User-Centric Explainable AI for Medical Image Diagnosis Prediction

ICIS 2022 – HITS-Health Information Technology Symposium

Abstract

Explaining increasingly complex machine learning will remain crucial to cope with risks, regulations, responsibilities, and human support in healthcare. However, extant explainable systems mostly provide explanations that mismatch clinical users' conception and fail their expectations to leverage validated and clinically relevant information. Meanwhile, research still lacks an understanding of how to effectively develop user-centric explanations for image diagnosis tasks. A promising paradigm to develop satisfying explanations that convey useful medical knowledge can be seen in informed machine learning, i.e., utilizing prior knowledge jointly with the patterns learned from data to generate predictions. Therefore, this study aims to contribute to the medical XAI challenge by enriching explanations for image diagnosis predictions with diagnostically relevant information that is congruent to a medical user's background and expertise. When accounting for multifaceted user requirements, explanations likely increase trust in clinical machine learning-based information systems.

User-Centric Explainable AI for Medical Image Diagnosis Prediction

ICIS 2022 – HITS-Health Information Technology Symposium

Introduction and Problem Statement

The well-known 'black box' challenge faced by medical AI seriously limits the chances of its effective translation into clinical practice (Rajpurkar et al. 2022). Risks, regulations, and responsibilities coin the need for specialized explainable artificial intelligence (XAI) that clinical users can fully understand and trust, instead of functioning as uninterpretable "black boxes" (Cutillo et al. 2020). Despite the recent surge of interest in developing methods to explain how AI-based systems produce a particular output, their utility in healthcare is questioned. One critic is that current XAI designs purely rely their results on the features, examples, or patterns found in the input data, for instance, through attribution-based methods. These mainly meet developers' demands, for instance, to debug and enhance models, but not their ultimate users' demands (Bauer et al. 2021). Such data-driven explanations were claimed by clinicians as inadequate to identify appropriate interventions (Jacobs et al. 2021) and mismatching human conception (Li et al. 2019). Instead, clinicians expect explainable systems to leverage validated and clinically relevant information (Jacobs et al. 2021) that supports their medical knowledge and reflect a similar analytic process to medical decision-making (Tonekaboni et al. 2019).

To develop more meaningful explanations, systems must be informed by useful information and usefully presented in a given context to a given user (Evans et al. 2022). Meanwhile, it has recently been concluded that integrating prior knowledge into machine learning (ML) systems is essential for better explanations of their functioning (Gaur et al. 2021). However, research lacks an understanding of how to integrate knowledge that is congruent to a user's background and expertise to design user-centric explanations for medical diagnosis predictions. At the same time, XAI lacks research that bridges the gap between implementation and user experience (Evans et al. 2022). Since it has been reported that the expectations of

system designers, which information will build users' trust, largely differ from the results in clinical practice (Lahav et al. 2018), justifying whether explanations achieve their goal in real-world settings is critical in healthcare (Das and Rad 2020). Thus, we aim to contribute to the general challenge of designing user-centric XAI and evaluate its effectiveness based on the interaction of real clinicians.

Prior Research and Research Goal

Data-driven approaches to explain image-based ML include visualizing the convolution filters of convolutional neural networks (CNNs) or influential image regions using saliency maps which, however, do not allow for the integration of additional modalities (Rajpurkar et al. 2022). While purely data-driven models might not conform to constraints and available knowledge from the domain and context, informed machine learning (IML) deals with using a separate source of information and integrating it into an ML pipeline (Beckh et al. 2021; Rueden et al. 2021). This has been shown to overcome the dependency on large amounts of data (Choi et al. 2017; Ma et al. 2018) and to increase performance, reliability, as well as robustness (Deng et al. 2020). It has also been recently revisited to provide users with more context-aware and more usable explanations (Beckh et al. 2021). The spectrum of IML approaches ranges from the integration of knowledge from scientific disciplines, common sense, or the respective field of application (e.g., medical ontologies or lexica), in various representations such as logical rules, equations, constraints, simulations, or knowledge graphs (Beckh et al. 2021; Deng et al. 2020; Rueden et al. 2021). The hierarchical semantic CNN developed by Shen et al. (2019), for instance, adds two prediction levels to a CNN: Multiple subnetworks each predict a semantic clinical indicator that is relevant to radiologists. The outputs of both the CNN and intermediate information are then concatenated into another subnetwork for the final diagnosis prediction task. Such concept learning approaches, however, present predicted clinical indicators without further explanation, i.e., leave the intervention to find effects on the final prediction to the user (Koh et al. 2020). Other concept learning approaches present global explanations (Bau et al. 2017) or rely the final diagnosis on a few clinical concepts without combining deep image-based features (LaLonde et al. 2020).

Thus, there are two types of predictive computer-aided diagnosis (CADx) systems in medical imaging (Hancock and Magnan 2016). (1) Images are directly used by an algorithm to produce a binary-valued outcome, for instance, whether lung nodules are malignant. (2) Diagnostic image features, such as the size or volume, are computationally quantified, whereas the end goal may be displaying them to users or using them as inputs to a downstream CADx. While the former approach may be less interpretable, if at all, the latter assumes that features are accurate and only these can be used to provide an accurate final diagnosis. This work aims to overcome the opaqueness of end-to-end ML for the first type of CADx. In contrast to purely data-driven approaches, we seek to expose how extant IML can be used to both learn from image data and integrate diagnostically relevant knowledge, to ultimately generate explanations that are user-centric, i.e., congruent to a clinical user's background and expertise. This results in our research question: "*How to design effective, user-centric XAI for medical image diagnosis using informed ML?*"

Research Approach

Design Science Research (DSR) is concerned with designing and evaluating innovative IT artifacts to solve organizational problems (Peffers et al. 2018). We apply Action Design Research (ADR) proposed by Sein et al. (2011). The approach combines DSR-based artifact design and construction with an evaluation as a continuous interaction between researchers and practitioners throughout each design cycle. The ADR method fully recognizes the role of the organizational context in shaping the design process of an explainable diagnostic imaging system. Moreover, it lets us incorporate clinical end-users in the research process to design and evaluate the artifact in a real application context with continuous feedback from radiology experts. Lastly, it allows us to produce theoretical outputs based on the theoretical foundations gained through the literature on user-centricity, XAI, as well as IML; from which we derive prescriptive design knowledge for general user-centric explainability in medical imaging.

Project Status

The project will be accompanied by clinicians from an associated radiology department. ADR suggests four sequential phases interwoven with the construction of the artifact: (1) Problem Formulation; (2) Building,

Intervention, and Evaluation (BIE); (3) Reflection and Learning; and (4) Formalization of Learning (Sein et al. 2011). The first phase serves to permeate the problem and define the objectives of the practice-inspired research. Therefore, we conducted a preliminary interview consisting of open questions concerning problematic diagnosis tasks, related CADx systems, possible explainability features, as well as their utility of the latter in medical practice. This revealed that ML-based decision support is highly important for the diagnosis of lung diseases and local (i.e., patient-level) explanations are highly favored over black-box predictions. It was mentioned that the classification of lung nodules involves diagnostic features that not only help a doctor to differentiate between benign and malignant (i.e., cancerous lesions) lesions but also judge the extent of follow-up screening activities and explain the severity of a patient's diagnosis. Since clinicians are familiar with these features (Shen et al. 2019), they are well suited to radiologist-level explainability. To this end, this study used lung nodule classification as a research case. To ensure a theory-ingrained artifact, literature about different approaches to ML-based automated ICD coding was reviewed. As outlined, research lacks IML that provides explanations based on diagnostically relevant features, for instance, the relevance to the final prediction or corresponding image annotations. Based on the formulated problem, we will develop a prototype to explore and evaluate multiple explanation types. For the first BIE loop, we will build on the model proposed by Shen et al. (2019) as it learns from hybrid information by implementing both diagnosis as well as semantic feature prediction tasks, and extend the model with additional explanation methods. We will adapt integrated gradients (Sundararajan et al. 2017) and discuss the feature attribution of (a) input pixels to the final diagnosis, (b) input pixels to diagnostic features, and (c) diagnostic features to the final diagnosis with clinicians. In subsequent BIE loops, we will address questions regarding further relevant information, such as screening decisions based on diagnostic features, and further knowledge representations, such as rule-based explanations about diagnostic features.

Potential Contributions

Overall, this study seeks to contribute to the medical XAI challenge by exposing how IML approaches can be used to generate explanations that are user-centric and align with prior knowledge. We envision our

contributions to be threefold as follows. First, we identify IML techniques to tackle the challenges of data-driven explanations and assess them according to their applicability to the case of medical image diagnosis. Second, we derive prescriptive design knowledge for XAI systems in the form of design principles, generalized to medical diagnosis prediction, as an important step toward user-centricity. Third, we evaluate a prototype in an application-grounded manner, such that the design is rigorously guided via expert interaction and likely inspire more trust in clinical settings.

References

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. 2017. "Network Dissection: Quantifying Interpretability of Deep Visual Representations," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, IEEE, pp. 3319-3327 (doi: 10.1109/CVPR.2017.354).

Bauer, K., Hinz, O., van der Aalst, W., and Weinhardt, C. 2021. "Expl(AI)n It to Me – Explainable AI and Information Systems Research," *Business & Information Systems Engineering* (63:2), pp. 79-82 (doi: 10.1007/s12599-021-00683-2).

Beckh, K., Müller, S., Jakobs, M., Toborek, V., Tan, H., Fischer, R., Welke, P., Houben, S., and Rueden, L. von. 2021. "Explainable Machine Learning with Prior Knowledge: An Overview,"

Choi, E., Bahadori, M. T., Le Song, Stewart, W. F., and Sun, J. 2017. "GRAM: Graph-based Attention Model for Healthcare Representation Learning," *International Conference on Knowledge Discovery & Data Mining* (2017), pp. 787-795 (doi: 10.1145/3097983.3098126).

Cutillo, C. M., Sharma, K. R., Foschini, L., Kundu, S., Mackintosh, M., and Mandl, K. D. 2020. "Machine intelligence in healthcare-perspectives on trustworthiness, explainability, usability, and transparency," *NPJ digital medicine* (3), p. 47 (doi: 10.1038/s41746-020-0254-2).

Das, A., and Rad, P. 2020. "Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey," arXiv:2006.11371.

Deng, C., Ji, X., Rainey, C., Zhang, J., and Lu, W. 2020. "Integrating Machine Learning with Human Knowledge," *iScience* (23:11), p. 101656 (doi: 10.1016/j.isci.2020.101656).

Evans, T., Retzlaff, C. O., Geißler, C., Kargl, M., Plass, M., Müller, H., Kiehl, T.-R., Zerbe, N., and Holzinger, A. 2022. "The explainability paradox: Challenges for xAI in digital pathology," *Future Generation Computer Systems* (133), pp. 281-296 (doi: 10.1016/j.future.2022.03.009).

Gaur, M., Faldu, K., and Sheth, A. 2021. "Semantics of the Black-Box: Can Knowledge Graphs Help Make Deep Learning Systems More Interpretable and Explainable?" *IEEE Internet Computing* (25:1), pp. 51-59 (doi: 10.1109/MIC.2020.3031769).

Hancock, M. C., and Magnan, J. F. 2016. "Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms: probing the Lung Image Database Consortium dataset with two statistical learning methods," *Journal of Medical Imaging* (3:4) (doi: 10.1117/1.JMI.3.4.044504).

Jacobs, M., He, J., F. Pradier, M., Lam, B., Ahn, A. C., McCoy, T. H., Perlis, R. H., Doshi-Velez, F., and Gajos, K. Z. 2021. "Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing*

Systems, Y. Kitamura, A. Quigley, K. Isbister, T. Igarashi, P. Bjørn and S. Drucker (eds.), Yokohama Japan. 08 05 2021 13 05 2021, New York, NY, USA: ACM, pp. 1-14 (doi: 10.1145/3411764.3445385).

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. 2020. "Concept Bottleneck Models," in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh (eds.), PMLR, pp. 5338-5348.

Lahav, O., Mastronarde, N., and van der Schaar, M. 2018. "What is Interpretable? Using Machine Learning to Design Interpretable Decision-Support Systems," arXiv:1811.10799.

LaLonde, R., Torigian, D., and Bagci, U. 2020. "Encoding Visual Attributes in Capsules for Explainable Medical Diagnoses," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu and L. Joskowicz (eds.), Cham: Springer International Publishing, pp. 294-304 (doi: 10.1007/978-3-030-59710-8_29).

Li, X., Qian, B., Wei, J., Zhang, X., Chen, S., Zheng, Q., and (Keine Angabe). 2019. "Domain Knowledge Guided Deep Atrial Fibrillation Classification and Its Visual Interpretation," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, W. Zhu, D. Tao, X. Cheng, P. Cui, E. Rundensteiner, D. Carmel, Q. He and J. Xu Yu (eds.), Beijing China, New York, NY, USA: ACM, pp. 129-138 (doi: 10.1145/3357384.3357998).

Ma, F., You, Q., Xiao, H., Chitta, R., Zhou, J., and Gao, J. 2018. "KAME: Knowledge-based Attention Model for Diagnosis Prediction in Healthcare," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, A. Cuzzocrea, J. Allan, N. Paton, D. Srivastava, R. Agrawal, A. Broder, M. Zaki, S. Candan, A. Labrinidis, A. Schuster and H. Wang (eds.), Torino Italy, New York, NY, USA: ACM, pp. 743-752 (doi: 10.1145/3269206.3271701).

Peffers, K., Tuunanen, T., and Niehaves, B. 2018. "Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research," *European Journal of Information Systems* (27:2), pp. 129-139 (doi: 10.1080/0960085X.2018.1458066).

Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E. J. 2022. "AI in health and medicine," *Nature medicine* (28:1), pp. 31-38 (doi: 10.1038/s41591-021-01614-0).

Rueden, L. von, Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Walczak, M., Pfrommer, J., Pick, A., Ramamurthy, R., Garcke, J., Bauckhage, C., and Schuecker, J. 2021. "Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems," *IEEE Transactions on Knowledge and Data Engineering*, p. 1 (doi: 10.1109/TKDE.2021.3079836).

Sein, Henfridsson, Purao, Rossi, and Lindgren. 2011. "Action Design Research," *MIS Quarterly* (35:1), p. 37 (doi: 10.2307/23043488).

Shen, S., Han, S. X., Aberle, D. R., Bui, A. A., and Hsu, W. 2019. "An Interpretable Deep Hierarchical Semantic Convolutional Neural Network for Lung Nodule Malignancy Classification," *Expert systems with applications* (128), pp. 84-95 (doi: 10.1016/j.eswa.2019.01.048).

Sundararajan, M., Taly, A., and Yan, Q. 2017. "Axiomatic Attribution for Deep Networks," in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh (eds.), PMLR, pp. 3319-3328.

Tonekaboni, S., Joshi, S., McCradden, M. D., and Goldenberg, A. 2019. "What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use,"