Evaluation of Explainable AI by Medical Experts: a Survey of the Existing Approaches

Nikolay Babakov[1,*], Elena Rezgova[2], Ehud Reiter[3] and Alberto Bugarín[1]

[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Galicia, Spain

[2]Health Technologies LLC, Saint Petersburg, Russian Federation

[3]University of Aberdeen, Aberdeen, UK

Abstract

In this survey, we examine the landscape of Explainable Artificial Intelligence (XAI) techniques evaluation within the medical domain, focusing on studies evaluated by medical practitioners. Our analysis delves into the prevailing trends and identifies notable deficiencies in the current evaluation methodologies. Notably, we uncover that significant details of evaluation studies — such as the user study interface, study location, and participant remuneration—are often disregarded in the final reports describing the evaluation studies. Furthermore, our findings reveal a concerning scarcity of statistical significance testing in evaluation results, leading to overly optimistic conclusions regarding the applicability of XAI. Additionally, we highlight a prevalent trend of assessing XAI in isolation, devoid of comparative analysis, and an unbalanced emphasis on XAI perception attributes like usefulness, human-AI performance, and clinical relevance, overshadowing other crucial properties. Another issue identified in our survey is the imprecise formulation of participant queries, resulting in an excessive number of similarly purposed questions. The culmination of our study is not only an exposition of these findings but also a curated set of recommendations consisting of the main steps to be done before, during, and after the evaluation study aimed at researchers endeavoring to deploy XAI techniques in real medical applications. These guidelines are designed to enhance the genuine usability evaluation of XAI tools by medical professionals, ensuring a robust and meaningful application of XAI in healthcare.

Keywords

explainable artificial intelligence, evaluation by medical practitioners

1. Introduction

Artificial intelligence (AI) is rapidly transforming the landscape of medical applications, offering unprecedented advancements in disease diagnosis, patient care, and treatment methodologies. The integration of AI within the healthcare sector promises to enhance the efficiency, accuracy, and accessibility of medical services, thereby improving patient outcomes across a variety of disease areas, including cancer, neurology, and cardiology [1]. However, the opaque nature of most AI algorithms poses a significant challenge, particularly in the critical domain of medicine, where decisions directly impact patient health and lives. This opacity has ignited a growing necessity for explainable AI (XAI) [2] that can demystify AI decision-making processes, ensuring transparency and accountability, and facilitating trust in AI-driven medical interventions [3]. As the demand for explainability in AI increases, so does the importance of developing AI systems that not only predict and diagnose with high precision but also articulate their reasoning in ways that medical practitioners can understand and trust, thereby safeguarding patient welfare [4].

XAI can greatly aid medical practitioners by clarifying AI decisions, supporting tasks from diagnosis to treatment planning [5]. In diagnostic imaging, XAI helps specialists interpret complex medical images, leading to more accurate diagnoses [6]. It also advances personalized medicine by analyzing patient data to recommend tailored treatments [7], and supports patient monitoring and predictive

EXPLIMED 2025 - Second Workshop on Explainable Artificial Intelligence for the Medical Domain - 25–30 October 2025, Bologna, Italy

*Corresponding author.

nikolay.babakov@usc.es (N. Babakov)

healthcare through explainable risk assessments, enabling early interventions [8]. Overall, XAI enhances practitioner capabilities and promotes a more transparent and patient-centric healthcare system.

Although many works discuss XAI evaluation properties [9, 10], there are no generally established practices for evaluating XAI techniques in the medical domain. This creates a significant gap in assessing how medical professionals perceive and understand XAI in real-world tasks.

The lack of standardized evaluation frameworks limits assessment of how well medical practitioners interpret, trust, and use XAI in clinical decision-making. The variety of medical applications and data types complicates unified evaluation, while XAI's interdisciplinary nature further challenges the development of metrics that capture both understanding and practical usability [11]. Without such standards, refining XAI models to meet practitioners' needs is difficult, hindering their integration and acceptance in healthcare. Addressing this gap is crucial to realizing XAI's potential to improve patient care and outcomes.

While a universal approach to XAI evaluation in medicine is challenging, aggregating current practices can be valuable. In this paper, we present a structured survey of XAI evaluation methods used by medical practitioners, analyzing methodologies to reveal trends and gaps. Our synthesis highlights the range, effectiveness, and applicability of evaluation strategies across medical applications, identifying strengths and weaknesses both within and across studies. Based on these findings, we propose recommendations to help standardize and improve XAI evaluation, aiming to better align future XAI tools with the needs of healthcare settings.

The main differences between our survey and the existing surveys discussing XAI evaluation in general [12, 13, 14, 15] and in the medical domain [16, 17] are as follows:

- Our paper addresses the evaluation of XAI without narrowing the scope of the studied papers to the particular disease or data type (unlike, e.g., [17] considering only cardiological diseases or [18] considering only tabular and time series data)
- The methodology of information extraction and analysis differs significantly from all papers related to similar topics, which allows us to obtain valuable insights and propose original recommendations to the medical XAI community

2. Related literature

2.1. Explainable AI

There are several ways XAI techniques gain insights into certain AI models' decision-making processes that can be divided into two groups: post-hoc explanations aiming to explain the decision of "black box" models and ante-hoc interpretability utilizing the transparent models to get the explanations [19]. Each approach provides a different pathway to achieve transparency and understanding in AI models, catering to the nuanced requirements of various applications.

Post-hoc explanation methods are designed to elucidate the decision-making of already trained predictive models. This category includes tools like SHAP (SHapley Additive exPlanations) [20], LIME (Local Interpretable Model-agnostic Explanations) [21], and GRAD-CAM (Gradient-weighted Class Activation Mapping) [22]. SHAP and LIME offer insights into the contribution of each feature to the prediction of a model on a case-by-case basis, allowing users to understand the reasoning behind specific decisions. GRAD-CAM, on the other hand, provides visual explanations for decisions made by convolutional neural networks, particularly useful in tasks like image classification, by highlighting the areas of the image most influential to the model's prediction. These post-hoc methods are invaluable for dissecting complex models after training, offering a lens through which the inner workings of opaque models can be examined and interpreted.

Ante-hoc interpretability involves integrating interpretability by design directly into the architecture of the predictive model. This can be achieved through models inherently interpretable, such as decision trees [23, 24], or by incorporating interpretability constraints such as sparsity [25] or attention mechanisms [26] within the training process. Decision trees provide a straightforward, rule-based framework

for decision-making, making their logic easily traceable. Attention mechanisms, commonly used in neural network architectures, highlight the importance of different parts of input data, such as words in a sentence or pixels in an image, directly linking model outputs to specific input features. By embedding interpretability into the model design, ante-hoc methods ensure transparency and ease of understanding from the outset, making these models particularly suited for applications where explainability is as critical as accuracy.

2.2. Existing surveys about explainable AI

The growing demand for explainability in various fields yields a vast amount of papers related to novel XAI techniques and their application. This results in many surveys about XAI, in general, covering many facets of XAI such as problem definitions, goals, etc. [27, 28, 29, 30]. [31] discussed various aspects related to the deployment of XAI-based models. [32] concentrated on the diverse existing literature on counterfactuals and causability in XAI. [19] presented a filed guide for AI novices to effectively use XAI techniques.

The importance of explainability and interpretability in AI-driven decision-making in medicine also results in numerous corresponding surveys [5, 33]. [34] discussed recent trends in medical diagnosis and surgical applications using XAI. [35] identified nine different types of interpretability methods that have been used for understanding deep learning models for medical image analysis applications based on the type of generated explanations and technical similarities. [36] discussed the problem of the trustworthiness of AI applications in medicine. The systematic reviews performed in [37]resulted in structured recommendations for XAI applications in the medical domain.

2.3. Evaluation of explainable AI

While established metrics are available to assess the performance of predictive models, a consensus on an evaluation framework for XAI remains elusive. This gap underscores the complexity of quantifying the effectiveness of explanations in AI systems [9]. [10] highlighted such important directions of XAI evaluation as the goodness of explanations, user satisfaction, trust, and mental model, the role of curiosity in the search of explanations, and the performance of collaboration between user and XAI system. [38] identified 12 conceptual properties of XAI evaluation, performed a structured survey, and analyzed what properties are scored more frequently in the papers engaging XAI evaluation. [39] proposed to adapt the user-centric evaluation framework used in recommender systems to contribute to the human-centered standardization of XAI evaluation. [40] presented a taxonomy that provides guidance for researchers and practitioners on the design and execution of XAI evaluations. [41] showed that the XAI evaluation setup must be based on the real decision task rather than on artificial proxy tasks. [42] discussed that, apart from collecting feedback from experts, there are still several ways to evaluate the XAI techniques objectively in a user-agnostic manner. [43] summarized a human-centered demand framework to categorize different groups of XAI users into five key roles with specific demands by reviewing existing research and then extracting six commonly used human-centered XAI evaluation measures that help validate the effect of XAI. [44] proposed a set of recommendations on designing user evaluations in the field of XAI and performed an extensive user evaluation on the effects of rule-based and example-based contrastive explanations. There are also numerous surveys aimed at collecting existing practices in XAI evaluation from different points of view [12, 13, 14, 15].

XAI evaluation in medicine has also been studied from different perspectives [45]. [46] proposed to design and evaluate clinical XAI systems based on five principal guidelines: understandability, clinical relevance, truthfulness, informative plausibility, and computational efficiency. [47] delved into the critical consideration of various XAI failures and their potential implications on decision-making for individual patients. [48] discussed a use case of a decision support system integrated into a platform for healthcare professionals and demonstrates the importance of human-centered evaluation methods and potential struggles with mixed methods as detected by differences between qualitative and quantitative approaches.

There are also several existing surveys dedicated to collecting and analyzing the ways of XAI evaluation in medicine. [16] analyzed both human-dependent and human-agnostic methods of XAI technique evaluation. [17] reviewed the existing XAI evaluation practices in cardiology. [49] studied essential properties of XAI and an evaluation of explanation effectiveness in the healthcare field in a systematic review setup, which resulted in a pretty limited number of papers (only six papers were analyzed). [18] performed a survey analyzing different sides of XAI evaluation by clinicians, but it was dedicated only to tabular and time series data. In our paper, we perform a structured survey of the papers engaging the evaluation of XAI by medical practitioners, not limiting the scope of the papers to any specific data type or disease. Moreover, the methodology of feature extractions and analysis of the found papers differs from the aforementioned papers in many regards (e.g., certain parts of the XAI taxonomy were inherited from [38], more details are available in Section 3), which allows us to obtain valuable insights and propose original recommendations to the medical XAI community.

3. Survey methodology

3.1. Literature search

We conduct the survey following the standards described in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines for Scoping Reviews (PRISMA-ScR) [50]. We used the following scientific papers databases: Scopus [51], Web of Science [52], IEEE Xplore [53], ACM [54], and Pubmed [55].

We aim to find papers dedicated to the application of XAI techniques to the medical domain, that involve evaluation of XAI technology used by medical practitioners. To find such papers we run the database search by the papers' abstracts using the following query: `(XAI OR (explain* OR interpret*) AND (artificial AND intelligence OR (deep OR machine) AND learning OR neural AND network*)) AND ((*medic* OR health* OR clinic* OR hospital) OR (disease OR ill* OR disorder OR infection OR sick*)) AND (user* OR expert* OR practitioner* OR physician*) AND (eval* OR asses* OR validat*)`. The "*" symbol (also referred to as a "wildcard" operator), is used to substitute for one or more unknown characters in a search term.

3.2. Study selection

We define the following inclusion and exclusion criteria for the collected papers. We consider papers written in English that either introduce the original XAI technique or use existing ones in the medical domain and evaluate the performance of the XAI with medical professionals. We disregard paper if it is not related to medicine, if it does not use the XAI technique, or if the XAI technique is used without verification of its effectiveness by medical professionals.

The time range of the paper selection is between 2018 and 2023, inclusively. We decided to use 2018 as a starting year because, from numerous studies, it was clear that the trend of XAI research started in 2018 [5, 38, 56], which is probably a consequence of the XAI 2016 DARPA challenge [57]. 2023 was the last year fully available for the literature extraction (we queried the databases in February 2024).

The papers retrieved from the database using the aforementioned query were screened for inclusion by title and abstract by two authors of the paper (one has experience in XAI, another in medicine). The first 300 titles and abstracts were reviewed by both authors, and the result in Cohen's Kappa score for inter-annotator agreement was 0.81. Other papers were divided equally and screened independently. In cases of doubt, the inclusion or exclusion of certain papers was discussed collectively. The full content of the papers considered after screening was analyzed and, if considered relevant to the inclusion criteria, the paper was included in the final scope of the survey's papers. Figure 1 shows the literature selection process corresponding to the aforementioned setup.

3.3. Data synthesis and summarization

We extract the following data from the papers considered for analysis:

- **General information:** Year of publication, Medical domain: disease or specific task
- **AI model information:** Type of predictive model (Deep Neural Network, Bayesian or Hierarchical Network, Support Vector Machine, Tree Ensemble, other), Type of data (graph, image, tabular, text, time series, user-item matrix, video, other), Type of AI model task (classification, regression, policy learning, other) [38]
- **XAI technique information:** Type of method used to explain (post-hoc explanation, built-in or ante-hoc interpretability), Type of explanation (decision tree, feature importance, heatmap, text, etc), Type of XAI task (model explanation, model inspection, outcome explanation, transparent box design) [38]
- **Medical professionals evaluation design:** Number of participants, Number of cases shown to each participant, Place of study (online, offline), Remuneration, Type of evaluation task (proxy or real), Is the XAI technique interface clearly defined?,Is the user study interface clearly defined?, Was evaluation pre-registered (e.g., with AsPredicted service [58])?, Experimental design (between, within, mixed) [59], General aim of the evaluation (Compare original XAI technique with existing ones, Compare different existing XAI techniques on a certain task, Compare explained and non-explained AI techniques, Compare the performance of medical professionals completing tasks on their own with the performance aided by XAI, Direct assessment of XAI technique (without alternatives), Explanation effect (positive, mixed, non-positive) [59], Statistical verification of evaluation results
- **Information about each task delegated to medical professionals:** Formulation of the task (e.g., the text of the question), Way of answer collection (e.g., Likert-scale or observing the accuracy of diagnosis), Task frequency (e.g., after each case, after all cases)

AI model information and *XAI technique information* are re-used directly from [38] as far as they provide a comprehensive understanding of the AI models and XAI techniques used in each paper.

Medical professionals' evaluation design contains the core information representing the main peculiarities of the evaluation studies. While most features in this group are self-explanatory, others may deserve additional clarification. We consider two principal types of evaluation tasks: proxy and real. The real task assumes that a medical professional decides on certain patients' data relying on the assistance of the XAI technique (or another technique proposed in the experiment). A proxy task does not engage a medical professional to make a decision but asks to assess the perceived quality of the system's performance from different perspectives. There are three common experimental design setups when conducting user evaluation: between-subjects (or between-groups) designs, within-subjects designs, and mixed designs that combine elements of both. The between-subjects setup assumes that one subject (i.e., participant) is only exposed to one condition of the experiment. In terms of within-subject setup, each participant sequentially passes through all conditions. Explanation effect may generally be classified as positive, mixed, or non-positive, and a certain value of this parameter is extracted from the self-reported analysis of the performed XAI evaluation experiments in the studied papers.

Information about each task delegated to medical professionals involves individual analysis of each task delegated to the medical professionals in terms of an evaluation study. Normally, such tasks are either answering questions about the perceived quality of the XAI technique or deciding a certain patient's

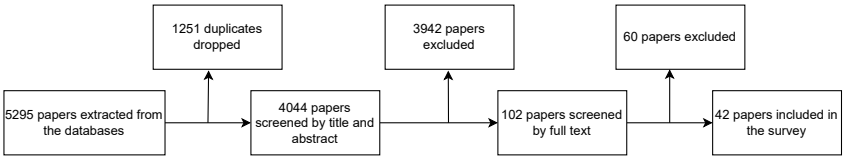

Figure 1: Literature selection diagram.

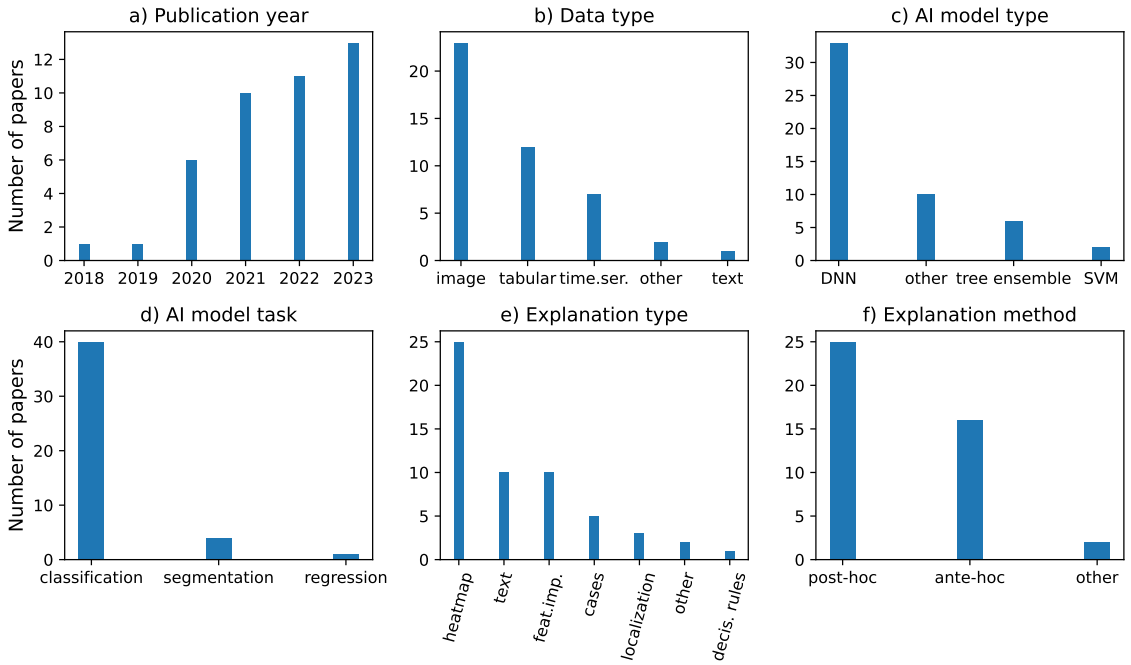

Figure 2: General statistics of the models and XAI techniques used in the papers included in the survey. If several types of reported entities (e.g., data type, AI model type, etc.) are used within one paper, then the paper is counted several times on the corresponding plot.

case with the assistance of the XAI technique. In most cases, the formulation of the task corresponds to the formulation of the question asked to a participant. The way of answer collection may be either direct collecting of the responses to the questions or observation of the participant's performance by checking the quality of performing tasks or analyzing the logs of interaction with the proposed XAI system. Task frequency analyzes the number of times certain tasks are demonstrated to the participants, which in most cases may be either after each demonstrated case or after using a system with all cases.

During the analysis of each task, we also aim to assign the task to one of the groups of XAI quality properties: Trust [10, 16, 60, 49], Mental model (how users understand a system) [10, 60], Agreement with AI decision, Human-AI performance [60, 49, 13], Plausibility of explanation, clinical relevance [17, 13], Usefulness [60, 16, 39], User confidence in explanation [38, 61], Complexity [17, 38, 39], Correctness [38, 39], Completeness [17, 38, 14], Satisfaction [60, 49, 39, 13]

This list of the XAI perceived quality properties is not comprehensive. It is formed dynamically during the analysis of the survey papers, relying both on the existing literature on XAI (the cases with the corresponding citations) and on the questions that are not frequently discussed in such literature but naturally form the questions cluster of considerable size (e.g., agreement with AI decision). We acknowledge the existence of some known taxonomies of XAI quality properties, such as a fundamental list proposed in [10] (trust, goodness, etc.) and a 12-Co list proposed in [38], however for the aims of this survey we found it more useful to create a list that would correspond to the XAI evaluation properties really occurring in the papers included in the survey.

4. Results

4.1. Literature selection

The keywords-based query to the databases returned 5295 papers. After deduplication, we drop 1251 papers and screen titles and abstracts of the rest 4044. This results in the disregarding of 3942 papers due to inconsistency with the stated inclusion and exclusion criteria (see Section 3.2). We read the full text of 102 papers and exclude 60 more from the survey, as far as they also did not meet the criteria.

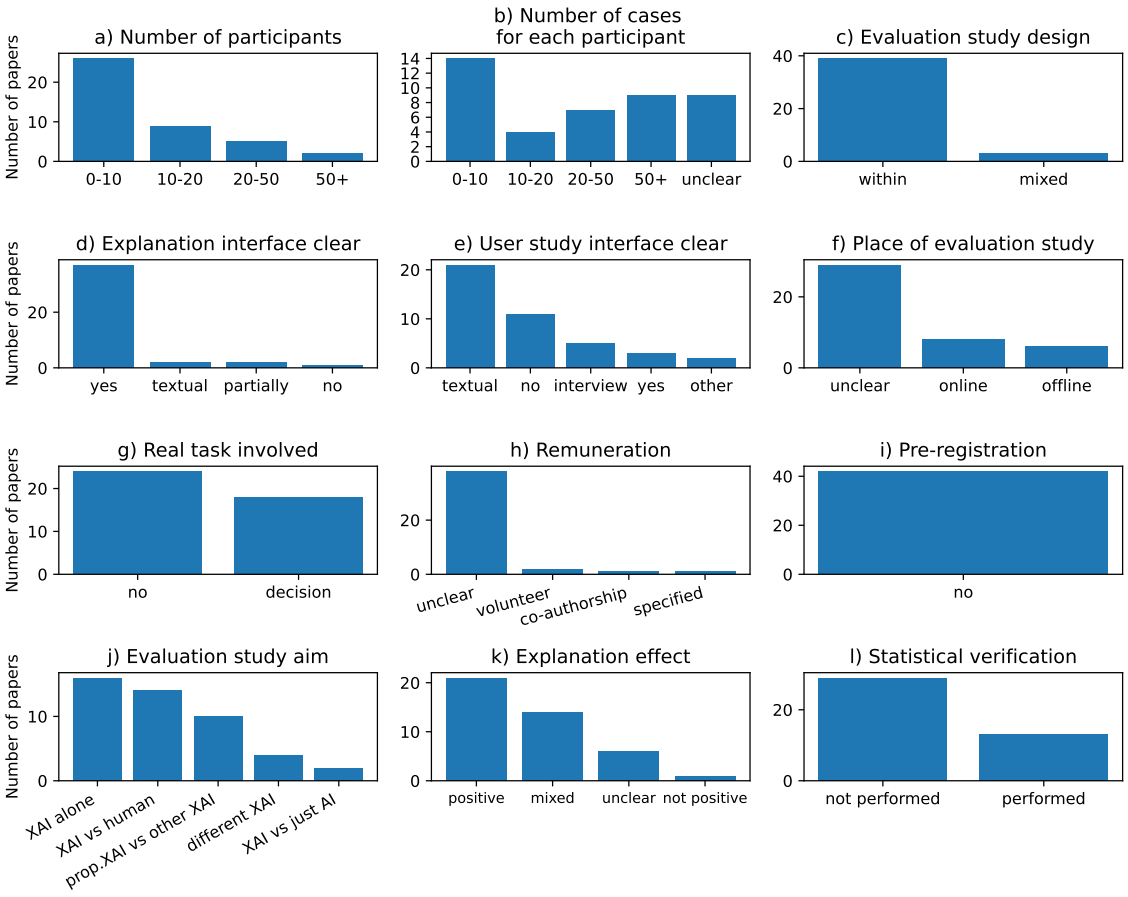

Figure 3: General statistics of the evaluation studies performed in the papers included in the survey.

Thus we consider 42 papers to perform further analysis. Refer to Figure 1 for the literature selection diagram.

4.2. General statistics of AI and XAI techniques

The table with the details of analyzed paper is available in our repository.[1] We present the main statistics corresponding to publication year, AI model, and XAI techniques in Figure 2. Figure 2a confirms the idea that the trend in XAI research started in 2018 and within the last 4 years (2020-2023), interest in XAI applications in the medical domain has been growing. Figure 2b shows that images, tabular data, and time series are frequently used for XAI applications. Deep Neural Networks are the most frequent choice of the models to be explained (see Figure 2c), which is natural due to the big part of image data indicated in Figure 2b and also to the general interest in Neural Networks and, in most cases, their better performance than tree ensembles or other AI model types. Figure 2d shows that XAI techniques are most frequently used for classification tasks. In some cases, classification is accompanied by a segmentation task [62, 63].Combination of DNNs for image data classification naturally yields frequent usage of post-hoc explanation generating heatmaps (e.g., GRAD-CAM) [64, 65].Even though post-hoc methods seem more popular, ante-hoc methods are also used quite often (e.g., DNN architectures combine classification and segmentation prediction [62, 63]). Apart from explanation in the form of heatmap textual and feature importance explanations are also used pretty often, especially with tabular data [66, 67].In this survey, all papers' XAI tasks were outcome explanations, which is natural in terms of the scope of the studied papers (i.e. when the medical professional makes the XAI-assisted decision, it is normally related to a certain case rather than, e.g., the whole model).

[1] Will be available upon acceptance

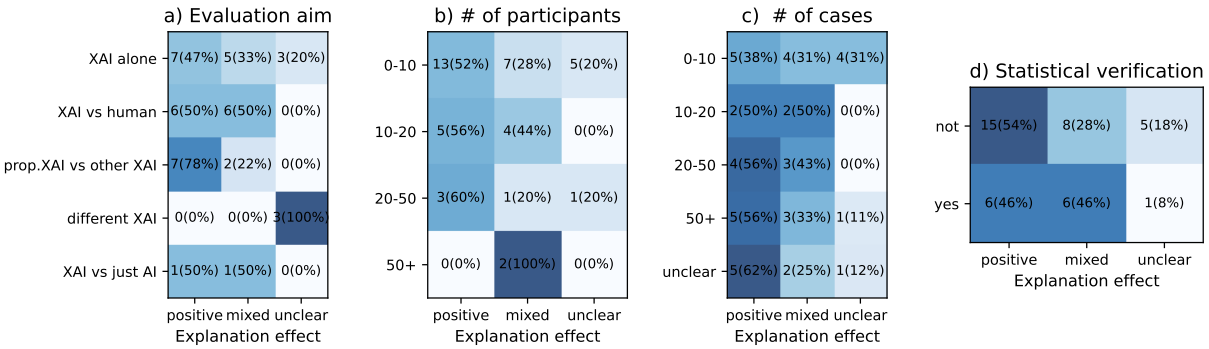

Figure 4: Analysis of the effect of application XAI technologies reported in the studied papers w.r.t. the aim of the evaluation study, number of participants, number of cases demonstrated to each participant, and statistical verification of the collected evaluation results. We disregard one paper with a "non-positive" effect on the compactness of visualization. The color map and percentage are visualized w.r.t. the share in the sum of values in each row of the matrix.

4.3. Statistics of evaluation studies

Figure 3 shows the general statistics of evaluation studies performed in the papers included in the survey. From Figures 3a,b we can see that most evaluation studies are prone to minimizing the number of participants and total cases shown to them - most papers involve up to 10 specialists who are demonstrated with up to 10 cases each. Figure 3c shows that the most prevalent evaluation setup is within-subject which assumes demonstrating all conditions (e.g., all XAI techniques) to each participant. Overall, these insights correspond to the general knowledge of the limited availability of medical professionals due to their high load on their work.

Figures 3d-i analyze the clarity of the evaluation study setup reported in the papers. Figure 3d shows that in the majority of cases, the interface of the explanation demonstrated to the medical practitioners is clearly defined, normally by means of screenshots (37 papers), but sometimes it is described either textually or shown only partially (2 papers for each case). However, the exact interface of human studies is normally described just textually (21 papers). Figure 3f shows that in most papers, the place of evaluation study performance is unclear (29 papers), whereas online and offline studies take place with almost equal frequency (6 and 8 papers correspondingly). Figure 3g shows that involving real decisions about certain patient data cases occurs less frequently (18 papers), than evaluation by proxy tasks, such as asking the participants about the perceived quality of the explanation (24 papers). Figure 3h shows that in most cases, the remuneration of participants is not discussed directly (38 papers). In rare cases, when it is discussed, it may be clearly volunteer participation (2 papers), co-authorship in the paper or clearly specified sum (one paper for each case correspondingly). Figure 3i shows that the practice of evaluation studies pre-registration seems to not be known in the medical XAI community, so none of the papers performed it.

Figures 3j,k,l show the core properties of the analyzed evaluation studies. Figure 3j shows that many papers (16) analyze the XAI technique in isolation without having any alternatives (e.g., other multiple XAI techniques). However, alternative evaluation options, even though they are used a bit less frequently, also take place (e.g., a comparison of the performance of medical professionals completing tasks on their own with the performance aided by XAI is performed in 14 papers, and a comparison of the original XAI technique with existing ones is performed in 10). Figure 3k shows the self-reported effect of the application of XAI, and in most cases, it is either positive (21 papers) or mixed (14 papers). Finally, Figure 3l shows that in most cases, the statistical verification of evaluation study results is not performed (in 29 papers).

Figure 4 shows the analysis of the self-reported effect of the application of XAI technologies with certain details of the evaluation setup. Figure 4a shows assessing XAI technologies alone results in positive results of the evaluation in almost half of cases (47%). Comparison of human-only and human-XAI performance evenly results in positive and mixed effects, and comparison of the original XAI

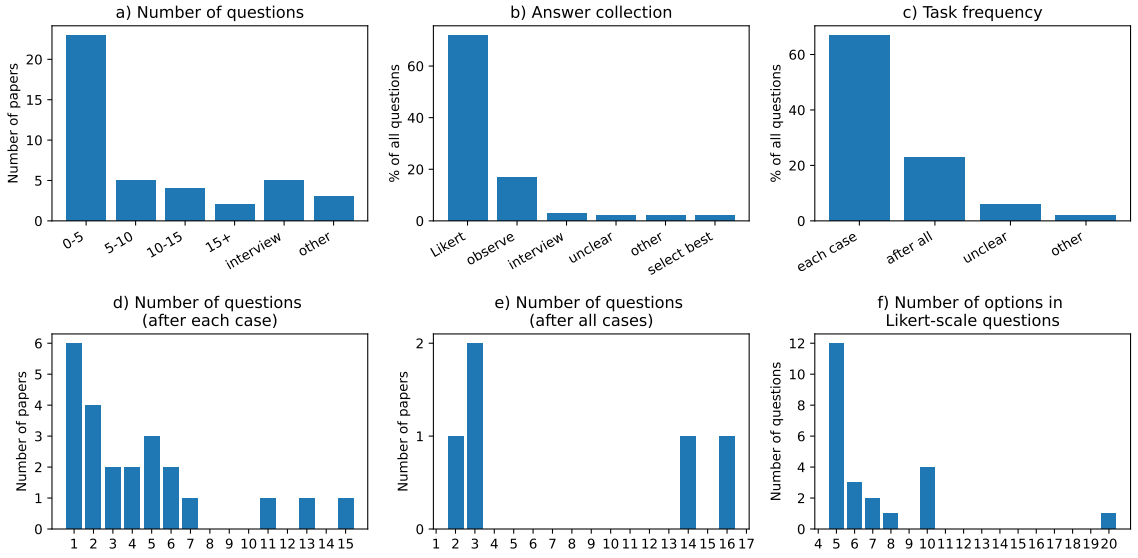

Figure 5: Statistics of the questions of the tasks involved in the evaluation studies.

technique in most cases is reported to have a positive effect (78%). Figure 4b visualizes the relation between a number of participants and the report effect of the XAI application. The tendency to report a positive effect is prevalent for all numbers of participants (i.e., for three of the four groups it is more than 50%). Figure 4c shows that the increase in the number of cases demonstrated to each participant increases the likelihood of the positive effect of XAI being reported by the authors, as far as it increases from 38% to 56%. Still, if the number of cases demonstrated by each evaluation study participant is unclear, the results are reported as positive with the biggest frequency (62%). Figure 4d shows that the evaluation studies without statistics analysis are more likely to have positive effects of XAI techniques application reported than those with statistical analysis (54% and 46% correspondingly).

Figure 5 shows the statistics of the tasks the evaluation studies consist of. In particular, in Figures 5a,b,c we can see that in most cases, the evaluation studies involve a small number of questions (up to 5) corresponding to each patient's case, asked using the Likert scale. Another relatively frequent choice of task answer collection is observation (17% of questions), which implies collecting the information through observing the participant's performing the task (e.g., comparing the participant performance with and without the assistance of XAI [68, 69, 70]). Whereas tasks are in most cases shown after each particular case, sometimes (23% of questions) they are shown after all cases. Figures 5d,e show the typical frequency of the questions depending on whether the questions are asked after each demonstrated case or after all cases. Figure 5f shows the distribution of several answer options for Likert scale-based questions. From these three Figures, we can see that in most studies, the participants get a relatively small number of questions, and if the questions are based on a Likert scale, the most typical number of options is 5.

We also analyze the exact properties of XAI technologies perception by medical professionals included in the evaluation. For further analysis, we rely on the taxonomy discussed in Section 3.3, that is why first of all we aim to analyze the coverage of this taxonomy w.r.t. the properties occurring in the papers (i.e., how many questions used in the papers' evaluation study can be related to one of the properties within this taxonomy and how many questions fall out of it). Figure 6 shows that in most cases (21 papers) 100% of questions included in evaluation studies of the papers included in our survey can be described with the proposed taxonomy. Note that we consider only the papers that provide clear questions. This means that the papers using, e.g., unstructured interviews of medical professionals are not considered in this analysis.

Figure 7 shows the detailed statistics of questions used in the evaluation studies w.r.t. our taxonomy of XAI perception properties. Figure 7a analyses the number of evaluated properties per paper. Most

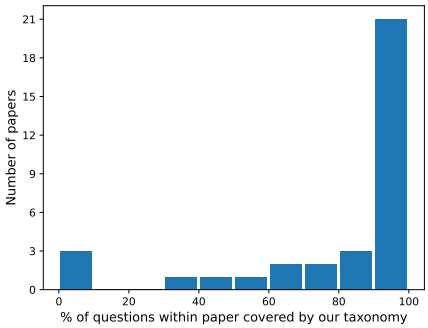

Figure 6: Analysis of the efficiency of XAI perception properties taxonomy used in our survey.

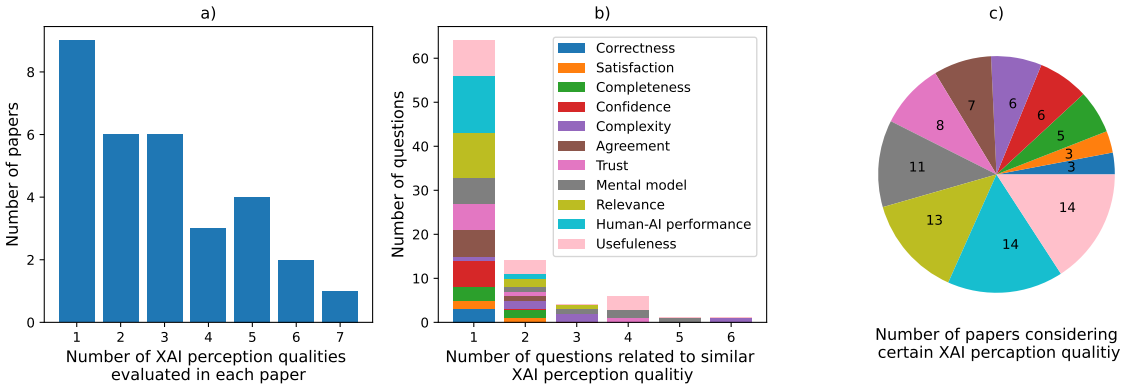

Figure 7: Detailed analysis of the tasks used in the evaluation studies.

papers evaluate from one to three properties. Figure 7b shows the statistics of the number of questions related to the similar XAI perception property asked within one evaluation study. In particular, this plot aims to show how often the medical practitioners participating in evaluation studies have to answer very similar or almost duplicated questions. The leftmost bar shows that in most cases, the duplicates do not take place, i.e., there is one question related to a certain property. However, in some cases, similar questions occur in the studies, resulting in two to four similar questions, which is particularly visible for usefulness (in the bars corresponding to two and four similar questions). Finally, Figure 7c shows that the most frequently evaluated properties are usefulness (14 papers), human-AI performance (14 papers), the clinical relevance of the provided explanation (13 papers), and mental model (11 papers).

5. Discussion

In this section, we discuss the general insights obtained from the studied papers.

5.1. Insufficient details of evaluation studies setups

Our analysis in Figure 3 shows that most researchers endeavor to conscientiously report on the number of participants and the explanation interface utilized in their studies. This practice facilitates a basic understanding of the user study framework and aids in assessing the study's scope and relevance to real-world applications. Despite this positive trend, a recurring issue is the lack of comprehensive details that are crucial for fully comprehending the user studies' methodology and, if needed, replicating them. Notably, the user study interface, particularly the description of tasks assigned to participants, is often articulated merely in verbal terms without any visual or demonstrative supplementation. This absence of a concrete demonstration may obscure the study's operational dynamics and the participant's interaction with the XAI system.

Furthermore, some studies occasionally omit the exact number of cases reviewed by each participant, a detail that is pivotal for evaluating the study's robustness and the reliability of its findings [63, 71]. Equally concerning is that some studies inadvertently impose a heavy burden on participants, presenting them with a rather large number of cases (e.g., over fifty patient cases). This could potentially lead to fatigue, affecting the quality of the feedback or decisions made by the medical practitioners involved [70, 72, 73]. Additionally, the rarity of testing the statistical significance of the results in these studies casts doubt on their validity and the generalizability of the conclusions drawn. Moreover, we also find that more than half of the evaluation studies do not involve making real decisions about the patient data, but instead rely only on proxy tasks (such as collecting the subjective feedback of medical practitioners about the explanation quality). Such tasks are valuable, but they do not always accurately reflect the complexities and pressures inherent in real clinical decision-making processes.

Moreover, finer details such as the place of the user study (whether conducted online or offline) and participant remuneration are seldom disclosed, despite their potential impact on participant engagement, fairness and the overall outcome of the study. The practice of pre-registration of evaluation studies (with online platforms like AsPredicted [58]), which can lend additional credibility and transparency to the research, appears to be largely overlooked in the field. This oversight suggests a gap in the adherence to rigorous scientific protocols, which, if addressed, could significantly enhance the quality and reproducibility of research in XAI applications in healthcare.

5.2. Poorly justified positive results

We also find a predominant trend among researchers to report positive outcomes from the application of XAI technologies in medical tasks (Figure 3k). While these optimistic conclusions might initially seem encouraging, they underscore the necessity for rigorous validation. Specifically, a common approach in evaluation studies is to assess the effectiveness of an XAI technology in isolation rather than in comparison with other XAI solutions or even against the baseline performance of medical practitioners without XAI assistance (Figure 3j). This methodology, despite yielding favorable results in nearly half of the cases (Figure 4a), raises concerns regarding the reliability of such outcomes. Without a benchmark or competitive framework, gauging the true impact and advancement offered by a specific XAI application becomes challenging, potentially overestimating its value in the absence of a comparative analysis.

Moreover, the rarity of testing for statistical significance further complicates the assessment of these studies' reliability. In instances where feedback from medical practitioners is garnered through unstructured or semi-structured formats, such as interviews [64, 74] or group discussions [75], the feasibility of applying statistical tests to validate the results diminishes. However, the absence of statistical verification, regardless of the reason, is noteworthy. Our analysis identified that over half of the studies reporting positive impacts of XAI evaluations failed to substantiate their claims with statistical significance tests. This gap not only casts doubt on the asserted benefits of XAI technologies but also highlights a critical need for more stringent methodological standards in research. The practice of affirming positive findings without robust statistical evidence can lead to a skewed understanding of XAI's efficacy, underscoring the imperative for more rigorous and quantitatively verified research to truly ascertain the effectiveness of XAI interventions in healthcare.

5.3. XAI perception properties

The analysis performed in Figure 7 showcases a broad spectrum of properties related to XAI perception among medical practitioners addressed in the evaluation studies. While there's a notable inclination towards assessing universally intuitive properties like usefulness, human-AI performance, clinical relevance, and mental model or understandability, it's imperative to recognize the importance of a more extensive array of perception properties. This includes, but is not limited to, the medical practitioner's confidence in the decision made with the help of XAI, the complexity of understanding the delivered explanation, and overall satisfaction with the XAI technology. The existing skew might limit our understanding of XAI's multifaceted impact on healthcare professionals. Addressing this issue requires

a more structured approach, potentially through the adoption of existing taxonomies of XAI perception properties [38, 10]. These frameworks can guide researchers in comprehensively evaluating XAI tools, ensuring that a wider range of significant attributes is considered during user studies with medical practitioners. This not only enriches the depth of the evaluation but also aligns the development of XAI technologies with the nuanced needs of healthcare settings.

Moreover, while the formulation of questions in most evaluation studies is generally precise—aiming to avoid redundancy and ensure clarity—instances of question duplication are not unheard of. Figure 7b shows that in some cases, similar properties may be addressed with two to four different questions. In some cases, several questions related to a similar property may be justified by the particular setup of the task. For example, in [75] two statements related to the clinical relevance of the explanations are shown for scoring: "pertinency of the map with respect to the diagnosis considered correct " and "coherence between the map and the suggestion provided by the machine". In [63] the user satisfaction of different parts of the proposed system is analyzed with two different questions: "I enjoyed being in the virtual environment?" and "I enjoyed using the COVIRplatform?".

However, there still are numerous cases when multiple questions about similar properties seem redundant. For example, [76] asks about the clinical relevance of the explanation with two pretty similar questions: "The predicted result is reasonable" and "The explanation of the prediction is clear and reasonable". The same paper involves two questions about complexity which also seem pretty similar: "The application is user-friendly and easy to navigate" and "I believe I would require technical assistance to use this system effectively". [63] asked participants about the complexity of the system addressing three questions: "Is it easy to learn to use it?", "I learned to use it quickly? ", and "I easily remember how to use it?".

The redundancy of these and some other questions not only affects the efficiency of data collection but also unnecessarily increases the cognitive load on participants. Ensuring the uniqueness and relevance of each question asked during a user study is crucial for maximizing the utility of the feedback collected and enhancing the overall efficiency of the study. As such, closer attention to the design of the evaluation study setup is warranted. By meticulously crafting the questions and avoiding overlaps, researchers can not only streamline the evaluation process but also foster a more engaging and less burdensome experience for medical practitioners participating in the study. This careful approach to study design underscores the importance of thorough preparation in achieving meaningful and reliable insights into the perception of XAI applications in healthcare.

6. Recommendation

Joining the existing XAI user evaluation study recommendations [59, 44], our findings from this survey and our experience in running similar evaluation studies, in this section we make the guideline for the XAI specialists aiming to evaluate the XAI techniques with medical practitioners.

6.1. Before the user study

Set a reasonable number of cases and questions

The questions asked to participants (normally related to the subjective perception of XAI) should be optimally selected. In particular, it is recommended to maximize the range of various XAI perception properties addressed by these questions, avoiding asking multiple questions related to a similar property. It may be particularly useful to learn about the commonly known XAI perception properties from the existing literature [38, 10, 39, 17, 13, 40].

At the same time, the cognitive load and the potential loss of interest of the participant, if the number of tasks included in the evaluation study is unreasonably high, must be taken into account. Therefore, it is necessary to select only those XAI perception properties that seem to be most relevant to a certain application in the particular study and also include a limited amount of patients that, at the same time, cover a maximal range of potential states (e.g., all possible diagnoses within the framework of the disease being studied are represented evenly in the sample of demonstrated patients).

Evaluate objective and subjective properties

The applicability of XAI in the medical field may equally depend on both subjective and objective factors (they can also be referred to as self-reported measures and behavioral measures, respectively [44]). Whereas subjective factors are normally studied by asking questions regarding personal perceptions of various aspects of XAI quality, the most frequent choice for measuring the objective quality of XAI, which still requires medical practitioners, is evaluating human-AI performance by asking to diagnose the patient first without and then with the help of the XAI technique. The cases demonstrated within these two groups could be both different and similar. Whereas in different cases (i.e., a medical practitioner diagnoses one group of patients with the help of XAI different group), the evaluation study may be performed within a single session, in the case of similar patients, it is recommended to make a break between evaluation sessions to prevent the fact that the doctor may stick to a previously made decision about the patient.

Objective properties may also be human-agnostic, and they are also worth being reported. For example, [38] discusses such objective properties as compactness (the size of the explanation) and confidence (the presence and accuracy of probability information in the explanation), and [77] proposes an objective metric named Degree of Explainability, which is directly proportional to the number of relevant questions that a piece of information can correctly answer.

Balance between real and proxy tasks

While subjective metrics of XAI quality perception provide valuable information it can also not be neglected that proxy tasks and subjective measures can be misleading in evaluating XAI systems [41]. Therefore, it is recommended to include in the design of the task the real action normally performed by medical practitioners in terms of the task XAI technologies are aimed to assist. In most cases, such a task is making the diagnosis regarding the particular patient.

Select the experimental settings proportional to the number of participants

It is generally expected that it would be unfeasible to engage numerous medical practitioners with sufficient experience in the specific medical field to participate in the evaluation studies. That is why the most frequent experimental setting chosen for such an evaluation study is within-subjects, where each participant sequentially passes through all conditions and provides feedback. For example, in such a setup, the practitioners may be first exposed to the patients' cases without explanations and then with explanations.

However, if it is possible to engage the number of evaluation study participants sufficient to form at least two groups of reasonable size, one can use a between-subjects setting when one subject is only exposed to one condition, e.g., one group gives feedback about unexplained AI predictions, and another group gives such feedback about the predictions with the explanations.

Avoid positive bias

It is necessary to foresee the potential positive bias caused by experimental design. One of the "risky" experimental designs that is likely to yield positive bias is when the explanation is evaluated in isolation without any alternatives, such as a demonstration of the patients' cases without any explanation or with an explanation generated with alternative XAI techniques. Such a setup may be particularly critical if the participants are asked about a single XAI technique knowingly developed by the authors of a paper in the interview format (when the authors are actually the interviewers).

Overall, it is recommended to measure explanation effects implicitly rather than explicitly. When participants are not aware of the evaluation's purpose, their responses may be more genuine. When measuring understanding or similar constructs, the participant's explicit focus on the explanations may cause skewed results not present in a more realistic scenario when a medical practitioner is concentrated on solving the actual task rather than on assessing the quality of the explanation.

In general, the phenomena of positive bias is known in the literature as a significant challenge across various domains of research, including clinical trials and expert evaluations [78]. Studies have shown that the lack of blinding and the absence of alternative options can lead to overly optimistic outcomes due to cognitive biases. For instance, in clinical trials, both patient and expert evaluations are subject to this bias, highlighting the necessity of implementing blinding techniques to mitigate its effects [78, 79, 80]. Overall, regardless of the nature of the study (whether it is a real clinical study

involving patients taking some medicines or medical experts assessing the quality of XAI technology), the objective remains to ensure the integrity and objectivity of research findings. These strategies are crucial for minimizing performance and detection biases, thereby preventing the overestimation of treatment effects and maintaining the credibility of scientific inquiry across different fields of study.

Prepare the definite plan of the core process of user study

While the preparation of the content of the evaluation study is important, the preparation of the process of the study itself is also crucial. It may slightly vary depending on the place of study.

In the case of offline participation, it may be helpful to provide information about meeting points, things necessary to bring, ways of preparation for the study, and to also send a reminder one day before. The plans during the evaluation study may also include finer details like where the participants are supposed to leave their belongings and plans for unexpected situations (e.g., uncooperative participants or malfunction of the demonstrated system).

In the case of online participation [81, 82], especially if the link for the study is distributed publically (e.g., in professional communities), it may be good to design some basic questionary that will ensure that the participant has the relevant experience for the certain medical task. This may involve asking about the years of experience dealing with certain diseases and asking about the standard diagnostic procedure this practitioner makes relative to the disease. Moreover, given that it is difficult to foresee how much time a certain participant is ready to devote to the user study, it is recommended to use tools that allow for objectively controlling the real engagement in the proper answers to the tasks (e.g., measuring the time spent on each page).

If it is initially clear that the study could take a long time, it is also recommended to give an explicit option to leave the study at a certain moment (e.g., when half of the cases are shown). This may guarantee the better quality of the collected feedback, avoiding the answers collected from the participants who became tired or lost interest in the study.

For whatever setup, it is also necessary to prepare the proper consent to participate in the study that will thoroughly discuss the whole process of the anticipated study and remind about the right to leave the study at whatever moment. The tasks and questions delegated to the participants should be clear and should also avoid inadvertent cues.

Run pilot study with small group

Before running the study with all available participants the good practice could be to engage a small pilot group of participants ready to test the proposed XAI system in a more thorough way by not only answering the pre-defined questions and performing fixed tasks but also by participating in a post-study interview and/or performing the proposed tasks following a think-aloud protocol [83]. This preliminary feedback can potentially improve not only the user study but also the evaluated XAI technology. However, the participation of such medical experts in the final evaluation studies is not recommended, because they may have a positive bias toward the system.

Pre-register of the evaluation study

Pre-registration of evaluation studies with online platforms like AsPredicted [58] has recently become a common practice [84]. The pre-registration involves submitting a document detailing the planned study (e.g., measured variables and hypotheses, data exclusion criteria, and the number of samples that will be collected) online before initiating the data collection. The pre-registration can be evidence against the findings being a result of selective reporting or p-hacking [85] and thus strengthen the credibility of a study.

6.2. During the user study

Collect the intermediate feedback

If the preparation of the evaluation study is performed correctly, no specific actions are expected from the organizers of the study. However, it may be necessary to foresee unexpected problems in the design of the study, which is why it is important to give the option for participants to leave a free-form comment (either in the text field or in a verbal manner depending on the place of study) to quickly

analyze it to correct the error promptly or, in the worst case, interrupt the study if the identified error is critical.

Exclude irrelevant or poorly-performing participants

It may also be useful to automatically exclude the participant from the study under some conditions. This could be the inconsistency of medical qualifications (e.g., if the participant indicates that one never or rarely works with the studied disease) or if the average time spent on the page is below some manually selected minimal threshold. Another criterion for automatic termination of the participant could be using attention-checking tasks that could contain relatively easy-to-answer tasks, so incorrect answers to such tasks may suggest that the participant can no longer continue to provide valuable answers.

6.3. After the user study

Filter irrelevant answers

Each participant's entry should be analyzed in terms of relevance. If the preliminary information about the participant's experience with certain diseases was collected in a free-form text, the analysis of such answers may give serious grounds to believe that this participant has insufficient experience in diagnosing the disease being studied. Moreover, in the case of collecting the answers by asking multiple Likert scale-based questions, similar answers to such questions (i.e., all questions on the page receive an equal response value) may also be the reason for excluding specific or all responses from a participant. The excluded answers, however, should be open-sourced as well as the considered answers to increase the credibility of the study, making it possible to ensure that the answers have been excluded for data-cleanness reasons rather than for adjusting the answers to the initial hypothesis.

Run the statistical verification of the collected results

If the evaluation setup involves a comparison of various XAI techniques with each other or a comparison of human performance with and without XAI aid (which is in general a recommended setup), the statistical verification of such a comparison is crucial to making a scientifically correct conclusion based on the results of such a study. Typical solutions for comparison between distributions obtained under different conditions are ANOVA tests and t-tests. The analysis of the data collected by the Likert scale may involve nonparametric tests such as the paired Wilcoxon signed-rank test or the Kruskal-Wallis H test to avoid normality assumptions.

Sometimes, the analysis of the data may involve the aggregation of one XAI perception property obtained from different sources (e.g., objective and subjective measures of understanding are combined into one distribution). In such cases, it is necessary to verify that there is sufficient agreement between the aggregated distribution, which is necessary to assess the validity of this aggregation with reliability measures such as tau-equivalent reliability (also known as Cronbach's α).

Moreover, the agreement between numerous participants providing feedback about similar items could also be valuable for analysis. It is normally calculated using Cohan's Kappa and Fleiß's Kappa [86].

Report all essential details of the study

When preparing the report about the evaluation study, it is important to disclose all details, including, but not limited to the ones studied in our survey (see Section 3.3, and Figures 2 and 3): comprehensive information about the backbone AI model and corresponding XAI technique, all details about the user study, such as number of participants and information about their background, number of tasks delegated to each participant, precise formulation of the tasks, etc. Visually report the most important findings to make it easier to understand the general findings of the study.

7. Conclusion

Drawing from the survey of XAI techniques evaluated by medical practitioners in the medical domain, this paper has illuminated several critical areas for improvement within current evaluation practices. The investigation underscored the frequent omission of essential evaluation study details, such as user study interfaces, locations, and participant remuneration, in published reports. A significant

finding of our survey is the pervasive lack of statistical significance tests in the evaluation of XAI applications, which has contributed to unwarranted optimism about the efficacy of XAI in medical contexts. Moreover, the analysis highlighted the tendency to evaluate XAI systems in isolation without benchmarking them against other methods or technologies. This approach risks overstating the utility and clinical relevance of XAI without a thorough comparative analysis.

A notable concern identified is the skewed focus on certain perception attributes of XAI, such as its supposed usefulness and performance in conjunction with human operators, at the expense of exploring a broader spectrum of important attributes. Additionally, the survey pointed out the issue of poorly constructed queries for study participants, which complicates the evaluation process with redundant or overly similar questions.

In response to these findings, the paper presents a set of actionable recommendations for future researchers. These recommendations emphasize the importance of comprehensive and transparent reporting of evaluation studies, the necessity for statistical rigor in assessing the results, and the value of comparative analysis to truly gauge the effectiveness of XAI systems. Furthermore, the recommendations advocate for a more balanced evaluation of XAI attributes and call for clearer, more distinct questioning techniques in participant studies.

The ultimate goal of these recommendations is to refine the approach to evaluating XAI applications in healthcare, ensuring that future studies can provide a more accurate, meaningful assessment of these technologies' real-world utility. By addressing the identified deficiencies, the research community can move toward a more nuanced and effective integration of XAI in medical practice, enhancing the potential for these technologies to contribute positively to patient care and medical outcomes.

Acknowledgments

This paper is part of the R+D+i project TED2021-130295B-C33, funded by MCIN/AEI/10.13039/501100011033/ and by the "European Union NextGenerationEU/PRTR". This research also contributes to the projects PID2020-112623GB-I00 and PID2023-149549NB-I00 funded by MCIN/AEI/10.13039/501100011033/ and by ERDF A way of making Europe. The support of the Galician Ministry for Education, Universities and Professional Training and the "ERDF A way of making Europe" is also acknowledged through grants "Centro de investigación de Galicia accreditation 2024-2027 ED431G-2023/04" and "Reference Competitive Group accreditation 2022-2025 ED431C 2022/19"

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5 to check grammar and spelling.

References

[1] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, Y. Wang, Artificial intelligence in healthcare: past, present and future, Stroke and Vascular Neurology 2 (2017) 230 – 243. doi:10.1136/svn-2017-000101.

[2] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, J. Zhu, Explainable AI: A brief survey on history, research areas, approaches and challenges, in: Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8, Springer, 2019, pp. 563–574.

[3] U. Pawar, D. O'Shea, S. Rea, R. O'Reilly, Explainable AI in healthcare, 2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA) (2020) 1–2. doi:10.1109/CyberSA49311.2020.9139655.

[4] A. Holzinger, C. Biemann, C. S. Pattichis, D. B. Kell, What do we need to build explainable AI systems for the medical domain?, arXiv preprint arXiv:1712.09923 (2017).

[5] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, U. R. Acharya, Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022), Computer Methods and Programs in Biomedicine 226 (2022) 107161.
[6] K. Borys, Y. A. Schmitt, M. Nauta, C. Seifert, N. Krämer, C. M. Friedrich, F. Nensa, Explainable AI in medical imaging: An overview for clinical practitioners–saliency-based XAI approaches, European journal of radiology (2023) 110787.
[7] M. Ivanovic, S. Autexier, M. Kokkonidis, AI approaches in processing and using data in personalized medicine, in: European Conference on Advances in Databases and Information Systems, Springer, 2022, pp. 11–24.
[8] D. Saraswat, P. Bhattacharya, A. Verma, V. K. Prasad, S. Tanwar, G. Sharma, P. N. Bokoro, R. Sharma, Explainable AI for healthcare 5.0: opportunities and challenges, IEEE Access 10 (2022) 84486–84517.
[9] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).
[10] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for explainable AI: Challenges and prospects, arXiv preprint arXiv:1812.04608 (2018).
[11] J. Jung, H. Lee, H. Jung, H. Kim, Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review, Heliyon (2023).
[12] S. Mohseni, N. Zarei, E. D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable AI systems, ACM Trans. Interact. Intell. Syst. 11 (2021). URL: https://doi.org/10.1145/3387166. doi:10.1145/3387166.
[13] R. Visser, T. M. Peters, I. Scharlau, B. Hammer, Trust, distrust, and appropriate reliance in (x) AI: a survey of empirical evaluation of user trust, arXiv preprint arXiv:2312.02034 (2023).
[14] J. Zhou, A. H. Gandomi, F. Chen, A. Holzinger, Evaluating the quality of machine learning explanations: A survey on methods and metrics, Electronics 10 (2021) 593.
[15] D. V. Carvalho, E. M. Pereira, J. S. Cardoso, Machine learning interpretability: A survey on methods and metrics, Electronics 8 (2019) 832.
[16] E. Pietilä, P. A. Moreno-Sánchez, When an explanation is not enough: An overview of evaluation metrics of explainable AI systems in the healthcare domain, in: Mediterranean Conference on Medical and Biological Engineering and Computing, Springer, 2023, pp. 573–584.
[17] A. Salih, I. B. Galazzo, P. Gkontra, E. Rauseo, A. M. Lee, K. Lekadir, P. Radeva, S. Petersen, G. Menegaz, A review of evaluation approaches for explainable AI with applications in cardiology, Authorea Preprints (2023).
[18] F. Di Martino, F. Delmastro, Explainable AI for clinical and remote health applications: a survey on tabular and time series data, Artificial Intelligence Review 56 (2023) 5261–5315.
[19] G. Ras, N. Xie, M. van Gerven, D. Doran, Explainable deep learning: A field guide for the uninitiated, J. Artif. Int. Res. 73 (2022). doi:10.1613/jair.1.13200.
[20] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.
[21] M. T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?": Explaining the predictions of any classifier, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016). doi:10.1145/2939672.2939778.
[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.
[23] A. Sivaprasad, E. Reiter, N. Tintarev, N. Oren, Evaluation of human-understandability of global model explanations using decision tree, 2023. arXiv:2309.09917.
[24] S. Maruf, I. Zukerman, E. Reiter, G. Haffari, Explaining decision-tree predictions by addressing potential conflicts between predictions and plausible expectations, in: A. Belz, A. Fan, E. Reiter, Y. Sripada (Eds.), Proceedings of the 14th International Conference on Natural Language Generation, Association for Computational Linguistics, Aberdeen, Scotland, UK, 2021, pp. 114–127. URL:

https://aclanthology.org/2021.inlg-1.12.
[25] L. Schneider, B. Bischl, J. Thomas, Multi-objective optimization of performance and interpretability of tabular supervised machine learning models, Proceedings of the Genetic and Evolutionary Computation Conference (2023). doi:10.1145/3583131.3590380.
[26] M. Tutek, J. Šnajder, Toward practical usage of the attention mechanism as a tool for interpretability, IEEE access 10 (2022) 47011–47030.
[27] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence, Information Fusion 99 (2023) 101805. doi:https://doi.org/10.1016/j.inffus.2023.101805.
[28] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, A. Hussain, Interpreting black-box models: a review on explainable artificial intelligence, Cognitive Computation 16 (2024) 45–74.
[29] W. Saeed, C. Omlin, Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities, Knowledge-Based Systems 263 (2023) 110273.
[30] A. Saranya, R. Subhashini, A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends, Decision analytics journal (2023) 100230.
[31] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, B. Zhou, Trustworthy AI: From principles to practices, ACM Computing Surveys 55 (2023) 1–46.
[32] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, J. Jorge, Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications, Information Fusion 81 (2022) 59–83. doi:https://doi.org/10.1016/j.inffus.2021.11.003.
[33] R. González-Alday, E. García-Cuesta, C. A. Kulikowski, V. Maojo, A scoping review on the progress, applicability, and future of explainable artificial intelligence in medicine, Applied Sciences 13 (2023) 10778.
[34] Y. Zhang, Y. Weng, J. Lund, Applications of explainable artificial intelligence in diagnosis and surgery, Diagnostics 12 (2022) 237.
[35] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, P. Lambin, Transparency of deep neural networks for medical image analysis: A review of interpretability methods, Computers in Biology and Medicine 140 (2022) 105111. doi:https://doi.org/10.1016/j.compbiomed.2021.105111.
[36] M. Kim, H. Sohn, S. Choi, S. Kim, Requirements for trustworthy artificial intelligence and its application in healthcare, Healthcare Informatics Research 29 (2023) 315.
[37] H. Chen, C. Gomez, C.-M. Huang, M. Unberath, Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review, NPJ digital medicine 5 (2022) 156.
[38] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI, ACM Computing Surveys 55 (2023) 1–42.
[39] I. Donoso-Guzmán, J. Ooge, D. Parra, K. Verbert, Towards a comprehensive human-centred evaluation framework for explainable AI, in: World Conference on Explainable Artificial Intelligence, Springer, 2023, pp. 183–204.
[40] M. Chromik, M. Schuessler, A taxonomy for human subject evaluation of black-box explanations in XAI., Exss-atec@ iui 1 (2020).
[41] Z. Buçinca, P. Lin, K. Z. Gajos, E. L. Glassman, Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems, in: Proceedings of the 25th international conference on intelligent user interfaces, 2020, pp. 454–464.
[42] A. Rosenfeld, Better metrics for evaluating explainable artificial intelligence, in: Proceedings of the 20th international conference on autonomous agents and multiagent systems, 2021, pp. 45–50.
[43] X. Kong, S. Liu, L. Zhu, Toward human-centered XAI in practice: A survey, Machine Intelligence Research (2024) 1–31.
[44] J. van der Waa, E. Nieuwburg, A. Cremers, M. Neerincx, Evaluating XAI: A comparison of rule-based and example-based explanations, Artificial Intelligence 291 (2021) 103404. doi:https:

//doi.org/10.1016/j.artint.2020.103404.
[45] G. Yang, Q. Ye, J. Xia, Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond, Information Fusion 77 (2022) 29–52.
[46] W. Jin, X. Li, M. Fatehi, G. Hamarneh, Guidelines and evaluation of clinical explainable AI in medical image analysis, Medical Image Analysis 84 (2023) 102684. doi:https://doi.org/10.1016/j.media.2022.102684.
[47] M. Ghassemi, L. Oakden-Rayner, A. L. Beam, The false hope of current approaches to explainable artificial intelligence in health care, The Lancet Digital Health 3 (2021) e745–e750.
[48] K. S. Kacafírková, S. Polak, M. S. Smitt, S. A. Elprama, A. Jacobs, Trustworthy enough? evaluation of an AI decision support system for healthcare professionals, in: The World Conference on eXplainable Artificial Intelligence, CEUR Workshop Proceedings, 2023, pp. 7–11.
[49] J. Jung, H. Lee, H. Jung, H. Kim, Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review, Heliyon 9 (2023) e16110. doi:https://doi.org/10.1016/j.heliyon.2023.e16110.
[50] A. C. Tricco, E. Lillie, W. Zarin, K. K. O'Brien, H. Colquhoun, D. Levac, D. Moher, M. D. Peters, T. Horsley, L. Weeks, et al., PRISMA extension for scoping reviews (prisma-scr): checklist and explanation, Annals of internal medicine 169 (2018) 467–473.
[51] Scopus, ???? URL: https://www.scopus.com/search/form.uri?display=basic#basic, accessed: 2024-02-15.
[52] Web of Science, https://www.webofscience.com/wos, ???? Accessed: 2024-02-15.
[53] IEEE Xplore, https://ieeexplore.ieee.org/Xplore/home.jsp, ???? Accessed: 2024-02-15.
[54] ACM digital library, https://dl.acm.org/, ???? Accessed: 2024-02-15.
[55] Pubmed, https://pubmed.ncbi.nlm.nih.gov/, ???? Accessed: 2024-02-15.
[56] M. A. Kadir, A. Mosavi, D. Sonntag, Assessing XAI: Unveiling evaluation metrics for local explanation, taxonomies, key concepts, and practical applications (2023).
[57] D. Gunning, E. Vorm, Y. Wang, M. Turek, Darpa's explainable AI (XAI) program: A retrospective, Authorea Preprints (2021).
[58] ASPREDICTED, https://aspredicted.org/, ???? Accessed: 2024-02-15.
[59] Y. Rong, T. Leemann, T.-T. Nguyen, L. Fiedler, P. Qian, V. Unhelkar, T. Seidel, G. Kasneci, E. Kasneci, Towards human-centered explainable AI: A survey of user studies for model explanations, IEEE Transactions on Pattern Analysis and Machine Intelligence (2023).
[60] S. Mohseni, N. Zarei, E. D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable AI systems, ACM Transactions on Interactive Intelligent Systems (TiiS) 11 (2021) 1–45.
[61] U. Ehsan, S. Passi, Q. V. Liao, L. Chan, I. Lee, M. Muller, M. O. Riedl, et al., The who in explainable AI: How AI background shapes perceptions of AI explanations, arXiv preprint arXiv:2107.13509 (2021).
[62] R. Lindsey, A. Daluiski, S. Chopra, A. Lachapelle, M. Mozer, S. Sicular, D. Hanel, M. Gardner, A. Gupta, R. Hotchkiss, et al., Deep neural network improves fracture detection by clinicians, Proceedings of the National Academy of Sciences 115 (2018) 11591–11596.
[63] K. Amara, A. Aouf, H. Kennouche, A. O. Djekoune, N. Zenati, O. Kerdjidj, F. Ferguene, COVIR: A virtual rendering of a novel nn architecture o-net for covid-19 ct-scan automatic lung lesions segmentation, Computers & Graphics 104 (2022) 11–23.
[64] S. Jadhav, G. Deng, M. Zawin, A. E. Kaufman, COVID-view: Diagnosis of COVID-19 using chest CT, IEEE transactions on visualization and computer graphics 28 (2021) 227–237.
[65] T. Chanda, K. Hauser, S. Hobelsberger, T.-C. Bucher, C. N. Garcia, C. Wies, H. Kittler, P. Tschandl, C. Navarrete-Dechent, S. Podlipnik, et al., Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma, Nature Communications 15 (2024) 524.
[66] S. Chari, P. Acharya, D. M. Gruen, O. Zhang, E. K. Eyigoz, M. Ghalwash, O. Seneviratne, F. S. Saiz, P. Meyer, P. Chakraborty, et al., Informing clinical assessment by contextualizing post-hoc explanations of risk prediction models in type-2 diabetes, Artificial Intelligence in Medicine 137

(2023) 102498.
[67] K. Aliyeva, N. Mehdiyev, Uncertainty-aware multi-criteria decision analysis for evaluation of explainable artificial intelligence methods: A use case from the healthcare domain, Information Sciences 657 (2024) 119987.
[68] S. Domínguez-Rodríguez, H. Liz-López, A. Panizo-LLedot, Á. Ballesteros, R. Dagan, D. Greenberg, L. Gutiérrez, P. Rojo, E. Otheo, J. C. Galán, et al., Testing the performance, adequacy, and applicability of an artificial intelligence model for pediatric pneumonia diagnosis, Computer Methods and Programs in Biomedicine 242 (2023) 107765.
[69] I. Neves, D. Folgado, S. Santos, M. Barandas, A. Campagner, L. Ronzio, F. Cabitza, H. Gamboa, Interpretable heartbeat classification using local model-agnostic explanations on ecgs, Computers in Biology and Medicine 133 (2021) 104393.
[70] A. Kumar, R. Manikandan, U. Kose, D. Gupta, S. C. Satapathy, Doctor's dilemma: evaluating an explainable subtractive spatial lightweight convolutional neural network for brain tumor diagnosis, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 17 (2021) 1–26.
[71] M. Dabass, S. Vashisth, R. Vig, MTU: A multi-tasking u-net with hybrid convolutional learning and attention modules for cancer classification and gland segmentation in colon histopathological images, Computers in Biology and Medicine 150 (2022) 106095.
[72] J. Pedrosa, P. Sousa, J. Silva, A. M. Mendonça, A. Campilho, Lesion-based chest radiography image retrieval for explainability in pathology detection, in: Iberian Conference on Pattern Recognition and Image Analysis, Springer, 2022, pp. 81–94.
[73] Y. Xu, M. Hu, H. Liu, H. Yang, H. Wang, S. Lu, T. Liang, X. Li, M. Xu, L. Li, et al., A hierarchical deep learning approach with transparency and interpretability based on small samples for glaucoma diagnosis, NPJ digital medicine 4 (2021) 48.
[74] A. Pirovano, H. Heuberger, S. Berlemont, S. Ladjal, I. Bloch, Improving interpretability for computer-aided diagnosis tools on whole slide imaging with multiple instance learning and gradient-based explanations, in: Interpretable and Annotation-Efficient Learning for Medical Image Computing: Third International Workshop, iMIMIC 2020, Second International Workshop, MIL3ID 2020, and 5th International Workshop, LABELS 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 3, Springer, 2020, pp. 43–53.
[75] D. Slijepcevic, F. Horst, S. Lapuschkin, B. Horsak, A.-M. Raberger, A. Kranzl, W. Samek, C. Breiteneder, W. I. Schöllhorn, M. Zeppelzauer, Explaining machine learning models for clinical gait analysis, ACM Transactions on Computing for Healthcare (HEALTH) 3 (2021) 1–27.
[76] R. Hendawi, J. Li, S. Roy, et al., A mobile app that addresses interpretability challenges in machine learning–based diabetes predictions: Survey-based user study, JMIR Formative Research 7 (2023) e50328.
[77] F. Sovrano, F. Vitali, An objective metric for explainable AI: how and why to estimate the degree of explainability, Knowledge-Based Systems 278 (2023) 110866.
[78] P. Probst, S. Zaschke, P. Heger, J. C. Harnoss, F. J. Hüttner, A. L. Mihaljevic, P. Knebel, M. K. Diener, Evidence-based recommendations for blinding in surgical trials, Langenbeck's archives of surgery 404 (2019) 273–284.
[79] A. Hróbjartsson, F. Emanuelsson, A. S. Skou Thomsen, J. Hilden, S. Brorson, Bias due to lack of patient blinding in clinical trials. A systematic review of trials randomizing patients to blind and nonblind sub-studies, International Journal of Epidemiology 43 (2014) 1272–1283. doi:10.1093/ije/dyu115.
[80] I. Boutron, C. Estellat, L. Guittet, A. Dechartres, D. L. Sackett, A. Hróbjartsson, P. Ravaud, Methods of blinding in reports of randomized controlled trials assessing pharmacologic treatments: a systematic review, PLoS medicine 3 (2006) e425.
[81] B.-Y. Mo, S. Nuannimnoi, A. Baskoro, A. Khan, J. Ariesta Dwi Pratiwi, C.-Y. Huang, Clusteredshap: Faster gradientexplainer based on k-means clustering and selections of gradients in explaining 12-lead ecg classification model, in: Proceedings of the 13th International Conference on Advances in Information Technology, 2023, pp. 1–8.

[82] A. Katzmann, O. Taubmann, S. Ahmad, A. Mühlberg, M. Sühling, H.-M. Groß, Explaining clinical decision support systems in medical imaging using cycle-consistent activation maximization, Neurocomputing 458 (2021) 141–156.
[83] P. Afflerbach, Verbal reports and protocol analysis, in: Methods of literacy research, Routledge, 2001, pp. 97–114.
[84] J. P. Simmons, L. D. Nelson, U. Simonsohn, Pre-registration: Why and how, Journal of Consumer Psychology 31 (2021) 151–162. doi:https://doi.org/10.1002/jcpy.1208.
[85] U. Simonsohn, L. D. Nelson, J. P. Simmons, P-curve: a key to the file-drawer., Journal of experimental psychology: General 143 (2014) 534.
[86] J. L. Fleiss, Measuring nominal scale agreement among many raters., Psychological bulletin 76 (1971) 378.