**Workshop: Explainable AI in medicine - A critical appraisal of limitations and insights for future developments**, Lugano, Switzerland, November 2-3, 2023, https://www2.idsia.ch/cms/xai-med23/

Abstract

### A framework for assessing and certifying explainability of health-oriented AI systems

Denzel, Philipp and Brunner, Stefan and Luley, Paul-Philipp and Frischknecht-Gruber, Carmen and Reif, Monika Ulrike and Schilling, Frank-Peter and Amini, Amin and Repetto, Marco and Iranfar, Arman and Weng, Joanna and Chavarriaga, Ricardo

Explainability has been recognized as one of the key tenets for the development of trustworthy AI systems for health-related applications. As regulation for AI is being developed, organizations deploying health-oriented AI systems will have to comply with requirements on transparency and explainability. However, despite the imminent introduction of these regulations, actionable guidelines for assessing compliance are still lacking.

We present ongoing work on developing a framework for assessing and certifying the transparency of AI systems. This framework makes an explicit link between the foreseen certification requirements and validated processes, algorithms, and methods for assessing compliance of AI systems; resulting in a concrete workflow to perform AI certification. It is based on analysis of the proposed AI regulation in the EU, recommended practices, and ISO standards. This is complemented by empirically validated state-of-the-art algorithmic methods for explainable AI in real-world applications.

Stakeholders have unique requirements for the explainability of AI systems. Meanwhile, a wide variety of explainable AI methodologies are available and may be appropriate for different stakeholder preferences, data modalities, applications, and purposes. As a result, the nuanced selection of relevant methodologies becomes an indispensable consideration in this framework. Therefore, within this framework, a taxonomy has been developed to guide the decision for selecting the appropriate and applicable set of methods. The framework and the application of the taxonomy is illustrated through several health-related use cases.

Take the case of a skin lesion classification system, involving as stakeholders the patient, the dermatologist, the developers, and the authorities. Here, several considerations guide the choice of methods: e.g., which stakeholder should receive the explanation, is the model directly interpretable or not, are intrinsic or post-hoc explanatory methods required, or whether explanations should be local or global. To mention some of the methods suitable for the dermatologist based on these considerations: In cases where local explanations are required and deep learning methods are used, our framework will point to methods such as SHAP or LIME that illustrate what features in the image led to the model's decision. Such methods might indicate that a lesion was classified as malignant due to its asymmetric shape, ill-defined border, or irregular colour.

Likewise, developers and regulators may require other types of explanations. This could include, for example, explaining global behavior in conjunction with local behavior, as well as identifying model weaknesses throughout the learning and verification stages by providing feature-level explanations.

In essence, the presented framework will provide a concrete guide for researchers, developers, and certification bodies in the development, validation, and certification of explainable, transparent AI systems and promote the adoption of best practices for responsible AI innovation.

https://digitalcollection.zhaw.ch/handle/11475/29258