Healthcare Explainable Artificial Intelligence-Possibilities, Challenges, and a Fresh Perspective on the Problem Space

[1]Sathishkumar M and [2]Dr. Raghavendran V

[1]Research Scholar, Department of Computer Science, School of Computing Sciences,
Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, Tamilnadu
600117 –India Email: sathishmohan355@gmail.com
[2]Assistant Professor, Department of Computer Science, School of Computing Sciences,
Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, Tamilnadu
600117- India Email: raganand78in@gmail.com

Abstract

Despite the potential that XAI can offer to the application of AI in this business, explainable artificial intelligence (XAI) is still in the early stages of acceptance in the healthcare sector. Standards for explanations, the level of interaction between various stakeholders and the models, the implementation of quality and performance metrics, the agreement on standards for safety and accountability, its integration into clinical workflows, and IT infrastructure are just a few of the issues that still need to be resolved. There are two goals for this Paper. The first one involves summarizing the findings of a literature review and highlighting the current state of explain ability, including any gaps, difficulties, and opportunities for XAI in the healthcare sector. We advise using a combined taxonomy to group explain-ability methodologies in order to facilitate understanding and onboarding into this field of study. The second goal is to determine whether using a novel strategy to analyze the explainability problem space through a particular problem or domain lens and automating that method in an AutoML-like manner would help reduce the issues outlined above. The literature has a propensity to view the explainability of AI through a model-first lens, which ignores real-world issues and domains. For instance, the explainability of a patient's survival model is handled similarly to the calculation of a hospital's procedure cost. We can (semi-)automatically find appropriate models, optimize their parameters and their explanations, metrics, stakeholders, safety/accountability level, and suggest ways of integrating them into clinical workflow when the problem or domain to which XAI should be applied is clearly defined.

Keywords: Artificial Intelligence, XAI, Explainable AI, Interpretability, Machine Learning, and AI in Healthcare.

1 INTRODUCTION

One of the primary reasons artificial intelligence (AI) is not yet (completely) trusted and is not therefore widely used in solving many real-world problems, including healthcare concerns, is the lack of explainability and transparency of state-of-the-art AI systems. In a recent set of recommendations for the use of AI in healthcare, the World Health Organization (WHO) calls for "ensuring explainability" [1]. Applications of AI in healthcare can save a lot of lives through early diagnosis or drug development, for example, but they also present special problems and risks if they are not applied and used properly. AI models that collaborate with people in the healthcare industry can help people make decisions, like diagnosis and prognosis. Holding AI models and assisted decision-making systems responsible can be extremely valuable [2], which is essential for their use in healthcare. Accountable another recent topic that is connected to the explainability of AI and ML systems is machine learning. ML systems include ML models, auxiliary software, hardware, and procedures, among other things. Explainability is utilized in this context to give the system the ability to defend its decisions, predictions, and in some cases even its thinking process, which is a crucial aspect of a system's accountability [3]. The initial goal and key contribution of this study is to describe the findings from our literature review and to develop a taxonomy that condenses as many XAI-related concepts as is practical into a "simple" structure. The overarching goal is to categorize and organize the vast body of knowledge so that it can be addressed more quickly, foster collaboration and make the subject more understandable, and highlight the state-of-the-art in explainability as well as the gaps, obstacles, and opportunities explainable artificial intelligence faces in the healthcare sector.

2 Current Outlook on XAI

A new field of study called explainable AI (XAI) and its applications to healthcare have been impacted by legislative initiatives and research initiatives including DARPA XAI [4], GDPR [5], and the upcoming Artificial Intelligence Act [6]. Although this subject is not entirely new, it has been rapidly expanding since the DARPA XAI program [4] was introduced in 2017. For instance, over 6,500 new articles were published between February and July of 2021 citing "Explainable AI" and almost 8,500 mentioning "Interpretable AI" [7, 8]. This astounding rate of publication growth demonstrates how tough it is for field researchers to keep up with everything. The recipient of the explanation is crucial in the multidisciplinary research field of XAI, which is affected by psychology, philosophy, sociology, the human-computer interaction, and education. However, this discipline still needs to develop in terms of how it applies to actual scenarios, and the research community must work to create consensus and standards for definitions, safety, measurements, etc. Let's examine the XAI definitions that have been presented. According to the literature, "Interpretability" is described as the characteristic of a model that offers sufficient expressive data to comprehend how the model operates [9], for example. The term "domain" is mentioned in the literature as being important for interpretability [9, 10], and some authors, like [10], contend that interpretability is innately domain-specific.

3 Possibilities, Challenges

In light of the aforementioned ideas, the following makes an effort to list the key topics that XAI research should focus on in the future. Since each point reflects a gap and presents an opportunity or difficulty depending on how well it is addressed, we haven't divided the points into opportunities, gaps, and challenges. For using XAI in the healthcare arena, all of the points listed below are pertinent:

- **Definition of the terms:** Agree upon definitions of interpretability and explainabil- ity, agree upon vocabulary and taxonomy of XAI methods, see [9-11, 14, 16-21, 29].
- **Explanations:** Reach an agreement what are good, human-friendly explanations, perform more exploration of the human-computer-interaction aspect of explana- tions such as visualization, using concepts or ontologies, etc. Create benchmarking models for the quality of explanations, automatic generations of explanations from the ML model, and reducing human subjectivity. Investigate legal implications of explanations provided, etc. See [12-15, 17, 29] for details.
- **Quality and performance metrics:** Create frameworks for the evaluation of per- formance and definition of agreed upon metrics for measuring and benchmarking XAI methods, see [9, 10, 13, 14]. An interesting interpretability challenge was found in the user study carried out with data scientists by [28] with the results ofthe study indicating that data scientists tend to over-trust and not correctly use the interpretability tools which can have dangerous implications.
- **ML Ops:** Integration and automation of XAI methods into ML model life cycle and deployment model, see [12, 20, 27, 28].
- **Safety:** We need to further research on security of explainability of AI, including methods to prohibit the fooling of XAI methods through perturbations and random- ized input as well as methods to mitigate inferring private and sensitive infor- mation through explanations, see [9, 10, 13, 17, 19-21]. In addition, XAI methods could be relevant to expose risks entailed in the large parameter natural language models such as GPT-3 [34] whose usage is growing. We are just becoming awareof additional security risks that could be impactful in healthcare. In [35] the authorsclaim that such models can memorize parts of the training data within their parameters. This means that it is possible to carry out attacks to retrieve potentially pri- vacy-sensitive information that were present in the training data.
- **Regulations:** Create regulatory and legislative framework for XAI involving fair- ness, protection of privacy, truthfulness of explanations and accountability.

4. Recommendations for Applying a Problem/DomainApproach to Interpretability and Explainability

The most recent method, as previously said, views explainability through a "model lens" that takes into account model-agnostic interpretability, model-specific interpretability, and inherently interpretable models. The same "model lens" is then used for every issue and every field. For examples of such earlier work. Because different problems demand varying levels of explainability, they have various stakeholders, and they frequently may be subject to different regulations, we believe that this method of using XAI frequently results in unnecessary complex circumstances. Require various justifications. A model's explainability in an emergency department scenario, which deals with life-or-death situations, will have different criteria than its explainability in a drug development scenario, for instance, where the scenarios can be extremely different. Authors have also mentioned the significance of the domain for XAI, for example in [10, 18]. To solve this, we propose applying a "problem/domain lens" approach to explainability from the beginning and using this approach to (semi-)automatically find the best XAI model for a problem in a domain and, if none is found, develop a new model by combining or redesigning existing ones or developing entirely new problem/domain explainable models taking relevant explanations, performance, and safety metrics, etc. into consideration.

Problem definition: Using AI to detect and classify type and potential abnormality ofwhite blood cells in patients' blood smear images with suspicion of haematological malignancy.

Potential machine learning algorithms: We know the common machine learning algorithms applied for such problems – e.g., CNN, Capsule Nets, Visual Transform- ers, etc. Note that we can also design intrinsically interpretable deep learning models for image recognition such as adapted versions of CNN or different architec-tures like .

Potential XAI methods: If we are not using an intrinsically interpretable model, we can use some of the common XAI methods like SHAP , LIME , gradient- based Grad-CAM [26] or attribution-propagation such as layer-wise relevance propa- gation (LRP) , attention map and others to verify that the model is look- ing at the right pixels/patches in the image for its decision. This is something that our approach could (semi-) automatically suggest given the problem/domain. Figure 1 below shows one of the experiments we conducted in exploring suitable XAI methodsfor the problem we defined above. The figure shows the results of using CNN and Grad-CAM for Acute Lymphoblastic Leukaemia (ALL) classifying lymphoblasts on ablood-smear image of ALL-IDB database.

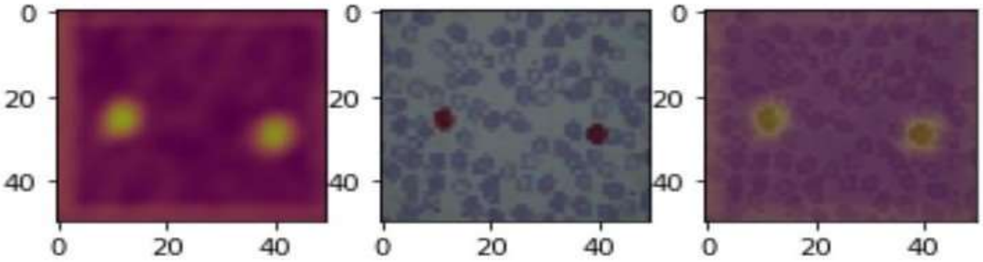

Fig. 1. An example of Grad-CAM implementation on ALL blood smears

Metrics: We can identify the KPIs that are relevant for this problem and domain, such as accuracy, number of patients that pathologist can perform this analysis per day, turnaround time from patient reception to diagnosis, response time for blood smear analysis. See [33] for more examples.

Stakeholders: For this problem the types of stakeholders such as doctors (pathologist, haematologist), nurses, patients, and insurance companies. Also, weneed to consider that we might need to discuss with stakeholders what XAI methods would be acceptable/preferable to them.

Safety and regulatory level: We need to be aware of the regulations, laws, and standards such as hospital regulations, insurance regulations, GDPR or future Artifi- cial Intelligence Act, etc.

Similar to automated machine learning (AutoML) [41], we are working on expand-ing this idea to conceptualize a framework/algorithm to automatize XAI for practi- tioners. The example above is just a short example, however having the knowledge of the problem and the domain lets us tailor a unique approach of the explainability and explanations that are relevant, using industry language and standards which makes them easy to use for the stakeholders. Using the AutoML or AutoXAI idea we could add a degree of automatism in selecting the right XAI method, parameter optimize- tion, metrics, stakeholders, and regulatory suggestions, and selecting the right level and type of explanations.

5. Conclusion

In order to enable a wider use and deployment of explainable machine learning assisted decision-making support systems in healthcare, it is critical to address the opportunities and challenges of explainable AI. The primary goal and contribution of this paper, as stated in the introductory chapter, is twofold: we communicated the summary of our field survey, and for simpler comprehension and onboarding to this research field, we suggesa synthesized taxonomy for categorizing explainabil- ity methods as well as a summary of opportunities, gaps, and challenges for applying explainability of AI to healthcare. Second, we are examining with our research topic how using ML models may be made simpler, more streamlined, and more customized by taking into account the problem, the context, and the domain. In order to assist ML practitioners in providing the proper "configuration settings" to a XAI application, we are also investigating whether this technique may be done in an AutoML-like manner, similar to AutoXAI.

References

1. World Health Organization (WHO). https://www.who.int/news/item/28-06-2021-who- issues-first-global-report-on-ai-in-health-and-six-guiding-principles-for-its-design-and-use. Last accessed on 2021/07/14.
2. Aurangzeb, A. M., Eckert, C., Teredesai, A.: Interpretable Machine Learning in Healthcare. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, pp. 559-560 (2018).
3. Pang, W., Markovic, M., Naja, I., Fung, C.P. and Edwards, P.: On Evidence Capture for Accountable AI Systems. In: SICSA Workshop on eXplainable Artificial Intelligence (XAI) (2021).
4. Gunning, D., Aha, D.: Explainable artificial intelligence (XAI) program. AI Magazine, 40(2), 44-58 (2019)
5. European Law General Data Protection Regulation. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02016R0679-20160504&qid=1532348683434. Last accessed 2021/07/27.
6. European Commission Artificial Intelligence Act. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206. lLst accessed on2021/07/18.
7. Dimensions query “Explainable AND Artificial Intelligence”. https://app.dimensions.ai/analytics/publication/overview/timeline?search_mode= content&search_text=explainable%20AND%20%22artificial%20intelligence%22&search_type= kws&search_field=full_search. Last accessed 2021/07/14.
8. Tjoa, E., Guan C . : A survey on explainable artificial intelligence (XAI): Toward medical XAI”. IEEE Transactions on Neural Networks and Learning Systems, 1–21 (2020).
9. Longo, L., Goebel, R., Lecue, F., Kieseberg, P, Holzinger., A.: Explainable Artificial Intel- ligence: Concepts, Applications, Research Challenges and Visions. In: International Cross- Domain Conference for Machine Learning and Knowledge Extraction, pp. 1–16. Spring- er (2020).
10. Adadi, A, Berrada. M.: Peeking inside the black box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access (6), 52138–52160 (2018).