 비상시 인공지능 시스템을 즉시 멈출 수 있는 '킬 스위치' 탑재에 관한 내용도 포함되었다.

한편, 유럽연합집행위원회(EC)에서는 2019년 4월 '신뢰할 수 있는(Trustworthy) 인공지능을 위한 윤리지침'을 발표했다.[25] 신뢰할 수 있는 인공지능을 개발하고 사용하기 위한 7가지 요구사항 중 하나가 기술적 '안전성'이다. 이에 가장 중요한 구성 요소는 기술적인 견고성이다. 의도하지 않은 위험을 최소화하고, 허용 가능한 수준으로 위험을 방지함으로써 인공지능 시스템이 안정적으로 작동되도록 개발하는 것이 기술적 견고성이다. 인공지능 안전은 기술적 안전성·정확성·신뢰성(Reliability)·재현성 등을 구현함으로써 확보할 수 있다. 신뢰성은 정의된 입력과 상황 범위에서 올바르게 작동하는 것을 의미한다. 재현성은 인공지능 실험이 동일한 조건에서 반복될 때 같은 행동을 보이는지 여부를 나타낸다. 대체계획이란 인공 지능 시스템에는 문제가 생기면 대체가 가능한 안전장치를 추가하는 것이다. 이것은 인공지능 시스템이 규칙 기반으로 전환하거나 행동을 계속하기 전에 인간 운영자의 개입을 요구한다는 것이다. 인공지능 안전 확보를 위해서는 모든 생명체와 환경을 해치지 않고 작업을 수행하는지 검토되어야 하며, 의도하지 않은 결과와 오류를 최소화하여야한다. 또한 다양한 응용 분야에 걸쳐 인공지능 시스템의 잠재적인 위험을 규정하고 평가하는 프로세스가 수립되어야 한다. 이러한 대체계획, 기술적 안전성, 정확성, 신뢰성, 재현성은 인공지능 시스템의 위험이 클수록 완성도가 높아야 한다.

민간활동으로는, 영국 옥스퍼드 대학과 FHI(Future of Humanity Institute)에서 공동으로 운영하는 전략 인공지능 연구소(Strategic AI Research Center)을 들수 있다. 여기서는 인공지능이 안전하고 유익한지 확인하기 위한 전략과 도구를 개발하고 있으며, 일반 인공지능(AGI, Artificial General Intelligence)이 자체 보상 기능을 수행할 때 생길수 있는 기능의 부작용에 따른 위험을 연구한 '실수 없는 시도 : 인간 개입을 통한 안전한 강화학습 구현(2017), 존재하는 위험과 희망의 정의(2015)' 등의 안전 관련 보고서를 발간했다.[26] 영국의 앨런튜링연구소(The Alan Turing Institute)는 2018년부터 8개 대학이 참여하여 데이터 과학과 인공지능에 대해 연구하고 있다. 앨런튜링연구소에서 발간한 '인공지능 윤리와 안전의 이해(2019)' 지침서에는 인공지능 시스템으로 인해 발생할 수 있는 잠재적인 위험을 식별하고 방지하기 위한 구체적이고 운영 가능한 조치를 제안하고 있다.[27] 여기서 제안하고 있는 FAST(Fairness, Accountability, Sustainability, and Transparency) 원칙은 인공지능 기술의 견고한 설계와 사용을 위한 원칙이다. FAST 중 "S"인 기술적 지속가능성(Sustainability)의 확보를 위해서는 안전성, 정확성, 신뢰성, 보안, 견고성이 보장되어야 한다는 주장은 EU 보고서와 같은 맥락이다.

3 일본

일본에서는 정부와 학계를 중심으로 인공지능 안전성 확보를 위한 대책들이 마련되고 있다. 2017년 1월 일본 총무성은 인공지능의 안전성과 정보 보안을 위해 공적인증제 도입을 추진하기로 결정하였다. IBM왓슨과 같은 인공지능 제품이나 서비스가 대상이 된다. 사용자 제어 가능 여부, 비상시 기능 정지 및 수정, 사이버 공격에 대한 보안 수준 등을 평가하여 인증 부여 여부를 결정한다. 2017년 2월에는 일본 인공지능학회에서 총 9조로 구성된 윤리지침을 발표하였다. 인간 사회에 가져올 부작용을 최소화하는 데 초점을 맞추고 있으며, 인공지능이 인간처럼 사회구성원으로서 윤리지침을 준수해야 한다는 내용도 포함되어 있다.

24) Guidelines on Regulating Robotics – RoboLaw, https://world.moleg.go.kr → cms → commonDown
25) Ethics Guidelines for Trustworthy Artificial Intelligence, 유럽연합(EU) 집행위원회(EC: European Commission), 2019년 4월 8일

26) William Saunders et al.(2017), Trial without Error: Towards Safe RL with Human Intervention, Owen Cotton-Barratt & Toby Ord(2015), Existential Risk and Existential Hope : Definitions
27) Leslie,D.(2019), Understanding artificial intelligence ethics and safety : A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute

4 대한민국

그림 8 지능정보 기술의 판단과정 (지능정보사회 윤리 가이드라인)[8]

2016년 발표된 '지능정보사회 중장기 종합대책'에는 인공지능을 비롯한 지능정보기술의 연구개발 전략뿐 아니라 산업 혁신, 사회정책 개선 방안 등이 총망라되어 있고, 중장기 정책 방향에서 프라이버시 침해 등에 대한 두려움 없이 국민이 안심하고 지능정보기술을 활용할 수 있는 제도적 기반 확보를 강조하고 있다[28]. 2018년 4월에는 과학기술정보통신부의 지원 아래 학계와 연구계, 민간 전문가 등 25명으로 구성된 정보문화포럼이 2년여간 연구한 결과로 '지능정보사회 윤리 가이드라인'을 발표하였다[29]. 공공성, 책무성, 통제성, 투명성 등 4대 원칙과 기술 개발·활용 단계별 38개 세부 지침을 수립하였다. 정부는 본 가이드라인을 통해 인공지능에 의한 잠재적 위험을 사전에 예방하고 구체적 행위 지침을 통해 관련 분야의 자율 규제 환경 조성에 기여하며, 이용자의 권한을 강화한다.

28) 제4차 산업혁명에 대응한 지능정보사회 중장기 종합대책, 미래과학부, 2016, https://www.korea.kr/archive/expDocView.do?docId=37384
29) 지능정보사회 윤리 가이드라인 - NIA 한국지능정보사회진흥원, 2018

Ⅳ 신뢰성 안전성 평가 가이드라인

의료 인공지능 적용에 대해 '신뢰성 안전성 가이드라인' (Guidelines for Trustworthy AI)는 아래와 같은 원칙과 이를 실현할 수 있는 방안 및 평가기준 이 필요하다.

1. 인간의 기본권 보호와 의료진의 감독

- 인공지능은 인간의 기본 권리와 자율성을 침해해서는 안 되고, 인공지능이 내리는 모든 결정에 대해 의료진이 개입, 감독할 수 있어야 한다.
  - 의료 인공지능의 개발과 적용 시에 인간의 기본권과 자율성을 침해하지 않도록 해야 한다.
  - 의료 인공지능의 결정이 바로 적용되지 않고, 의료진의 감독권을 보장하는 형태로 시스템을 구성해야 한다.

2. 견고함과 안정성

- 의료 인공지능은 안전하고 정확해야 한다. 특히 의료의 불확실성 등에 강인하고, 적대적 외부 공격에 의해 쉽게 손상되지 않아야 하며 합리적으로 신뢰할 수 있어야 한다.
  - 의료 인공지능의 정확도 평가는 external validation 결과가 있어야 하며, prospective validation을 추천한다.
  - 병원이나 진료마다의 스펙트럼 바이어스 등의 특이성을 반영하기 위하여 특정 상황에서 적용하기 전에 정확도를 평가하고, 그 상황이나 병원의 특성을 반영하는 데이터로 fine-tuning을 권장한다.

– 시간에 따른 데이터의 변화를 반영하기 위한 주기적 fine-tuning을 하는 continual learning을 권장한다.
– 의료 인공지능은 적대적 또는 악의적인 공격 등에 최소한의 방어가 되어야 하고, 이에 대한 별도의 테스트 결과를 명시해야 한다.
– 의료 데이터 및 의료 과정의 불확실성을 인지하고, 이에 견고하게 대응해야 한다.

3. 개인 정보 보호 및 데이터 관리

- 의료 인공지능 시스템이 수집한 개인 데이터는 안전하게 개인정보가 보장되어야 하며 권한 없이 쉽게 접근, 강탈되지 않아야 한다.
  – 의료 인공지능이 윤리적으로 개발되었는지에 대하여 명시해야 한다.
  – 개발 과정에 모인 개인 데이터 등에 대한 보안 방법이 명시되어야 한다.

4. 투명성

- 의료 인공지능 시스템의 알고리즘, 데이터에 접근할 수 있어야 하며 의료진은 의료인공지능 시스템이 내리는 결정에 대해 이해하고 설명할 수 있어야 한다.
  – 의료인공지능의 판단에 대해 의료진이나 환자가 이해할 수 있는 추가적인 정보를 제공해야 한다.

5. 다양성, 공정성, 차별 금지

- 의료인공지능이 제공하는 서비스는 연령, 성별, 인종 등 기타 구분에 관계없이 모든 사람에게 사용되거나, 일부에 사용될 경우 이유를 명시해야 한다. 시스템 구성 역시 편향되지 말아야 한다.
  – 의료 인공지능은 특별한 이유 없이 보편적으로 사용될 수 있게 개발되어야 하며, 특정 계층을 차별하지 않게 학습되어야 한다.
  – 의료 인공지능이 일부 계층을 위해 개발될 경우 그 이유와 적용 범위에 대해 명시해야 한다.

6. 의료 발달

- 의료 인공지능 시스템은 긍정적인 의료와 의학의 변화와 지속성을 위해 사용되어야 한다.
  – 의료 인공지능은 의학과 의료의 발달에 사용되어야 한다.

7. 책임

- 의료 인공지능 시스템과 그로 인해 발생하는 결과에 책임져야 하고 시스템 오류, 부정적 결과 등을 인지하고 책임질 수 있는 시스템을 구축해야 한다.
  – 의료인공지능의 오작동 등에 대한 법적 책임 소재 및 보상 방법을 정해야 한다.
  – 의료인공지능을 주기적으로 모니터링 하여 시스템 오류나 부정적 결과를 검출하는 별도의 시스템을 구축한다.
  – 의료인공지능의 예외적 예측을 감지할 수 있는 시스템을 구축한다.

참고문헌

1) Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." science 313.5786 (2006): 504–507.
2) Krizhevsky, Alex. cuda–convnet. https://code.google.com/p/cuda–convnet/, 2012. .
3) Lee, June–Goo, et al. "Deep learning in medical imaging: general overview." Korean journal of radiology 18.4 (2017): 570–584.
4) Jonas, Eric, and Konrad Paul Kording. "Could a neuroscientist understand a microprocessor?." PLoS computational biology 13.1 (2017): e1005268.
5) https://namu.wiki/w/표본조사
6) https://liuziwei7.github.io/projects/LongTail
7) L. Von Bertalanffy, The History and Status of General Systems Theory., Academy of Management Journal. 15 (1972) 407.426.
8) Goldstein, Alex, et al. "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation." Journal of Computational and Graphical Statistics 24.1 (2015): 44–65.
9) https://arxiv.org/abs/1602.04938
10) https://www.ieee802.org/1/files/public/docs2017/tsn–chen–RAP–whitepaper–1117–v02.pdf
11) https://www.ieee802.org/1/files/public/docs2017/tsn–chen–RAP–whitepaper–1117–v02.pdf
12) https://arxiv.org/abs/1704.02685
13) https://arxiv.org/abs/1412.6806
14) https://arxiv.org/abs/1610.02391
15) https://arxiv.org/abs/1704.05796
16) https://arxiv.org/abs/1811.10597
17) https://arxiv.org/abs/1608.05745
18) https://arxiv.org/abs/1903.10176
19) https://arxiv.org/abs/2002.07434
20) http://arxiv.org/abs/1412.6572
21) https://papers.nips.cc/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b–Paper.pdf
22) The White House Fact Sheets(2019.2.11.), President Donald J. Trump Is Accelerating America's Leadership in Artificial Intelligence
23) FDA(2019), Proposed Regulatory Framework for Modification to Artificial Intelligence/Machine Learning (AI/ML) based Software as a Medical Device(SaMD)
24) Guidelines on Regulating Robotics – RoboLaw, https://world.moleg.go.kr → cms → commonDown
25) Ethics Guidelines for Trustworthy Artificial Intelligence, 유럽연합(EU) 집행위원회(EC: European Commission), 2019년 4월 8일
26) William Saunders et al.(2017), Trial without Error: Towards Safe RL with Human Intervention, Owen Cotton–Barratt & Toby Ord(2015), Existential Risk and Existential Hope : Definitions
27) Leslie,D.(2019), Understanding artificial intelligence ethics and safety : A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute
28) 제4차 산업혁명에 대응한 지능정보사회 중장기 종합대책, 미래과학부, 2016, https://www.korea.kr/archive/expDocView.do?docId=37384
29) 지능정보사회 윤리 가이드라인 – NIA 한국지능정보사회진흥원, 2018

KHIDI 한국보건산업진흥원
Korea Health Industry Development Institute

28159 충청북도 청주시 흥덕구
오송읍 오송생명2로 187 오송보건의료행정타운
TEL 043-713-8000~5

본 내용은 연구자의 개인적인 의견이 반영되어 있으며, 한국보건산업진흥원의 공식견해가 아님을 밝혀둡니다.
본 간행물은 보건산업통계포털(http://www.khiss.go.kr)에 게시되며 PDF 파일로 다운로드 가능합니다.