설명 가능한 인공지능(XAI) 동향 및 주요 기법

김대영

Explainable AI (XAI): current status and methods

Daeyoung Kim

December 30, 2021

Abstract

Recently, with the development of deep learning, artificial intelligence has developed so much that it can be used in various industries. However, as the reliance on AI's judgment and decision-making increases, the social demand for explainability and transparency of AI has also increased. As the result, the interest and expectations in the field of explanatory artificial intelligence (XAI) that enables users to understand the judgment results and processes of artificial intelligence systems are also increasing, and many related studies have been conducted. This paper introduces a brief history, research trends, and major methodologies for explainable artificial intelligence.

I. 서론

최근 은행들은 고객들의 신용을 확대할 지나 대출을 승인할 지를 결정하는 데 인공지능(AI)을 사용한다. 의료계에서는 체내의 건강한 조직과 암을 구분하거나 각종 질환을 예측하고 진단하는 데 AI을 활용하고 있다. 또한, 기업 내에서는 AI를 사용해 수천 개의 이력서를 처리하고 인적자원을 관리하거나 경영이나 마케팅과 관련된 중요한 의사결정에 AI를 활용한다.

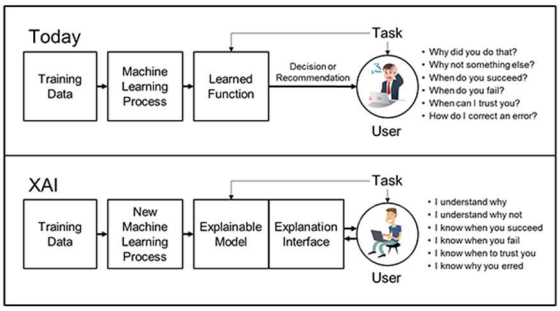
위와 같은 일들은 산업 전반에서 AI가 활용되는 일부 사례에 불과하다. 인공지능 분야는 컴퓨터 비전, 음성 인식, 자연어 처리 등 머신러닝과 딥러닝 기술의 발전으로 다양한 분야에서 탁월한 성능을 보여주고 있으며, 이젠 학술적 연구 단계를 넘어 비즈니스에 적용하는 수준으로 빠르게 발전하며 여러 산업 분야에서의 활용도와 기대가 급격히 높아졌다.

하지만 지금까지의 인공지능은 인지, 의사결정, 예측 등의 정보를 제공할 때 결과만 알려주고 도출한 최종 결과의 근거, 도출 과정의 타당성 등을 논리적으로 설명할 수 없는 점이 한계로 지적되어 왔다.[1]

골드만 삭스(Goldman Sachs) 매니징 디렉터인 찰스 엘칸(Charles Elkan)은 AI에 대한 신뢰도와 AI 시스템에 관한 난관을 극복하는 법에 대해, AI의 현주소를 다음과 같이 표현했다.

"우리는 폭발물 탐지견이 어떻게 임무를 수행하는지 정확히 알 지 못하지만, 탐지견이 내리는 결정을 굳게 신뢰합니다."[2]

머신러닝과 인공지능을 사용하는 기업과 기관의 인공지능에 대한 의존도가 점차 커지면서, 베일에 가려진 인공지능 모델이 어떻게 의사결정을 내리는지 이해할 필요성이 더욱 커졌고,[3] 이에 따라, 인공지능 모델의 의사결정 과정을 사용자가 더욱 잘 이해하고 인공지능 기술의 신뢰성을 높이기 위한[4] '설명 가능한 인공지능' 분야에 대한 관심이 커지게 되었다.



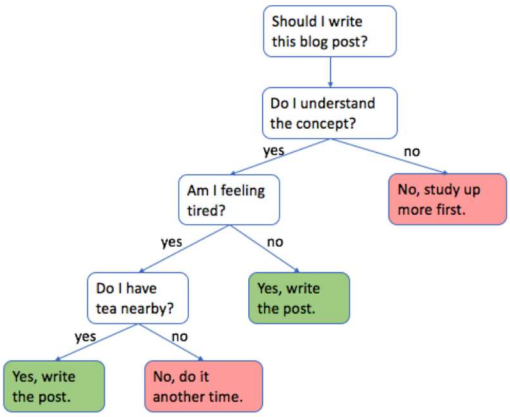
[그림 1] DARPA(Defense Advanced Research Projects Agency)에서 정의한 설명가능 인공지능[5]

[1] 이동진. "설명 가능한 인공지능(EXplainable AI, XAI) 동향." 한국과학기술정보연구원, November 2018. doi:10.22800/KISTI.KOSENEXPERT.2018.49.
[2] NVIDIA Korea, "설명 가능한 AI란 무엇인가?" NVIDIA Blog Korea, July 27, 2021, https://blogs.nvidia.co.kr/2021/07/27/what-is-explainable-ai/.
[3] NVIDIA Korea, "설명 가능한 AI란 무엇인가?"
[4] 정승준, 변준영, 김창익, "설명 가능한 인공지능 기술의 소개," 전자공학회지, 2019, 46(2), 55-63.
[5] M. Turek, "Explainable Artificial Intelligence (XAI)," Defense Advanced Research Projects Agency (DARPA), 2017.

'설명 가능한 인공지능'이란 주어진 데이터에 대해서 분류·예측할 뿐만 아니라 결정에 대한 인과관계를 분석하여 적절한 근거를 찾고 ICT, 심리, 언어 분야 등의 학제 간 융·복합적 기술 개발을 통해 AI 모델의 의사결정 결과를 사용자 레벨에서 설명하는 일련의 AI 기술이다.[6] 인공지능의 설명가능성과 투명성에 대한 커진 필요성과 함께, 설명 가능한 인공지능 기술은 인간과 인공지능 상호간의 신뢰할 수 있는 의사결정을 가능케 할 것으로 기대되고 있다. 이 글에서는 설명 가능한 인공지능의 동향과 대표적인 기법들을 소개하고자 한다.

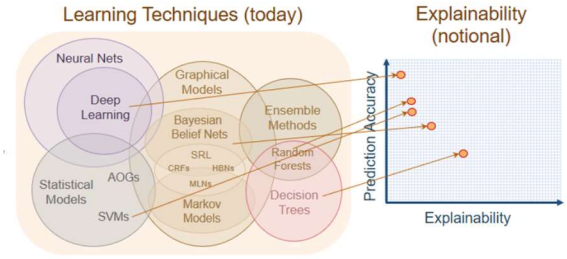
Ⅱ. 역사 및 동향

설명 가능한 인공지능은 새로운 주제는 아니다. 40년 전에 출판된 문헌에서 규칙을 통해 결과를 설명한 것처럼 설명 가능한 인공지능에 대한 초기 작업은 이미 오래 전부터 존재했음을 알 수 있다.[7] AI에 대한 연구가 시작되면서, 과학자들은, 특히 의사결정에 있어서, 지능 시스템이 판단 결과를 설명해야 한다고 주장해 왔다.[8] 예를 들어, 규칙 기반의 지능 시스템이 신용 카드 결제를 거부한 경우, 이러한 부정적인 결정에 대한 이유를 설명해야 한다는 것이었다.



[그림 2] 결정트리 예시. 위에서 아래로 레벨별로 정의된 논리에 따라 결정[9]

규칙 기반의 지능 시스템처럼 기존 머신러닝 모델들의 알고리즘은 덜 복잡했고, 규칙이나 지식이 인간에 의해 정의되고 만들어졌기 때문에 의사결정 과정을 조금이라도 이해하고 예측 할 수 있다.[10] 그래서 초기에는 목표로 하는 인공지능 모델과 비슷한 성능을 유지하면서 의사결정 과정을 보여줄 수 있는 결정 트리나 베이지안 접근법 같은 고전적인 머신러닝 모델을 새롭게 제시하는 방향으로 연구가 진행되었다.[11]



[그림 3] 모델의 예측 성능이 좋을수록 설명가능성(explainability)은 낮다.[12]

이후 알고리즘 복잡성이 높은 딥러닝 모델들이 기존 머신러닝 모델들보다 성능에 있어 더 우위를 점하게 되면서,[13] 딥러닝 모델을 이해하기 위한 노력이 많아지게 되었다. 최근에는 설명 가능한 딥러닝 모델을 구축하는 것에 대한 연구와 모델의 의사결정 과정에 대한 설명을 사용자가 이해할 수 있도록 표현하는 것에 대한 연구가 주를 이루고 있다.[14]

Ⅲ. 주요 방법론

설명 가능한 인공지능이 여러 접근법을 통해 연구되어 왔듯이, 설명 가능한 인공지능과 관련된 여러 기법이 존재한다.

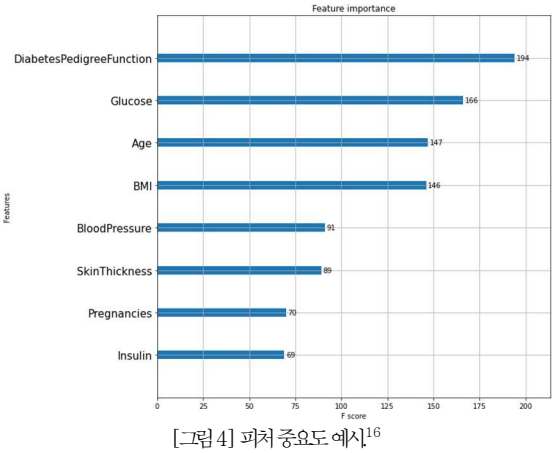
방법론	장점	단점
피처 중요도	- 피처 가중치에 기반한 설명을 제공 - 쉽고 직관적	- 설명은 로컬 피처에 대해서만 제한됨 - 스케일에 따라 모델에 미치는 영향을 파악하기 어려움 - 피처 간 의존성이 큰 경우 신뢰하기 어려움
LIME	- 모델과 무관하게 적용 가능하며, 입력과 출력만으로 확인이 가능 - 특정 샘플에 대해 설명이 쉬워 실무에 적합 - 다른 설명 가능한 인공지능 기법과 비교하여 가벼움	- 설명 단위가 그때그때 달라지고 비결정적이기 때문에, 입력 값이 같아도 출력 결과가 일정하지 않음 - 입력과 출력만을 간접적으로 설명할 뿐 인공지능 모델에 대해 설명하지 않으므로 모델의 본질을 설명할 수 없음
LRP	- 비교적 직관적	- 기여도를 히트맵으로 표현하는 것으로는 신경망 모델이 학습한 추상적인 개념을 알 수 없음

[6] 한지연, 최재식, "설명가능 인공지능," 소음진동 27, no. 6 (2017): pp. 8-13.
[7] A. Carlisle Scott, Explanation Capabilities of Production-Based Consultation Systems (Stanford: Depts. of Computer Science and Clinical Pharmacology, Stanford University, 1977).
[8] Feiyu Xu et al., "Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges," Natural Language Processing and Chinese Computing, 2019, pp. 563-574, https://doi.org/10.1007/978-3-030-32236-6_51.
[9] Jeremy Jordan, "Decision Trees.," Jeremy Jordan, May 19, 2019, https://www.jeremyjordan.me/decision-trees/.
[10] Feiyu Xu et al., "Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges."
[11] 정승준, 변준영, 김창익, "설명 가능한 인공지능 기술의 소개," 전자공학회지 46, no. 2 (2019): pp. 55-63.
[12] M. Turek, "Explainable Artificial Intelligence (XAI)," Defense Advanced Research Projects Agency (DARPA), 2017.
[13] 정승준, 변준영, 김창익, "설명 가능한 인공지능 기술의 소개," 전자공학회지 46, no. 2 (2019): pp. 55-63.
[14] 최재식, "설명가능 인공지능의 연구 동향," 전자공학회지 48, no. 9 (September 2021): pp. 16-22.

		- 은닉층 내부의 기여도를 확인할 수 있어, 해당 은닉층이 무엇을 감지했는지 알 수 있음
SHAP	- 모델에 무관하게 적용 가능 - 속도 향상에 최적화	- 최적의 설명 크기를 결정하지 않음 - 피처 종속성을 고려하지 않음 - 하나의 예측에 대해서만 관련됨

[테이블 1] 주요 방법론 장점 및 단점

첫번째는, 데이터의 피처가 알고리즘의 정확한 분류에 얼마나 큰 영향을 미치는지 분석하는 기법인 피처 중요도(Feature Importance, Permutation Importance)이다. 특정 피처의 값을 임의의 값으로 치환했을 때 원본보다 에러가 얼마나 더 커지는지를 판단하여 중요도를 측정한다. 특정 피처를 변형했을 때 모델의 예측 결과가 크게 달라졌다면, 모델은 이 피처에 대한 의존도가 높고, 이 피처는 중요도가 높다고 볼 수 있다. 하지만 스케일에 따라 모델에 미치는 영향을 파악하기 어려우며, 피처 간 의존성이 큰 경우 신뢰하기 어려운 부분이 있다.[15]

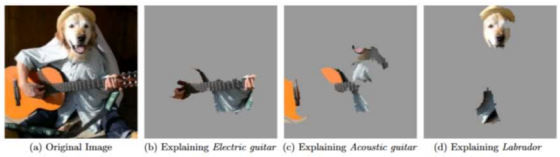


[그림 4] 피처 중요도 예시.[16]

계산 속도는 느리지만 피처 중요도 기법의 단점을 보완해 주는 부분 의존성 플롯(Partial Dependence Function)이 사용되기도 하는데, 특정 피처의 값을 선형적으로 변경하면서 알고리즘 해석 능력이 얼마나 증가하고 감소하는지를 관찰 할 수 있다.[17]

두번째는, Local Interpretable Model-agnostic Explanations (LIME)으로 데이터 하나에 초점을 맞추어 블랙박스 모델이 해석하는 과정을 분석하는 로컬 대리 분석(local surrogate)의 대표적인 기법이다. LIME은 임의의 블랙박스 모델을 이미 설명이 가능한 데이터 주변에서 희소 선형 결합을 통해[18] 현재 데이터의 어떤 영역을 집중해서 분석했고, 어떤 영역을 분류 근거로 사용했는지 국부적으로 설명가능하게 한다.[19] 예를 들어, 이미지를 분류하는 블랙박스 모델이 어떤 이미지를 개라고 판단했다면 이미 설명 가능한 다른 모델의 개에 대한 설명, 즉 개를 표현하는 픽셀들을 주어진 이미지와 대조하여 어느 부분이 개라고 판단한 근거인지 제시할 수 있다.[20]

LIME은 분류에 사용한 모델에 관계없이 적용할 수 있고 다른 설명 가능한 인공지능 모델에 비해 가볍다는 장점이 있다. 하지만 슈퍼 픽셀(특정 이미지의 관심 영역을 주변으로 영역을 확대해 가며 동일한 정보를 가진다고 간주할 수 있다고 찾은 영역을 의미) 알고리즘에 따라 마스킹 데이터가 달라져 랜덤한 결과를 보일 수 있으며, 데이터 하나에 대해 설명을 하기 때문에 모델 전체에 대한 일관성을 보전하지 못해 전체 데이터 신뢰도를 떨어뜨릴 수 있다.[21]

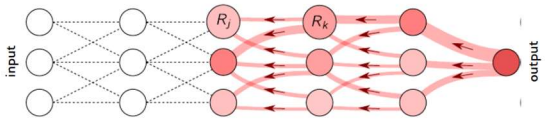


[그림 5] LIME 기반 이미지 분류예측 설명 예시. 원본 이미지, '전자 기타의 LIME, '어쿠스틱 기타의 LIME, 'Labrador'의 LIME.[22]

Layer-wise Relevance Propagation (LRP)은 딥러닝 모델에서 예측 결과로부터 분해(Decomposition)와 타당성 전파(Relevance Propagation)를 사용하여 역전파 형태로 신경망의 각 계층별 기여도를 측정할 수 있는 방법으로, 각 계층의 기여도를 히트맵 형태로 시각화하여 어느 부분이 판단에 영향을 미쳤는지를 확인할 수 있는 방법이다.[23] 기여도 계산 방법은 테일러 급수(Taylor series)를 응용한 심층 테일러 분해(Deep Taylor decomposition)를 지표로 삼았는데, 테일러 급수는 잘 모르거나 복잡한 함수를 다루기 쉽고 이해하기 쉽게 바꾸기 위해 쓰이는 방법이다. 예를 들어, 특정 신경망 모델이 왜 이미지를 고양이로 판단했는지 확인하고자 할 때, 특정 은닉층의 기여도를 토대로 표현한 히트맵을 확인하다 시각적으로 고양이 형상으로 보이는 히트맵이

[15] 안재현, XAI 설명 가능한 인공지능, 인공지능을 해부하다 (위키북스, 2020).
[16] 안재현, XAI 설명 가능한 인공지능, 인공지능을 해부하다 (위키북스, 2020).
[17] 안재현, XAI 설명 가능한 인공지능, 인공지능을 해부하다 (위키북스, 2020).
[18] 최재식, "설명가능 인공지능의 연구 동향," 전자공학회지 48, no. 9 (September 2021): pp. 16-22.
[19] 안재현, XAI 설명 가능한 인공지능, 인공지능을 해부하다 (위키북스, 2020).
[20] 최재식, "설명가능 인공지능의 연구 동향," 전자공학회지 48, no. 9 (September 2021): pp. 16-22.
[21] 안재현, XAI 설명 가능한 인공지능, 인공지능을 해부하다 (위키북스, 2020).
[22] Marco Ribeiro, Sameer Singh, and Carlos Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, 2016, https://doi.org/10.18653/v1/n16-3020.
[23] 이동진. "설명 가능한 인공지능(EXplainable AI, XAI) 동향." 한국과학기술정보연구원, November 2018. doi:10.22800/KISTI.KOSENEXPERT.2018.49.

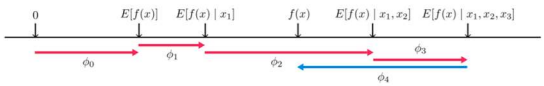
있다면 이러한 픽셀이 고양이로 분류하게끔 기여했다고 해석한다.[24] LRP는 특히 질병 진단 등에 유용하게 쓰일 수 있는데, MR과 같은 의료 영상을 보고 질병을 진단할 때 어떤 부분이 그 질병으로 진단하는 데에 근거가 되었는지 의료 영상에 표시해 진단 의사가 참고할 수 있도록 하여 진단의 효율을 높일 수 있다.[25]



[그림 6] LRP 과정. 각 뉴런은 상위 계층에서 받은 만큼 하위 계층에 재분배[26].

비교적 직관적이고 은닉층 내부의 기여도를 확인 할 수 있어 해당 은닉층이 무엇을 감지했는지 확인 할 수 있지만, 기여도를 히트맵으로 표현하는 것으로는 신경망 모델이 학습한 추상적인 개념을 해석하기 어렵다는 단점이 존재한다.[27]

마지막으로, Shapley Additive Explanations (SHAP)는 게임 이론에 기반한 모델로, 섀플리(Shapley) 값과 피처 간 독립성을 활용하여 전체 성과를 창출하는데 각 피처가 얼마나 공헌했는지 파악한다.[28] 피처들을 추가 및 제거하는 데이터셋을 만들어 이를 선형 모델로 구성하고 이렇게 구성된 선형 모델의 가중치를 가지고 해석하는 방식으로, 특정 변수가 제거되면 얼마나 예측에 변화를 주는지 살펴보고, 그에 대한 답을 SHAP 값으로 표현하는데 이때 SHAP 값은 한 예측에서 변수의 영향도를 방향과 크기로 표현한다.[29]



[그림 7] SHAP 설명에 대한 예시. 오른쪽 화살표(파란색)은 원점으로부터 모델이 높은 예측 결과를 낼 수 있게 도움을 주는 특성이며, 왼쪽 화살표(빨간색)은 모델의 예측에 방해가 되는 요소[30]

LIME과 비슷하게, SHAP 또한 최적의 설명 크기를 결정하지 않고, 피처 종속성을 고려하지 않으며, 하나의 예측에 대해서만 표현된다는 단점이 존재하며,[31] 새로운 아웃라이어 데이터에 취약하다.[32]

Ⅳ. 결론

최근 설명 가능 인공지능은 딥러닝 내부의 노드를 분석하고 그 오류를 수정하는 할 수 있고, 또한 복잡한 모델의 상관관계를 분석하여 원인요소와 결과요소로 나누어 설명할 수 있을 정도로 발전하였다.[33] 광범위한 분야에서 인공지능이 활용됨에 따라, 인간과 인공지능 간의 상호작용을 보다 효율적이고 투명하고, 또 편리하게 하는 기법에 대한 관심이 증가하고 있으며, 그 일환으로 설명 가능한 인공지능 기술에 대한 연구가 더욱 활성화되고 있다.[34] AI 설명가능성의 발전을 통해 인간과 인공지능 간의 신뢰성과 보다 밀접한 상호작용으로 금융, 보험, 군사 등 다양한 분야가 더욱 발전할 수 있게 되리라 기대한다.

---

[24] 최형규, “설명가능한 인공지능(Explainable AI; XAI) 연구 동향과 시사점 학습이 완료된 딥러닝 모델에 대한 설명을 중심으로,” SPRi 소프트웨어정책연구소, September 13, 2021, https://spri.kr/posts/view/23296?code=industry_trend.
[25] 이동진. “설명 가능한 인공지능(EXplainable AI, XAI) 동향.” 한국과학기술정보연구원, November 2018. doi:10.22800/KISTI.KOSENEXPERT.2018.49.
[26] Grégoire Montavon et al., “Layer-Wise Relevance Propagation: An Overview,” Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, 2019, pp. 193-209, https://doi.org/10.1007/978-3-030-28954-6_10.
[27] 최형규, “설명가능한 인공지능(Explainable AI; XAI) 연구 동향과 시사점 학습이 완료된 딥러닝 모델에 대한 설명을 중심으로,” SPRi 소프트웨어정책연구소, September 13, 2021, https://spri.kr/posts/view/23296?code=industry_trend.
[28] 안재현, XAI 설명 가능한 인공지능, 인공지능을 해부하다 (위키북스, 2020).
[29] 김홍비, 이태진, “정보보호 분야의 XAI 기술 동향,” 정보보호학회지 31, no. 5 (October 2021).
[30] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee, Consistent Individualized Feature Attribution for Tree, 2019, https://doi.org/https://arxiv.org/pdf/1802.03888v3.
[31] Prashant Gohel, Priyanka Singh, and Manoranjan Mohanty, “Explainable AI: Current Status and Future Directions,” July 2021, https://doi.org/10.1109/ACCESS.2017.DOI.
[32] 안재현, XAI 설명 가능한 인공지능, 인공지능을 해부하다 (위키북스, 2020).
[33] 최재식, “설명가능 인공지능의 연구 동향,” 전자공학회지 48, no. 9 (September 2021): pp. 16-22.
[34] 이동진. “설명 가능한 인공지능(EXplainable AI, XAI) 동향.” 한국과학기술정보연구원, November 2018. doi:10.22800/KISTI.KOSENEXPERT.2018.49.

References

NVIDIA Korea. “설명 가능한 AI란 무엇인가?” NVIDIA Blog Korea, July 27, 2021. https://blogs.nvidia.co.kr/2021/07/27/what-is-explainable-ai/.

M. Turek, “Explainable Artificial Intelligence (XAI),” Defense Advanced Research Projects Agency (DARPA), 2017.

한지연, 최재식. “설명가능 인공지능.” 소음진동 27, no. 6 (2017): 8-13.

Scott, A. Carlisle. Explanation Capabilities of Production-Based Consultation Systems. Stanford: Depts. of Computer Science and Clinical Pharmacology, Stanford University, 1977.

Xu, Feiyu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. “Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges.” Natural Language Processing and Chinese Computing, 2019, 563–74. https://doi.org/10.1007/978-3-030-32236-6_51.

정승준, 변준영, 김창익. “설명 가능한 인공지능 기술의 소개.” 전자공학회지 46, no. 2 (2019): 55-63.

최재식. “설명가능 인공지능의 연구 동향.” 전자공학회지 48, no. 9 (September 2021): 16–22.

안재현. XAI 설명 가능한 인공지능, 인공지능을 해부하다. 위키북스, 2020.

이동진. “설명 가능한 인공지능(EXplainable AI, XAI) 동향.” 한국과학기술정보연구원, November 2018. https://doi.org/10.22800/KISTI.KOSENEXPERT.2018.49.

최형규. “설명가능한 인공지능(Explainable AI; XAI) 연구 동향과 시사점 학습이 완료된 딥러닝 모델에 대한 설명을 중심으로.” SPRi 소프트웨어정책연구소, September 13, 2021. https://spri.kr/posts/view/23296?code=industry_trend.

김홍비, 이태진. “정보보호 분야의 XAI 기술 동향.” 정보보호학회지 31, no. 5 (October 2021).

Gohel, Prashant, Priyanka Singh, and Manoranjan Mohanty. “Explainable AI: Current Status and Future Directions,” July 2021. https://doi.org/10.1109/ACCESS.2017.DOI.

Ribeiro, Marco, Sameer Singh, and Carlos Guestrin. “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier.” Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, 2016. https://doi.org/10.18653/v1/n16-3020.

Jordan, Jeremy. “Decision Trees.” Jeremy Jordan, May 19, 2019. https://www.jeremyjordan.me/decision-trees/.

Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. Consistent Individualized Feature Attribution for Tree, 2019. https://doi.org/https://arxiv.org/pdf/1802.03888v3.

Montavon, Grégoire, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. “Layer-Wise Relevance Propagation: An Overview.” Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, 2019, 193–209. https://doi.org/10.1007/978-3-030-28954-6_10.