환자 중심의 의료 **XAI** 시스템 설계 및 검증

김유영O, 김민정, 김새별, 김진우
주식회사 하이, 연세대학교 HCI Lab
kimmmyoung90@gmail.com, minjungkim618@gmail.com, kyoakehoshi@gmail.com,
jinwoo@haii.io

Advancing Healthcare AI: Patient-Centered Design and Validation of XAI Systems

Yuyoung KimO, Minjung Kim, Saebyeol Kim, Jinwoo Kim
Haii corp., Yonsei University HCI Lab

요 약

최근 의료 인공지능(AI)의 발전은 설명 가능한 인공지능(XAI)의 중요성을 강조한다. 본 연구는 환자의 요구에 맞춘 XAI 시스템 설계의 필요성을 강조하며, 다양한 설명 형식과 철저한 평가 방법을 제안한다. XAI 설계에 환자 관점을 반영함으로써, 본 연구는 의료 AI 관행의 격차를 해소하고 환자의 참여와 AI 시스템에 대한 신뢰를 향상시키는 것을 목표로 한다. 특히 실험을 통해 전통적인 임상 설명과 환자를 위해 설계된 XAI 인터페이스의 설명성을 비교했다. 결과적으로 XAI 시스템을 사용할 때 환자의 만족도, 신뢰도 및 이해도가 크게 향상됨을 확인하였다.

1. 서 론[1]

최근 질병의 진단, 환자 관리, 의료 데이터 해석 등 다양한 분야에 의료 인공지능(AI)이 활용되고 있으며 [1, 2], 이러한 시스템은 의료 품질과 비용 효율성을 향상시키는데 도움이 되고 있다 [3]. 그러나 의료 데이터의 민감성, 의료 AI의 오용 등은 환자에게 해를 끼칠 수 있기 때문에 신중한 접근이 필요하다 [4].

이러한 문제를 해결하기 위한 한 가지 방법은 해석 가능한 AI 모델을 채택하는 것이다 [5]. 설명가능한 인공지능(eXplainable AI, XAI)는 AI 시스템의 의사 결정 과정을 인간이 이해할 수 있도록 하여 사용자 신뢰를 높이고 잠재적 오류나 편향을 해결하는 데 도움을 준다 [6, 7]. 또한, XAI는 AI 모델의 예측을 이해하고 디버깅하며, 추가적인 의료 인사이트를 제공할 수 있도록 돕는다 [8, 9].

최근 XAI 연구는 사용자 요구를 더 잘 이해하고 투명성을 개선하는 방향으로 나아가고 있다 [10, 11, 12]. 특히 의료 AI에서는 모델 개발자, 연구자, 의료 전문가, 환자 등 다양한 이해 관계자의 요구를 충족시키는 것이 중요하다 [13, 14, 15]. 예를 들어, 임상의는 환자의 진단, 치료 예후 정보를 원하므로, AI 시스템의 기본 논리와 기능을 이해하려고 할 수 있으며, 환자는 본인의 상태에 대한 이해를 높이기 위해 예측의 이유와 예후를 알고자 한다. 이러한 다양성은 XAI에서 사용자 중심 접근의 필요성을 강조한다 [16, 17]. 그러나 의료 AI은 역사적으로 환자의 관점을 간과해왔다 [18].

본 연구는 주로 환자를 위한 XAI 설계 및 검증에 중점을 둔다. AI 시스템이 전통적인 임상 환경에서 제공하는 설명과 동일한 수준의 만족도와 신뢰를 제공할 수 있는지 평가하고, 환자가 더 나은 자기 관리와 정보에 입각한 결정을 내리도록 돕고자 한다. 우리의 연구는 환자 중심의 XAI 설계를 통해 환자의 참여와 신뢰를 높이는 것을 목표로 한다. 또한 디자이너와 연구자에게 의료 분야에서 환자를 위한 XAI 채택을 향상시키기 위한 제안과 지침을 제공하고자 한다 [19, 20, 21].

2. 환자를 위한 **XAI** 설계

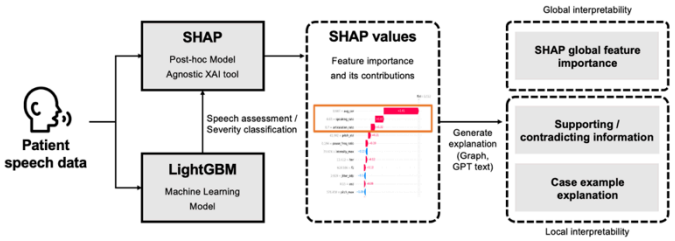
본 연구는 뇌졸중 후 말 언어 장애를 겪는 환자를 위해 설계된 XAI 시스템을 제시한다. 뇌졸중 후 말 언어 장애는 의사소통의 능력을 저하시키고 환자의 삶의 질을 떨어트리는 주요 장애 중 하나이다 [26]. 표준화된 평가 프로토콜에도 불구하고 [27], 언어 병리학자(SLP)의 전통적인 평가 방법은 종종 신뢰성 문제에 직면한다 [28]. 임상 환경에서 AI를 이용한 언어 평가의 사용이 증가하고 있지만, AI 기술의 복잡성은 특정 임상 결정의 기초를 이해하는 데 어려움을 줄 수 있다 [29]. 그럼에도 불구하고, 해석 가능한 모델은 환자 관리를 향상시키기 위한 실용적인 인사이트를 제공할 수 있기 때문에 최근 이를 접목한 사례가 늘고 있다 [30].

우리의 XAI 시스템은 기존 문헌을 바탕으로 구축되었다 [14, 24, 25, 31]. 그림 1은 XAI 시스템의 아키텍처를 보여준다. 머신러닝 모델(LightGBM)이 생성한 뇌졸중 후 말 언어장애

[1]* 이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2022-0-00621,대화 기반 설명가능성을 멀티모달로 제공하는 인공지능 기술 개발)

중증도 분류 결과는 SHAP(Shapley Additive exPlanations) 프레임워크를 사용해 해석된다. SHAP 값을 사용함으로써, 중증도 분류 결과에 대한 각 특징의 기여도를 확인할 수 있으며, 이를 통해 사용자는 말 언어 오류에 기여하는 가장 중요한 요소와 상대적 중요성을 이해할 수 있다[32, 33].

먼저, 우리는 뇌졸중 후 언어 장애의 심각도를 정량화하고, 색상 코딩을 사용해 결과에 긍정적 또는 부정적으로



기여하는 특징을 구분했다 [34] (그림 2A). 예를 들어, 빨간색은 언어에 부정적인 영향을 미치고, 파란색은 긍정적인 영향을 미친다. 다음으로, 인터페이스는 각 특징을 언어 하위 시스템(호흡, 발성, 공명, 조음, 운율)로 나눈다. 이 기능은 각 영역의 상대적 기여도를 반영하고 중요한 문제를 강조하는 데 필수적이다 (그림 2B). 마지막으로, 사례 예시를 통해 왜 이러한 분석이 결정되었는지에 대한 자세한 설명을 제공한다. 예를 들어, 음량 문제는 정상 범위와 비교하여 시각화 되고 설명된다 (그림 2C). 특징 기반 접근과 예시 기반 접근을 연결하여 각 특징의 중요성을 유사한 예시에서 설명한다 [35]. 사용자는 음량을 변경했을 때 심각도 점수가 어떻게 변하는지 시뮬레이션 할 수 있으며, 이는 증상과 심각도를 업데이트해 사용자가 이러한 특징이 진단에 미치는 영향을 이해하도록 돕는다 [24].

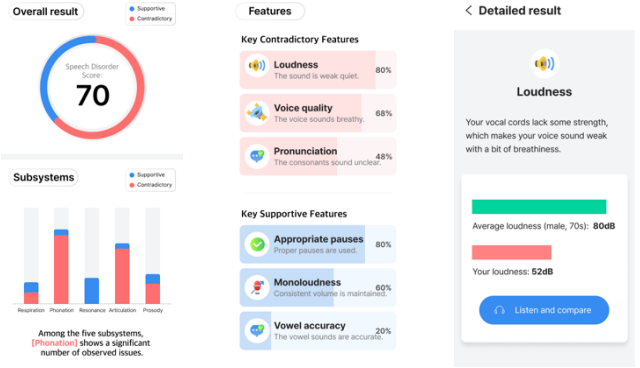
그림 2. 뇌졸중 후 말 언어장애 환자를 위한 XAI 인터페이스

3. 연구 방법

뇌졸중 후 말 언어장애 환자를 대상으로 전통적인 임상 설명과 XAI 인터페이스의 설명력을 비교했다. 실험 설계는 언어재활사의 설명과 XAI의 설명 두 가지 셀을 가진 피험자 내 설계로 이루어졌다. 순서 효과를 방지하기 위해 참가자들은 설명 모드의 제시 순서를 무작위로 다른 그룹에 할당했다. 전통적인 임상 설명은 언어치료사에게 샘플 환자의 말 언어 장애 중증도, 문제 요인, 장점을 평가하도록 하여 환자에게 텍스트와 보이스 설명으로 제시 했다.
우리는 언어재활사의 말 평가 결과 설명과 XAI 인터페이스에 대한 설명 만족도와 신뢰도를 평가하기 위해 두 가지 설문조사를 사용했다. "설명 만족도 척도"는 5점 리커트 척도(1 = "전혀 그렇지 않다" ~ 5 = "매우 그렇다")로 구성된 8개의 항목으로, 이해 가능성, 만족도, 세부 사항의 충분성, 완전성, 유용성, 정확성 및 신뢰성을 측정했다 [23]. "설명 신뢰도 척도" 역시 5점 리커트 척도로 구성된 8개의 항목으로, XAI 시스템의 예측 가능성, 신뢰성, 효율성 및 신뢰성을 평가했다 [23]. 이 척도는 이전 신뢰 연구 [36]에서 항목을 수정하여 XAI 설명 만족도 척도의 리커트 형식에 맞춘 것을 사용하였다. 추가적으로 XAI 설명에 대한 참가자들의 의견을 구하기 위해 반 구조화된 인터뷰를 실시했다. 인터뷰에서 얻은 응답을 주제별로 코딩하여 범주를 생성했다.

4. 연구 결과

뇌졸중 후 말 뇌졸중 후 말 언어장애를 가진 6명의 남성 환자가 실험에 참여했다. 이들의 평균 나이는 49.5세였으며, 나이는 23세에서 77세까지 다양했다. 두 명의 참가자(P1, P3)는 뇌졸중 발생 후 한 달 이내의 급성기 환자였으며, 나머지 네 명은 발생 후 6개월이 지난 만성기 환자였다. P2와 P5는 경미한 인지 장애가 있어 의사소통에 어려움이 있었으며, 이를 보조하기 위해 보호자가 실험에 참석했다. 통계 분석은 SPSS 29 버전을 사용하여 Wilcoxon Signed-Rank Test로 분석했다.
설명 만족도 변수를 분석한 결과, 모든 참가자들은 언어재활사의 설명 (M=27.83, SD=5.91)보다 XAI 설명 (M=36.50, SD=3.02)에 더 높은 만족도를 보였다 (Z=-2.023, p=.043). 이는 XAI 시스템이 환자 이해와 참여를 향상시킬 수 있다는 잠재력을 강조한다.
신뢰도 변수에서는 통계적 유의성을 찾지는 못하였으나 (Z=-1.625, p=.104), P3를 제외한 모든 참가자가 임상 설명(M=32.67, SD=3.83)보다 XAI 설명에 대해 더 높은 신뢰도를 보고했다(M=36.00, SD=4.40).
대부분의 참가자는 두 설명 모두 이해하기 쉽고 도움이 되며 신뢰할 수 있다고 평가했다. 그러나 많은 참가자들이 XAI 설명에 대해 더 긍정적인 피드백을 남겼다.
건강 정보에 대한 접근 용이성: 뇌졸중 후 환자들은 복잡한 정보를 한꺼번에 이해하는 데 어려움을 겪는다. XAI 설명은 시각적 도구와 사례 기반 설명을 사용하여 주요 문제를 이해하기 쉽게 만들었다.
환자들이 원하는 설명의 차이: 급성기 환자들은 자신의 강점을 인정하는 설명을 선호했다. 이는 치료 동기를 촉진하는 데 도움이 되었다. 반면, 만성기 환자들은 예후와



(A) 전체 결과 (B) 특징 중요도 (C) 사례 예시

개선 가능성에 더 관심을 가졌다.
의료 설명에 대한 비판적 수용 부족: 참가자들은 전문가와 XAI 시스템의 설명을 무비판적으로 수용하는 경향이 있었다. 이는 의료 전문가에 대한 과도한 신뢰를 나타내며, 설명의 정확성이나 관련성을 충분히 이해하지 못하는 경우가 많았다.

5. 논의

AI 시스템이 의료 분야에서 중요한 결정을 내리는 데 점점 더 많이 사용됨에 따라, XAI의 중요성이 커지고 있다 [6, 7]. XAI는 시스템의 행동과 과정을 인간이 이해할 수 있는 방식으로 설명하는 데 중점을 둔다. 이는 많은 의료

시스템을 배포할 때의 규제 요구 사항과도 일치한다 [37, 38]. 본 연구는 뇌졸중 후 언어 장애 환자를 위한 언어 평가 인터페이스를 설계하는 것 이상의 의미를 가진다. 우리는 의료 XAI에 대한 보다 광범위한 함의를 제시한다.

1. 사용자 이해를 위한 설명 설계**:** XAI 설명은 사용자의 전문 지식 수준과 특정 요구에 맞추어야 한다. 특히 뇌졸중 환자와 같은 인지적 제한이 있는 사용자에게는 AI 추론과 의료 개념에 대한 쉽게 이해할 수 있는 설명이 필요하다 [22, 40, 41]. 설명의 제시 방식은 그 명확성과 해석 가능성에 큰 영향을 미친다 [42]. 고수준 정보 [39], 적절한 용어 [43], 인식 가능한 패턴을 사용하면 더 이해하기 쉬운 설명을 설계할 수 있다. 이를 위해, 대형 언어 모델 (LLM)을 사용하여 정보의 복잡성을 조정함으로써 환자의 수준에 맞춘 설명을 제공할 수 있다.

2. 설명에 다양한 형태와 방식 사용**:** 환자들은 자신의 상태를 이해하고, 치료 동기를 유지하며, 예후를 개선하기 위한 행동 변화를 위해 다양한 형태의 설명을 필요로 한다 [14, 15]. 따라서 설명은 한 가지 형태나 스타일로 제한되어서는 안 된다. 설명 생성 메커니즘, 유형, 범위 또는 이러한 특징의 조합에 기반한 다양한 설명 방식이 제안되었다 [9, 44]. 특징 기반 설명은 예측에 기여한 환자의 특성을 식별할 수 있으며, 사례 기반 방법은 유사한 환자의 결과를 보여줄 수 있다 [16]. 반 사실 분석은 환자가 예후를 개선하기 위해 어떤 변화를 해야 하는지 이해하는 데 도움을 준다 [16]. XAI를 설계할 때, 그래프, 설명 텍스트, 음성 상호작용 등 다양한 증거를 사용하는 환자의 다양한 요구를 고려하는 것이 중요하다 [20]. 설명의 설계 공간을 확장하면 다양한 환자 요구에 맞춘 보다 효과적인 XAI 방법을 개발할 수 있다.

3. 설명의 철저한 평가**:** XAI 시스템의 설명이 환자에게 과도한 신뢰를 주지 않도록 하는 것이 중요하다 [45]. AI 기술의 능력과 한계에 대해 환자에게 가이드를 주어 AI 시스템의 오용이나 과도한 의존을 피해야 한다 [46]. 설명의 정확성이 예측 모델의 정확성과 항상 일치하지 않을 수 있다. 안전이 중요한 영역에서는 AI 결정에 대한 회의적인 접근이 맹목적인 신뢰보다 더 적절할 수 있다 [47]. XAI 방법의 정확성과 신뢰성을 평가하는 것은 개발 과정에서 중요하다. 이전 연구는 해석 가능성을 평가하기 위한 응용 기반, 인간 기반, 기능 기반 평가를 포함한 3단계 접근 방식을 제안했다 [22]. 이러한 평가 방법들은 다양한 목적에 맞추어져 있으며 다양한 상황에 적용된다. 이 단계적 접근 방식은 AI 모델의 정확성과 설명의 질을 지속적으로 검증하는 중요성을 강조한다. 이 반복적인 검증 과정은 의료 분야에서 AI 시스템의 신뢰성과 효과를 보장하는 데 필수적이다 [48].

6. 결론

본 연구는 뇌졸중 후 언어 장애 환자를 위해 XAI 인터페이스를 설계 하고, 그 효과를 평가했다. XAI는 환자의 설명 만족도와 신뢰도를 크게 향상시킬 잠재력을 가지고 있다. 그러나 참여자 수의 제한과 단일 설명 인터페이스 사용은 연구 결과의 일반화에 한계를 두었다. 향후 연구는 더 다양한 환자 그룹을 포함하고 다양한 설명 인터페이스를 탐구해야 한다.

7. 참고문헌

1. Dreyer, K. and B. Allen, Artificial Intelligence in Health Care: Brave New World or Golden Opportunity? Journal of the American College of Radiology, 2018. 15(4): p. 655-657.
2. Topol, E.J., High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 2019. 25(1): p. 44-56.
3. Higgins, D. and V.I. Madai, From Bit to Bedside: A Practical Framework for Artificial Intelligence Product Development in Healthcare. Advanced Intelligent Systems, 2020. 2(10): p. 2000052.
4. Volovici, V., et al., Steps to avoid overuse and misuse of machine learning in clinical research. Nature Medicine, 2022. 28(10): p. 1996-1999.
5. Rudin, C., Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 2019. 1(5): p. 206-215.
6. Barredo Arrieta, A., et al., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 2020. 58: p. 82-115.
7. Ehsan, U., et al., Automated rationale generation: a technique for explainable AI and its effects on human perceptions, in Proceedings of the 24th International Conference on Intelligent User Interfaces. 2019, Association for Computing Machinery: Marina del Ray, California. p. 263–274.
8. Shortliffe, E., Computer-based medical consultations: MYCIN. Vol. 2. 2012: Elsevier.
9. Guidotti, R., et al., A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys, 2018. 51(5): p. Article 93.
10. Ehsan, U., et al., Human-Centered Explainable AI (HCXAI): Beyond Opening the Black-Box of AI, in Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems. 2022, Association for Computing Machinery: New Orleans, LA, USA. p. Article 109.
11. Liao, Q.V. and K.R. Varshney, Human-centered explainable ai (xai): From algorithms to user experiences. arXiv preprint, 2021.
12. Liao, Q.V., et al., Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 2022. 10(1): p. 147-159.
13. Romero-Brufau, S., et al., A lesson in implementation: A pre-post study of providers' experience with artificial intelligence-based clinical decision support. International Journal of Medical Informatics, 2020. 137: p. 104072.
14. Kim, M., et al., Do stakeholder needs differ? - Designing stakeholder-tailored Explainable Artificial Intelligence (XAI) interfaces. International Journal of Human-Computer Studies, 2024. 181: p. 103160.
15. Amann, J., et al., Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making, 2020. 20(1):

p. 310.
16. Imrie, F., R. Davis, and M. van der Schaar, Multiple stakeholders drive diverse interpretability requirements for machine learning in healthcare. Nature Machine Intelligence, 2023. 5(8): p. 824-829.
17. Ehsan, U. and M.O. Riedl. Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. in International Conference on Human-Computer Interaction. 2020. Cham: Springer International Publishing.
18. Whittlestone, J., et al., Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. London: Nuffield Foundation, 2019.
19. Liao, Q.V., D. Gruen, and S. Miller, Questioning the AI: Informing Design Practices for Explainable AI User Experiences, in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020, Association for Computing Machinery: Honolulu, USA. p. 1–15.
20. Doshi-Velez, F. and B. Kim, Towards a rigorous science of interpretable machine learning. arXiv preprint, 2017.
21. Islam, S.R., W. Eberle, and S.K. Ghafoor. Towards quantification of explainability in explainable artificial intelligence methods. in The thirty-third international flairs conference. 2020.
22. Doshi-Velez, F. and B. Kim, Considerations for Evaluation and Generalization in Interpretable Machine Learning, in Explainable and Interpretable Models in Computer Vision and Machine Learning, H.J. Escalante, et al., Editors. 2018, Springer International Publishing: Cham. p. 3-17.
23. Hoffman, R.R., et al., Metrics for explainable AI: Challenges and prospects. arXiv preprint, 2018.
24. Schoonderwoerd, T.A.J., et al., Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. International Journal of Human-Computer Studies, 2021. 154: p. 102684.
25. He, X., et al., What Are the Users' Needs? Design of a User-Centered Explainable Artificial Intelligence Diagnostic System. International Journal of Human–Computer Interaction, 2023. 39(7): p. 1519-1542.
26. [Duffy, J.R., Motor speech disorders : substrates, differential diagnosis, and management. Fourth edition. ed. Evolve. 2020, St. Louis, Missouri: Elsevier.
27. Enderby, P., Frenchay dysarthria assessment. British Journal of Disorders of Communication, 1980. 15(3): p. 165-173.
28. Carmichael, J. and P. Green. Revisiting dysarthria assessment intelligibility metrics. in Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP'04). 2004.
29. Kale, D.C., et al., Causal Phenotype Discovery via Deep Networks. AMIA Annual Symposium Proceedings, 2015. 2015: p. 677-86.
30. Xu, L., J. Liss, and V. Berisha, Dysarthria detection based on a deep learning model with a clinically-interpretable layer. JASA Express Letters, 2023. 3(1).
31. Amershi, S., et al., Guidelines for Human-AI Interaction, in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019, Association for Computing Machinery: Glasgow, Scotland Uk. p. Paper 3.
32. Pandl, K.D., et al., Trustworthy machine learning for health care: scalable data valuation with the shapley value, in Proceedings of the Conference on Health, Inference, and Learning. 2021, Association for Computing Machinery: Virtual Event, USA. p. 47–57.
33. Lundberg, S.M. and S.-I. Lee, A unified approach to interpreting model predictions. Advances in neural information processing systems, 2017. 30.
34. Ribeiro, M.T., S. Singh, and C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016, Association for Computing Machinery: San Francisco, California, USA. p. 1135–1144.
35. Crabbé, J., et al., Explaining latent representations with a corpus of examples. Advances in Neural Information Processing Systems, 2021. 34: p. 12154-12166.
36. Cahour, B. and J.-F. Forzy, Does projection into use improve trust and exploration? An example with a cruise control system. Safety Science, 2009. 47(9): p. 1260-1270.
37. Administration, F.D., Proposed regulatory framework for modifications to Artificial Intelligence/Machine Learning (AI/ML)-based Software as a Medical Device (SaMD). 2019: Department of Health and Human Services (United States).
38. Mourby, M., K. Ó Cathaoir, and C.B. Collin, Transparency of machine-learning in healthcare: The GDPR & European health law. Computer Law & Security Review, 2021. 43: p. 105611.
39. Alvarez Melis, D. and T. Jaakkola, Towards robust interpretability with self-explaining neural networks. Advances in neural information processing systems, 2018. 31.
40. Miller, T., Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 2019. 267: p. 1-38.
41. Silva, W., et al. Towards Complementary Explanations Using Deep Neural Networks. 2018. Cham: Springer International Publishing.
42. Huysmans, J., et al., An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. Decision Support Systems, 2011. 51(1): p. 141-154.
43. Swartout, W.R. and J.D. Moore. Explanation in Second Generation Expert Systems. 1993. Berlin, Heidelberg: Springer Berlin Heidelberg.
44. Tjoa, E. and C. Guan, A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE

transactions on neural networks and learning systems, 2021. 32(11): p. 4793-4813.
45. Simkute, A., et al., XAI for learning: Narrowing down the digital divide between “new” and “old” experts, in Adjunct Proceedings of the 2022 Nordic Human-Computer Interaction Conference. 2022, Association for Computing Machinery: Aarhus, Denmark. p. Article 39.
46. Jacovi, A., et al., Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI, in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021, Association for Computing Machinery: Virtual Event, Canada. p. 624–635.
47. Parasuraman, R. and V. Riley, Humans and Automation: Use, Misuse, Disuse, Abuse. Human Factors, 1997. 39(2): p. 230-253.
48. Nauta, M., et al., From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. ACM Computing Surveys, 2023. 55(13s): p. Article 295.