Computers in Biology and Medicine 180 (2024) 108908

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Intelligent systems in healthcare: A systematic survey of explainable user interfaces

João Cálem [a,c,*], Catarina Moreira [b,c], Joaquim Jorge [a,c]

[a] Instituto Superior Técnico, Universidade de Lisboa, Portugal
[b] Data Science Institute, University of Technology Sydney, Australia
[c] INESC-ID, Portugal

ARTICLE INFO

Keywords:
Explainable artificial intelligence
Human machine teaming
Human–computer interaction
AI-in-the-loop
XUI

ABSTRACT

With radiology shortages affecting over half of the global population, the potential of artificial intelligence to revolutionize medical diagnosis and treatment is ever more important. However, lacking trust from medical professionals hinders the widespread adoption of AI models in health sciences. Explainable AI (XAI) aims to increase trust and understanding of *black box* models by identifying biases and providing transparent explanations. This is the first survey that explores explainable user interfaces (XUI) from a medical domain perspective, analysing the visualization and interaction methods employed in current medical XAI systems. We analysed 42 explainable interfaces following the PRISMA methodology, emphasizing the critical role of effectively conveying information to users as part of the explanation process. We contribute a taxonomy of interface design properties and identify five distinct clusters of research papers. Future research directions include contestability in medical decision support, counterfactual explanations for images, and leveraging Large Language Models to enhance XAI interfaces in healthcare.

Contents

* Corresponding author at: Instituto Superior Técnico, Universidade de Lisboa, Portugal.
E-mail address: joao.calem@tecnico.ulisboa.pt (J. Cálem).

https://doi.org/10.1016/j.compbiomed.2024.108908
Received 3 May 2024; Received in revised form 5 July 2024; Accepted 15 July 2024
Available online 26 July 2024

1. Introduction

The growing global disparities in radiology are expected to significantly impact over half of the world's population, resulting from a lack of trained professionals and limited resources [1]. While Deep Learning (DL) models have demonstrated high levels of accuracy in predictions [2], their opaque nature has led to concerns about their decision-making processes, particularly in safety-critical applications. Studies have shown that DL models may perpetuate biases and discrimination [3–5]. In medical applications, machine learning patterns may not align with expert annotations, suggesting that they may be based on spurious correlations rather than medical domain knowledge [6–8]. Explainable Artificial Intelligence (XAI) is a field that aims to develop systems that enable human users to understand and validate the decision-making processes of algorithms.

To foster user trust and encourage domain experts to use DL models, XAI research aims to develop methods for explaining model predictions. While the field has witnessed notable advancements in XAI algorithms and techniques [9–11], it is also important to explore the interface design properties of intelligent systems. The emerging field of explainable user interfaces (XUI) emphasizes the importance of effectively communicating XAI results to end users [12]. While some XUI design guidelines are applicable across domains, the unique user requirements of medical professionals necessitate a specific exploration of XUI design properties of medical systems. For example, preferences for different explanation types vary based on user familiarity with DL models [13], and some explanations are found to be uninformative to medical practitioners specifically [14]. To our knowledge, no systematic literature review has been conducted in the healthcare domain of the XUI.

The systematic literature review conducted in this paper follows the guidelines outlined in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement [15]. The rest of the paper is organized as follows: Section 2 discusses related surveys in the XAI literature. Section 3 outlines the PRISMA methodology used to select the 42 most relevant articles. Section 4 examines the interface design properties found in the surveyed interfaces. Section 5 presents a taxonomy of XAI interfaces, categorizing papers based on their output and interaction types. Section 6 summarizes the answers to our research questions. Section 7 suggests future research directions for developing effective XAI interfaces, and Section 8 concludes the paper.

2. Related work

Researchers have conducted a growing number of surveys on XAI literature. They introduce the concept of explainability in AI [16,17], identify methods and metrics [18], and bring insights from fields outside of computer science [10]. While many surveys in this area tackle the technical aspects of explainable algorithms, human-centred design considerations are also a clear focus of the XAI field. Mohseni et al. [19] categorize XAI design goals and evaluation methods, mapping these goals and methods for different user groups. The paper also offers step-by-step design guidelines and summarizes ready-to-use evaluation methods for different goals in XAI research.

The intersection of XAI with the information visualization field underscores the critical importance of how explanations generated by intelligent systems are effectively communicated to end-users. Surveys exploring this intersection often delve into visual analytics in a broad sense [20,21]. While these surveys shed light on the general principles of visual communication, they often fail to address how the unique characteristics and requirements of different application areas may influence the design and efficacy of visualization strategies for conveying AI interpretability.

The field of XUI also explores interface design with a focus on user interaction. Chromik and Butz [12] review XAI literature from an interaction design perspective, categorizing tools by type of interaction and the interaction goal. They provide a structured framework for understanding the design space of XUI, thereby emphasizing the variability and adaptability of interface solutions in accommodating diverse user needs and preferences within the realm of XAI.

XAI properties can have significant differences across different domains [22]. Health sciences are a common area where XAI research is applied, and researchers conducted literature reviews with this particular focus [23]. Petch et al. [24] aimed to categorize explainable models in health sciences and identify challenges faced in this field. Albahri et al.[25] investigate XAI through a bias risk perspective. Nazar et al. [26] explore applying interpretable models to medicine with a user-centred focus. A major problem explored in this survey was the lack of domain-specific expert knowledge used to train models. Models trained this way are more susceptible to biases, less trustworthy to experts, and incapable of providing high-level explainability. Especially relevant to our work is the identified challenge of effectively communicating explanations to the domain experts.

While researchers have conducted surveys of XAI in the medical domain, none specifically addressed interface design. To the best of our knowledge, none of the existing literature reviews of XAI focus on both interactive design and the medical domain. Narrowing the scope this way allowed us to make a more comprehensive literature review and extract insights that can more effectively empower health professionals.

3. Methodology

This systematic literature review follows the PRISMA guidelines [15]. The study consists of three phases: defining research questions, identifying articles from scientific databases, and screening the initial search result based on inclusion and exclusion criteria.

3.1. Research questions

As we transition from the broader context to a more focused examination, exploring the three research questions below becomes pivotal in analysing explainable interface design trends in the existing literature. These questions dissect the output modalities, user interaction styles, and desirable properties of XAI systems, ultimately contributing to developing insights for effective design and implementation in medical decision support.

RQ1: What output modalities are found in existing XAI interfaces?

Different XAI algorithms produce explanations in varying formats, such as charts, networks, and textual explanations. Investigating these formats is important to identify trends and preferences among researchers. Moreover, the nature of the output greatly influences the comprehensibility and interpretability of AI-driven decisions [12,27]. This research question aims to uncover patterns in the presentation of outputs that have been most effective in aiding explainability and other desirable properties.

RQ2: What user interaction styles and methods are used in explainable interfaces?

User interaction plays an important role in the usability and effectiveness of XAI interfaces. As these interfaces aim to facilitate interaction between AI-generated insights and human understanding, the interaction methods must be carefully designed to facilitate effective knowledge transfer. By identifying successful interaction techniques, this paper aims to contribute to the development of best practices for interface design that encourage deeper engagement, improved comprehension, and enhanced decision-making.

RQ3: How can the properties of explainable interfaces be tailored to prioritize human-centric design, promoting interactivity and accessibility in medical domain tasks?

Designing interfaces that effectively support clinical decision-making requires a thorough understanding of the needs and requirements of healthcare professionals. Specifically, this question will focus on identifying how XAI algorithms have been integrated into interfaces in a way that prioritizes the human-centred approach to clinical decision support or model development.

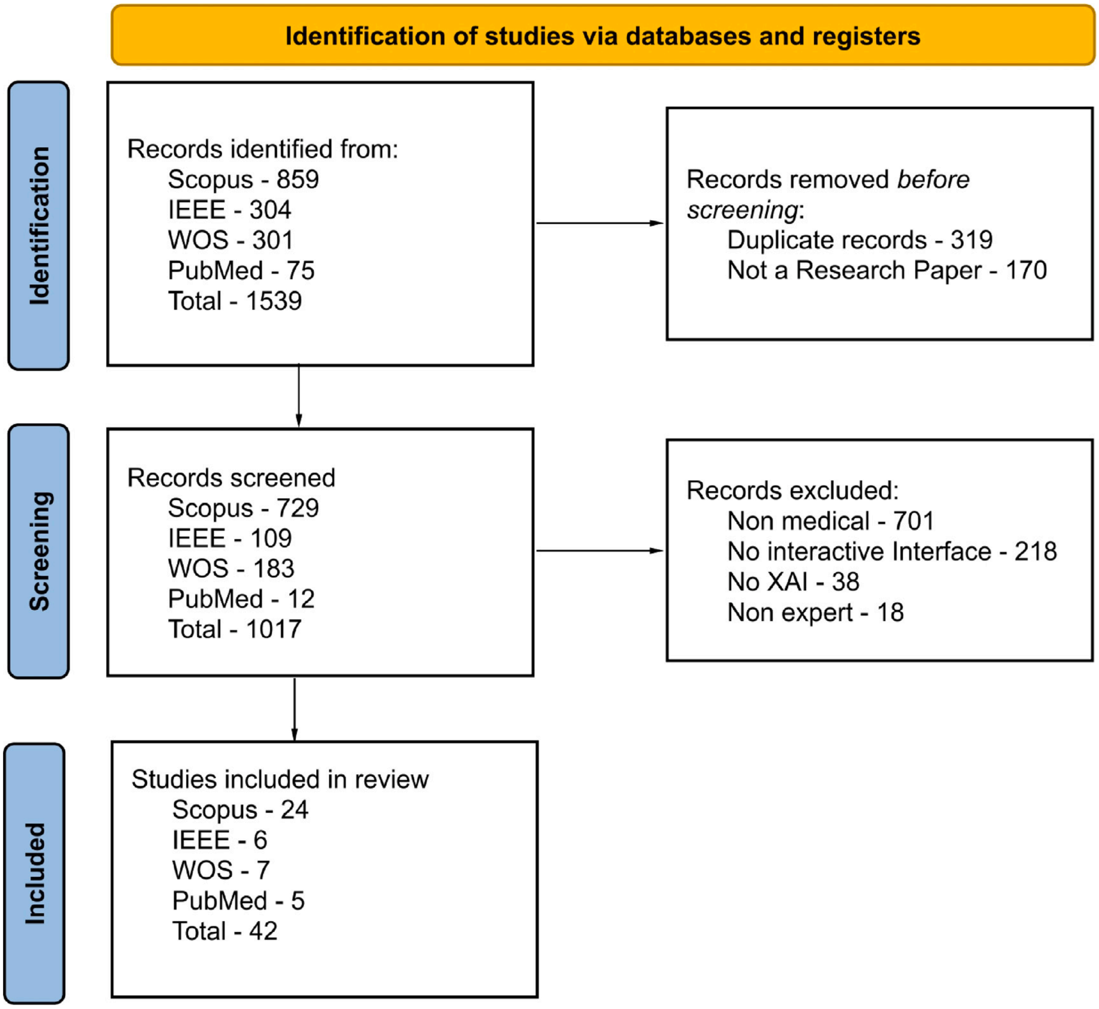


Fig. 1. Identification of studies via databases.

3.2. Initial search strategy

Our search strategy aimed to collect scholarly articles on XAI interfaces. We achieved this by constructing search queries that combined two themes: papers related to explainable artificial intelligence and those relating to interactivity, such as visual analytics, user interfaces, and dialogue. We utilized several APIs, including Scopus, IEEE, WOS, and PubMed, to search for papers based on their titles, abstracts, and author keywords.

Our final query is shown below. We iteratively refined the query, ensuring that important papers remained in the search results while capturing new results related to our research questions. We made adjustments such as shortening words like `dialogue` to dialo* and using the W/and PRE/ operators to limit the distance between keywords or groups of keywords. These strategies helped us reduce the number of irrelevant papers while preserving relevant ones.

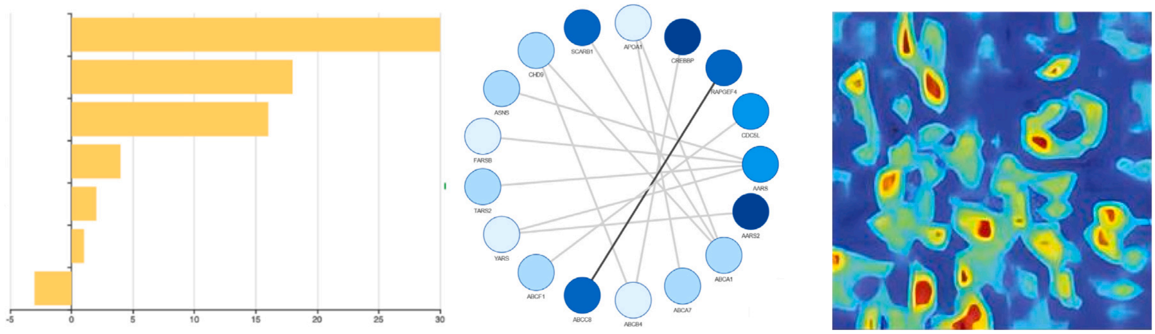
```
( xai OR counterfact* OR ( ( explain* OR inter-
pret* OR interactive ) W/1 ( ai OR "artificial
intelligence" OR "machine learning" OR "deep
learning" ) )

                    AND

(( user OR visual* ) PRE/ ( interface* OR
analytic* )) OR dialo* OR ( user PRE/ cent* ))
```

We faced limitations with some of the APIs that, for example, did not support the equivalent operators to $W/$ and $PRE/$. We used less strict strategies to avoid missing important papers in these cases. For instance, instead of using the $W/$ and $PRE/$ operators, we used the operator AND.

3.3. Inclusion and exclusion criteria

We used the following inclusion criteria when selecting papers:

1. **Focus on interactivity or user interfaces**. The first selection criterion is important to ensure that the papers chosen for the survey provide relevant insights into the design and evaluation of interfaces. This survey can gain insights into the key design principles and evaluation methods used to develop effective interfaces for clinical decision support by prioritizing interactivity and user interfaces.
2. **Applicable to medical diagnosis**. Although XAI algorithms and techniques from other domains can potentially be applied to the medical domain, we argue that valuable insights for medical interface design will mostly be found in intelligent systems applicable to medical tasks. Instead of including keywords relating to healthcare in our initial search, we decided only to limit this in our inclusion criteria. This is because some of the interfaces we surveyed were domain agnostic, but still potentially relevant to healthcare tasks. Domain-agnostic interfaces do not necessarily have any healthcare-related keywords in their title or abstract.



(a) Charts [58] (b) Networks [50] (c) Heat maps [53]

Fig. 2. Different modalities for the feature contribution output type.

3. **XAI algorithm must be used**. This inclusion criterion is important to ensure that the papers selected for the survey are focused on XAI-based interfaces and provide insights into the algorithms and techniques used in these systems. By identifying the XAI algorithms used in interactive tools for clinical decision-making, this survey can provide insights into the current state of the art and identify the most promising approaches for the development of effective XAI-based interfaces.

The following exclusion criteria were used when selecting papers:

1. **Focus on non-experts**. This criterion is relevant to exclude papers focusing on non-experts such as patients or the general public. By focusing on papers that apply to healthcare professionals, this survey can provide insights directly relevant to the target audience and help improve the development of interfaces that can better support healthcare professionals in their clinical decision-making processes.
2. **Non-Medical Use Cases**. A significant number of papers were excluded due to this criteria. Any paper that was specifically focused on an area that was not relevant to medical diagnosis was excluded. These included engineering, automation, education, economics, politics, and many other research domains irrelevant to our research questions.

3.4. Results and limitations

Results obtained in March of 2024 led to 1539 papers in the initial search, as shown in Fig. 1. The manual process using the inclusion and exclusion criteria narrowed these results to 42 explainable interfaces relevant to answering our research questions. Our search strategy was designed to be comprehensive, including as many relevant papers as possible. However, due to this survey's focus on the medical domain, the search results were narrowed down extensively. See Table 2 in Section 5 for a summary with all the references to the 42 papers we surveyed.

Like any human-driven task, our process of finding relevant research is influenced by biases. This review recognizes that sticking to just four databases (Scopus, Web of Science, PubMed, and IEEE) may have caused us to miss some articles. We could have broadened our search by including Google Scholar and SpringerLink. We also chose not to extract references from collected papers to enhance our results since our search approach was already extensive without them. The search query focused on keywords related to our papers of interest, but this approach might have limited our search and caused us to overlook relevant articles.

4. Explainable interface design

Building upon the groundwork established through our research questions, search strategy, and selection criteria, this section delves into the specific intricacies of explainable interfaces. The objective is to systematically examine how these interfaces convey information and facilitate user interaction. By exploring diverse explanation formats and interaction methodologies, we aim to discern patterns and extract insights into interface desirable properties.

4.1. Information output

User interfaces encompass diverse goals, each employing distinctive approaches to convey information effectively. Our analysis of 42 interfaces revealed recurring patterns in the presentation of information across multiple articles, outlined throughout this section.

Within human–computer interaction (HCI), the same information can be conveyed through various modalities. Bernsen proposed a taxonomy classifying output modalities in HCI into natural language (linguistic) or visual (analogue) categories [28]. In the XAI literature, linguistic outputs predominantly manifest as typed text in natural language. However, our comprehensive survey indicated that most examined works employed visual outputs. Relevant analogue modalities from Bernsen's taxonomy in the XAI literature include charts like bar plots and line plots, networks with nodes and edges, and images. Our investigation aims to determine the prevalence of these modalities in the literature concerning different types of information that interfaces commonly present.

Feature Contribution explanations are the most common in the interfaces we surveyed. These explanations aim to highlight which features in the data were most important for a given prediction. Algorithms such as SHAP [29], LIME [30] or LINDA [11] are examples of model-agnostic algorithms that extract feature contributions from input data. Moreover, contributions can also be shown for features learnt for a specific black box model, such as neuron activations in a neural network [31]. The majority of interfaces present this information using charts [31–34], primarily with bar plots [13,14,27,35–43] such as the one in Fig. 2(a). The same information has also been summarized in natural language text [27,38,44]. User interfaces displaying networks tend to show feature contributions with the thickness of graph edges [45–48], seen in Fig. 2(b). In image classification tasks, the contribution of individual pixels is commonly shown using heat maps in images as in Fig. 2(c). These methods aim to highlight important regions in an image based on how influential they are for the current prediction [14,34,43,49–54]. Despite feature contributions being the most common explanation type we found, it shows mixed results in user studies with medical domain experts. These explanations were commonly misunderstood by users without ML familiarity [13,40] or less preferred compared to other explanation types [40,41]. Experts

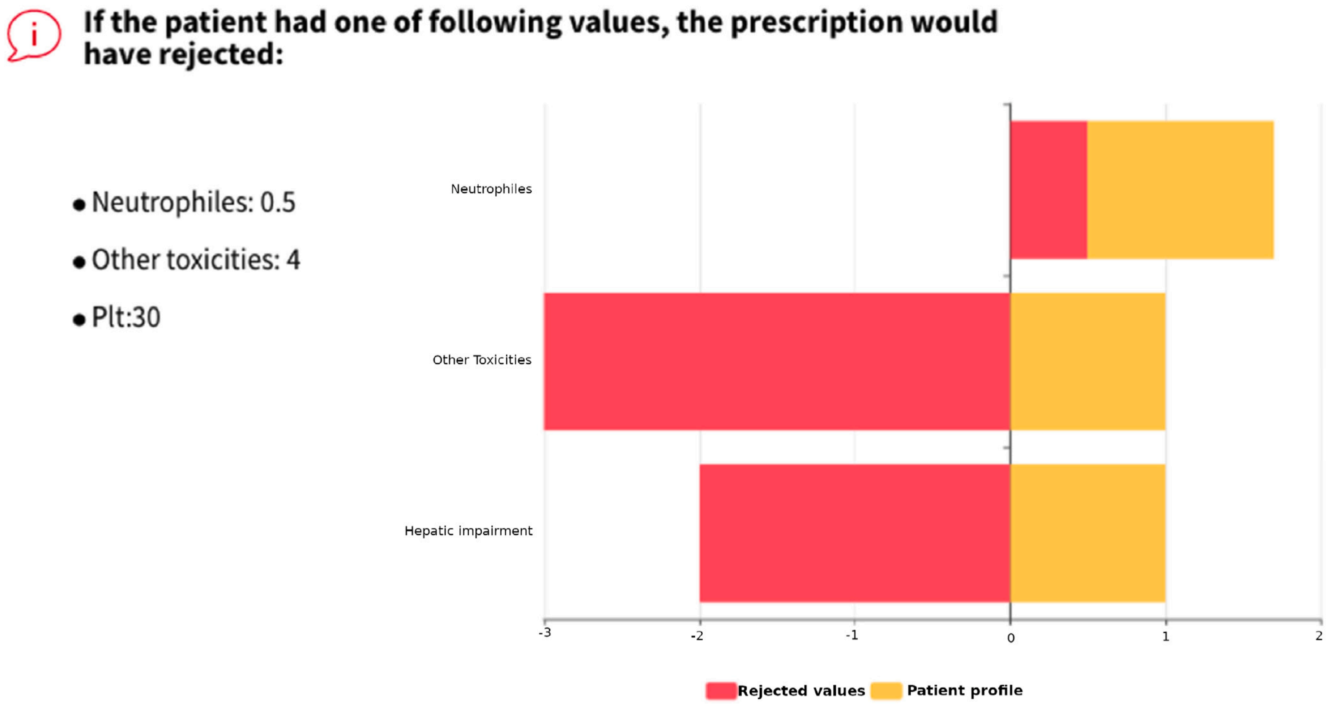


Fig. 3. Example of a counterfactual explanation using a bar plot [40].

also found feature contributions uninformative, particularly concerning heat map explanations [14]. However, they show greater potential for novice or junior-level users [43].

Partial Dependence shows how the prediction of a model changes across the range of up to two specific input features. For an individual instance, all other features are held constant. Users can visualize the potentially complex non-linear relationships between features and the model output. Since this information is calculated locally for a given instance, it may not capture the relationship in the global scope of the model. Averaging the results over the whole data set can capture this global scope and highlight if a feature's behaviour is inconsistent [55]. Partial Dependence line plots are generally used to display this information [27,32,35,44,55,56]. This explanation type was mostly found in the domain-agnostic articles of our survey and, therefore, not often explored in the context of health sciences. As with feature contribution explanations, we expect these explanations to be difficult to interpret for users without ML familiarity.

Counterfactual explanations aim to point out changes necessary to result in an alternative prediction. Rather than focusing on *why* a certain event occurs, counterfactual reasoning allows users to examine what changes need to occur to get the desired outcome [57]. According to Miller, the rationale behind presenting explanations contrastively is grounded in users' cognitive inclination to consider causal factors about alternative counterfactual events [10]. These counterfactual scenarios are commonly articulated as lists enumerating distinctions between two instances. These distinctions can be presented through textual descriptions [58], numerical values [41,55], natural language representations [13,59] or graphs [40,60]. Additionally, visualizing counterfactuals can be achieved through charts such as through bar charts as seen in Fig. 3 [40], or by employing arrows to signify changes in the values of specific features [61–63]. Counterfactuals are naturally interpretable by clinicians, even without ML familiarity [40]. Another application for counterfactuals is to provide customer recommendations to improve their outcomes such as with loan applications [62]. Yet, practitioners have shown concerns that these recommendations can be overgeneralized in the context of health care [13].

Rule-Based methods use IF-THEN-ELSE statements, allowing users to learn logical representations of a given prediction that can be generalized to any instance. This helps users gain a global understanding of how a predictive model makes its decisions. It is possible to represent a rule-based model visually, for example, using decision tries [64]. However, interfaces we found in our survey tend to display rules using natural language [65,66]. A prediction model can itself be a rule-based algorithm, which is therefore inherently explainable [65]. Alternatively, a rule-based algorithm can be learned using model induction, by approximating the results of a *black box* model [66]. While this approach is more flexible, with possible uses in more complex tasks, we have not found user studies with medical experts for the latter explanation type.

Neural Network Concepts are representations of features learnt from DL models by analysing the behaviour of a single neuron or group of neurons in a neural network. These are especially useful when working with unstructured data, such as for image classification tasks. Representations start from lower-level concepts such as shapes and textures in an image, that are combined to form more complex concepts at each layer of the model, eventually leading to the final prediction. Concepts are commonly represented in interfaces by displaying examples of image segments with high activation for the concept [45,46,67], but synthetic images can also be generated to represent a concept [37,46]. In the medical domain, we found it more common for concepts to be manually labelled by experts [68]. Such a system allows the integration of domain knowledge in the explanations provided, but poses a much higher burden on healthcare expert's time.

Feature Context refers to information on how features in data compare across a sub-population or data set. For example, the statistical distribution of features can communicate to users the mean value of a specific feature and its expected ranges. This context can enhance the explainability of the different outputs discussed previously. With feature contribution explanations, not only is it useful to know if features are important for a prediction, but also to know if feature values are relatively low or high or check if they are outliers in the data. Individual feature context is most commonly represented with charts such as histograms [32,36,61,62,66,67], scatter plots [63] and box plots [67], but can also be summarized through natural language [27,44]. Contextualizing an instance with its neighbours can also help with case-based reasoning. In medical applications, for example, it is common for experts to compare a patient with other similar cases to

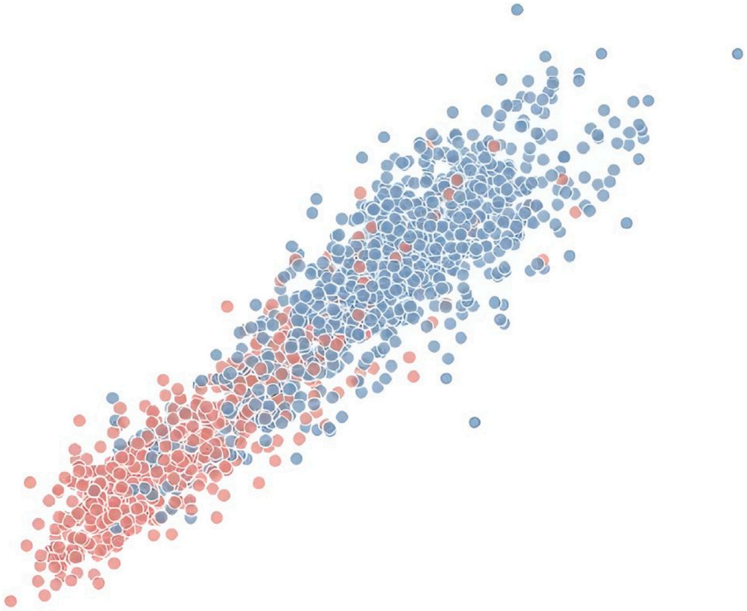


Fig. 4. Example of 2D projection of data where similar patients are positioned next to each other [33]. Patients are coloured differently depending on their diagnosis. Users can click on any point to explore each patient.

gain further insight or solve conflicting information [69]. A common interface design tool is to use an algorithm such as t-SNE [70] to project features into two dimensions for visualization purposes [31–33,36,39,43,45–47,55,71,72]. By using these two-dimensional projections, as in Fig. 4, users can visualize neighbouring instances to the one they are analysing.

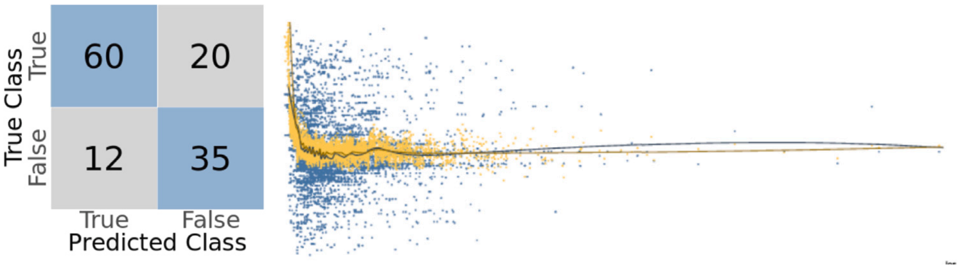
Model Performance refers to any performance metric calculated for the predictive model on a test set. Since most of the interfaces we surveyed were for classification tasks, the most common performance metrics were taken from confusion matrixes (Fig. 5(a)), showing misclassification through bar plots [31,55,58,61,62,66,67]. Many of these visualizations are interactive, letting users click on the different misclassified instances to investigate potential problems in a model. Evaluating a model using information extracted from XAI algorithms is also possible. For example, Li et al. develop a scatter plot visualization that helps users visualize the consistency of feature contribution explanations [71], seen in Fig. 5(b). A feature with low consistency has increased randomness in its feature contribution for different instances regardless of feature values. This is showcased in their charts when the points in the scatter plot are spread out and show more noise. Prince et al. [14] found that performance metrics are more informative for engineers developing or evaluating models, but are not as useful for medical experts.

To measure the accessibility of an interface, we identify its **Visual Complexity (VIS CPX)** on a three-point scale. We define low complexity as an interface with fewer than six possible unique visual outputs and whether less than four are shown simultaneously. If only one of these constraints were met, it was marked as medium complexity. These numbers were based on the mean values across the interfaces we surveyed. We also verify if more than one output type was used to communicate explanations in an interface. As argued by Chromik and Butz [12], using multiple modalities can enhance the explainability of an interface. Hence, even though an interface may have one main output type, interfaces should aim to have **Multi-Modal Explanations (MM EXP)**.

4.2. User interaction

While effective information communication remains a crucial component of explainable interfaces, how users engage with these interfaces holds equal significance. Drawing upon Shneiderman's taxonomy of user interaction types [73], we have identified three primary methods of interaction with XAI interfaces in the existing literature: direct manipulation, encompassing actions like clicking on visual elements, obtaining information on hover, and utilizing drag-and-drop functionality; indirect manipulation, involving the use of sliders, drop-down menus, and buttons; and lastly, natural language dialogue. Specifically, indirect manipulation involves interaction elements that are placed on a separate panel or menu. We have categorized user interaction styles with explainable interfaces into four distinct types, discussed below.

Exploration (EXP) is the most common user interaction style in the literature. Exploration is an interaction style that provides ways for a



(a) Confusion matrix [50] (b) Scatter plot measuring consistency of feature contributions [43]

Fig. 5. Examples of model performance visualizations.

user to navigate the information output of an interface from different perspectives to reach their own understanding and conclusions. Users should be able to engage in a meaningful conversation with the XAI system, pose questions, and receive relevant information in response. This approach, advocated by Miller [10], places interactivity and dialogue at the forefront of explanation delivery. However, it is important to note that a conversational explanation dialogue need not be confined to natural language. Alternative representations of information, such as utilizing interactive elements, diagrams, or other non-linguistic forms of communication, can be equally effective. We explore this interaction style in greater detail in Section 4.3.

Control (CTRL) is an interaction style in which the user can directly adjust the predictive model through the interface. This type of interaction is also referred to in the literature as human-in-the-loop or interactive machine learning [74]. In the medical domain, this control is important so that experts can align prediction models with their own domain knowledge. Literature exploring algorithms that can take human input is extensive, but we also found examples of this interaction style incorporated into explainable interfaces directly. Generally, through indirect manipulations, users can change the classification thresholds [55], the structure of a model [47], retrain a model with new feature weights [36] or fine-tune a model with corrected data [68].

Contestability (CNTS) allows users to disagree with outputs given by the explainable interface. To achieve this, the interface needs to give the user alternative predictions and explanations. However, this alone does not provide contestability. As argued by Vaccaro et al. [75], contestability is achieved through an appeal process where the user's *voice* is heard. This means that the interface requires active interaction mechanisms that can interpret the user's disagreement and solve it. Users are shown to desire contestability the higher the impact of a decision [76], meaning this interaction is increasingly important in the medical domain. However, none of the interfaces we have surveyed in this article implements interaction mechanisms matching these properties.

The **Supportive (SUP)** interaction style aims to integrate explainability into an expert's existing workflow. Such systems let users analyse data using their expert knowledge before showing prediction results from a DL model. Users can prompt the interface for predictions and explanations on an as-needed basis. Not only does this help shift the interface agency to the user, but it can also reduce potential biases such as anchoring and confirmation bias [59]. We found only two examples in our survey utilizing this interaction style [40,65], where predictions and explanations are designed to be integrated directly into clinician software.

To measure the accessibility of an interface, we identify their **Interaction Complexity (INT CPX)** on a three-point scale. We define medium complexity when the interface has five to seven unique interactive elements. Lower or higher than this range of values is marked as low or high complexity, respectively. These numbers were based on the mean values across the interfaces we surveyed.

4.3. Exploration

Since most of the interfaces we surveyed used the exploration interaction style, we further investigated common themes for types of information users can explore. While analysing each interface, we quantify in Table 2 in Section 5 how many of these themes are present in the interface.

Dynamic **Explanation Detail** is a design tool that researchers have implemented in many explainable interfaces. Through direct manipulation, users can hover their mouse over explanations to get further information [59,66,77,78], or click interactive elements to expand into more detailed explanations [43,66,67,67,68]. Other interfaces implement buttons [34,42,58,63,72] or sliders [38,46] to achieve similar results through indirect manipulation. Natural language dialogue is also an accessible way for users to ask for further details [44].

Table 1
Taxonomy of explainable interfaces based on output and interaction types.

ine Paper	Cluster	Output type	Interaction type
[33,34,37–39,58,61–63] [13,14,40–42,55,65,72]	Visual analytics	Charts	Indirect manipulation
[27,59,66,71,77–79] [31,32,36,43,67,68]	Interactive analytics	Charts	Direct manipulation
[50–54,65]	Image visualization	Images	Indirect manipulation
[45–48]	Networks	Networks	Direct manipulation
[44]	Natural language	Text	Dialogue

What-if analysis allows users to explore how manually changing input values affects the outcome of a prediction model. This allows users to probe a prediction model to explore how it reacts to different scenarios. Usually through indirect manipulation [14,33,36,41,48,55], users can edit features individually and visualize the resulting changes. For direct manipulation, Outcome Explorer is an example where users can interactively edit nodes in a structural graph to perform the what-if analysis [47].

Prototypes are especially important to explore in medical applications. This exploration style refers to investigating real instances that are similar to the one being analysed. It is also important to let users explore instances of different classes to mitigate representativeness bias [59]. Allowing users to click on elements in 2D projections of populations mentioned previously is a common way to add prototype exploration [31,36,43,55,68,71,72]. However, users can also get the top matches through indirect manipulation [14,40,41], search for prototypes through search boxes or drop-down menus [32,38] or ask the interface through natural language dialogue [44].

Sub-Population creation gives users control over how instances are contextualized or how to cluster datasets based on domain knowledge. This is usually achieved through menus using indirect manipulation [39,55,62,63,66,79]. One way to introduce this exploration style through direct manipulation is by using a lasso tool so that users can draw clusters of data in a 2D projection [32,36,43,68,71,72].

Understanding the underlying **Model Architecture** is crucial for users seeking in-depth insights. Interfaces that incorporate this exploration style allow users to delve into the specifics of the employed machine learning model. Users can inspect the model's layers, nodes, and connections to understand its structure comprehensively. Interfaces with visual model representations can add direct manipulation for this exploration style [31,32,45,46,66]. Drop-down menus or buttons are commonly used [34,37,47,49,52,67].

Explanation customization empowers users to modify the provided explanations' content, format, or level of detail. Tailoring explanations to meet user preferences or specific requirements is essential to many interfaces. This is usually achieved through indirect manipulation in the settings of the application [34,39,45,50,55,61,63,66,67].

5. Taxonomy of explainable interfaces

Based on our findings relating to information outputs and interaction types, we propose a taxonomy of explainable interfaces, dividing the existing literature into five clusters of papers. We based these clusters on two properties, as in Table 1. First, the main information output type was identified as charts, networks, images, and natural language text. Second, the main interaction type between direct manipulation, indirect manipulation and natural language dialogue is chosen. Many of these interfaces have multiple information and interaction types. We chose to identify only the main one to get clearly defined clusters in our taxonomy. Charts are the main output type used in the current literature. Moreover, few interfaces use images or natural language as their main output types. This contrasts the output types that medical professionals most commonly work with.

Table 2 summarizes our systematic review findings for each cluster, with example interfaces shown in Fig. 6. Visual and interactive analytics display information primarily through charts, differing only in whether the main interaction is through indirect or direct manipulation. These interfaces vary significantly in terms of visual and interaction complexities. The image visualization cluster only contains interfaces utilizing heat maps for image explanations, generally found in low-complexity interfaces. The applications in the network cluster also differ from the others due to their main output type. In our final cluster, we only found one interface using natural language dialogue in combination with XAI algorithms.

Domain-agnostic interfaces stand out in our visual analytics cluster. Spinner et al. [34] showcase how providing users with multiple choices of XAI algorithms enhances exploration by allowing users to customize the explanation outputs of the interfaces. Similarly, Context Sight [63] uses feature contribution and counterfactual explanation in the same interface. By presenting information using charts, both explanation types can be combined visually. Despite its interaction complexity, it presents much information in a single output panel. Through its model performance panel, the What-If tool [55] allows users to control the classification thresholds of the predictive model. As its name implies, the interface offers what-if analysis tools, resulting in a greater exploration interaction style. A similarly interactive was developed in the health care domain by Wang et al. [72]. Its predictive model is based on a graph neural network that provides path-based explanations, a less common explanation type that the paper argues resembles how medical experts explain phenomena to their peers.

Both Tarnowska et al. [65] and Naiseh et al. [40] developed tools that use the supportive interaction style. Even though they do not incorporate a lot of exploration or a large variety of explanation types, they are examples of how explainable interfaces can be integrated directly into an expert user's existing workflow. Tarnowska et al. [65] designed an interface for diagnosing tinnitus in patients and allows healthcare experts to record visits, diagnoses, and treatment plans. Naiseh et al. [40] incorporate and evaluate several different explanation types into a medical prescription system. These interfaces showcase how explainable interfaces can be used as a decision-support tool where predictive models and explanations are added bonus functionality.

There is a lack of interactivity in image visualization interfaces. All six interfaces from this cluster provide similar explanations: heatmaps that show the most important areas of an image for the current prediction. Yet very few of the interactivity properties we mentioned in our review are used in these interfaces. Including domain-agnostic interfaces in our comprehensive survey allowed for the identification of promising image classification systems. In the Interactive Analytics cluster, ConceptExplainer [35] showcases the neural network concept output type. As one of the most visually complex interfaces we surveyed, it does present several unique visual output types, such as an interactive hex chart that contextualizes the concepts it identifies in the DL model. NeuroCartography [45] and Summit [46] adapt this information type to Networks. Without losing a lot of exploration options, they drastically reduce the visual complexity of the interface by using this output modality.

Outcome-Explorer [47] is another example of an interface using the Network output modality. Unlike the previous two, this interface does not use an XAI algorithm. Instead, it uses a causality-based predictive model that is inherently explainable. Since users train the causal model directly in the interface, it is a clear example of the control user interaction style. After exploring the model structure and using the what-if analysis tool provided, expert users can change the structure of the predictive model to reflect their domain knowledge better. In the healthcare domain, RetainVis [36] is an example of how explainability can enhance the control interaction style. By analysing the feature contributions of different instances, users can retrain the DL model with alternate feature weights, resulting in a more aligned model with expert knowledge.

In the interactive analytics cluster, three healthcare interfaces offer simple interfaces with multi-modal explanations. This simplicity comes at the cost of very low user interaction. The interface developed by Panigutti et al. is a clear example of this trade-off [77]. The interface only presents the feature contribution output type through a scatter plot and natural language summary. It is an efficient method of communicating a single explanation method, but users cannot interact with the information meaningfully. The interface proposed by Barda et al. offers similar information outputs but adds tabular data and a time series chart so users can better visualize the input data [78]. Wang et al. offer a wider variety of explanation types by incorporating counterfactuals in natural language alongside feature contributions [59]. Nevertheless, all three interfaces are mostly static, where their only interactive mechanism displays extra explanation details by hovering over the different outputs.

On the other hand, tools such as RetainVis [36] forego visual simplicity to enable a higher degree of user interaction. Users can interactively create sub-populations and investigate instances of interest through their two-dimensional projection of the data set. The visual complexity adds to its learning curve, but the exploration interaction style shifts agency to the user. Rulematrix is another such example [66]. Through its rule-based explanations, this interface also lets users interactively probe the predictive model and get a better global understanding of its decision-making process. Even though its main interaction type is direct manipulation, a significant part of the user interaction is also through sliders and buttons. This makes the interface complex in terms of visual and interaction complexity.

Our Natural Language cluster stands out as having only one interface. The tool developed by Kusba and Biecek [44] presents visualizations such as partial dependence plots and feature contribution plots. However, it distinguishes itself from all the other interfaces we surveyed since its only interaction mechanism is through natural language dialogue. The paper's main contribution is training a natural language understanding model that assigns user prompts into a set of predefined question types. By only showing its different outputs as the users ask follow-up questions, the interface uses the progressive disclosure principle advocated by Chromik et al. [80]. This style of interaction helps keep the visual and interaction complexities low. Even though there is room for a much greater degree of user interaction styles, this interface shows the potential of using natural language dialogue in explainable interfaces.

Across all interfaces we surveyed, we found no standardized method of explainable interface evaluation. A few papers used the think-aloud protocol to gather expert user comments about an interface [59,67], while others did interviews with experts after they used the interfaces [36,65,78,79]. Proposed interface evaluation methodologies exist, such as the System Usability Scale [81], and others specifically developed for explainability such as the System Causability Scale [82]. It is a testament to the incipient nature of the field that standardized evaluation techniques have yet to be developed.

Our analysis shows that even if most of the interfaces we surveyed have user interactivity relevant to medical applications, overall use of interaction styles is limited. None of the interfaces we surveyed offer user contestability, and very few incorporate control or supportive interaction styles. For exploration, interfaces that offer multiple exploration types tend to have much higher visual and interaction complexities. This showcases the trade-off between simplicity and in-depth user interaction in explainable interface design.

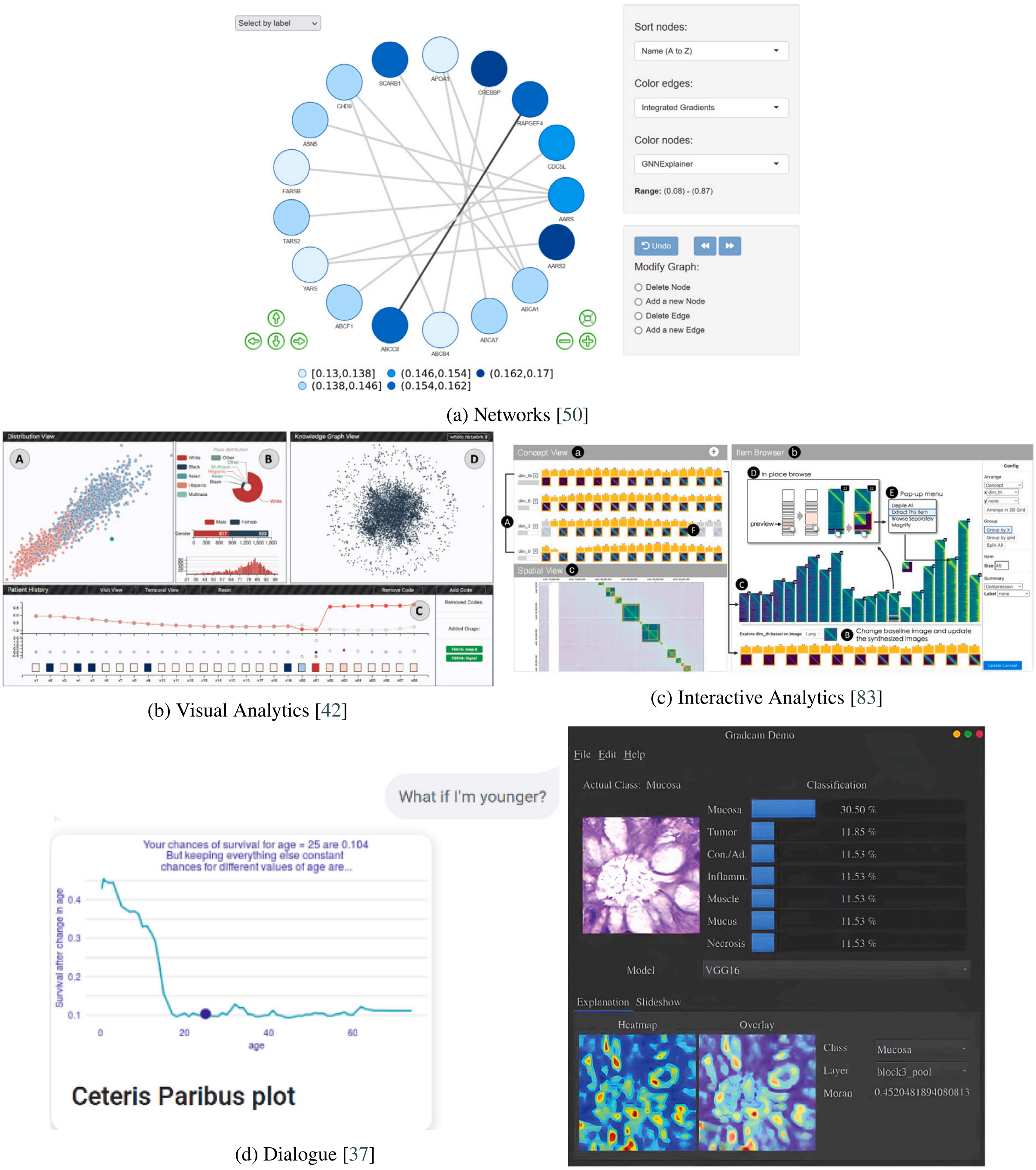
6. Answers to research questions

Building upon the proposed taxonomy and the analysis of information output and interaction styles, we now summarize our key findings and address the overarching research questions that guided our analysis.

RQ1: What output modalities are found in existing XAI interfaces?

Table 2
Table summarizing all papers by their interaction and output properties. Titles abbreviated to Exploration (EXP), Control (CTRL), Contestability (CNTS), Support (SUP), Visual Complexity (VIS CPX), Interaction Complexity (INT CPX) and Multi-Modal Explanations (MM EXP). Model input data abbreviated to Sequential (Seq), Structured (Struc), and Images (Img).

Cluster	Paper	Domain	Input	EXP	CTRL	CNTS	SUP	VIS CPX	INT CPX	MM EXP
Visual analytics	[13]	Healthcare	Struc	0	●	●	●	Med	Low	●
	[40]	Healthcare	Struc	0	●	●	✓	Low	Low	✓
	[33]	Healthcare	Seq	1	●	●	●	Med	Low	●
	[65]	Healthcare	Seq	1	●	●	✓	Low	Low	●
	[41]	Healthcare	Struc	2	●	●	●	Low	Low	●
	[14]	Healthcare	Imgs	2	●	●	●	Med	Low	●
	[42]	Healthcare	Seq	2	●	●	●	High	Med	✓
	[58]	Healthcare	Struc	3	●	●	●	Med	High	✓
	[72]	Healthcare	Graph	4	●	●	●	Med	Med	●
	[37]	Agnostic	Img Seq	1	●	●	●	Med	Low	✓
	[38]	Agnostic	Struc	2	●	●	●	Low	Low	✓
	[61]	Agnostic	Struc	1	●	●	●	Low	Low	●
	[62]	Agnostic	Struc	2	●	●	●	Low	Med	✓
	[39]	Agnostic	Struc	1	●	●	●	Med	High	●
	[34]	Agnostic	Img Seq Struc	3	●	●	●	Med	High	✓
	[63]	Agnostic	Struc	4	●	●	●	Med	High	✓
	[55]	Agnostic	Img Struc Seq	4	✓	●	●	Med	High	✓
Interactive analytics	[77]	Healthcare	Seq	1	●	●	●	Low	Low	✓
	[79]	Healthcare	Seq	1	●	●	●	Low	Low	●
	[78]	Healthcare	Struc	1	●	●	●	Low	Low	✓
	[59]	Healthcare	Struc	1	●	●	●	Low	Low	✓
	[67]	Healthcare	Struc Seq	1	●	●	●	High	Low	●
Interactive analytics	[71]	Healthcare	Struc	2	●	●	●	Med	Low	●
	[43]	Healthcare	Img Struc Seq	3	●	●	●	High	Med	✓
	[66]	Healthcare	Struc	4	●	●	●	High	High	✓
	[68]	Healthcare	Imgs	3	✓	●	●	Med	Med	●
	[36]	Healthcare	Seq	3	✓	●	●	High	Low	●
	[27]	Agnostic	Struc	0	●	●	●	Low	Low	✓
	[31]	Agnostic	Img Struc Seq	3	●	●	●	Med	Low	●
	[32]	Agnostic	Struc	3	●	●	●	High	Med	●
	[35]	Agnostic	Imgs	3	●	●	●	High	High	✓
Networks	[48]	Healthcare	Graph	2	✓	●	●	Low	Med	●
	[45]	Agnostic	Imgs	2	●	●	●	Low	Med	✓
	[46]	Agnostic	Imgs	2	●	●	●	Low	Med	✓
	[47]	Agnostic	Struc	2	✓	●	●	Low	Med	●
Img visualization	[51]	Healthcare	Imgs	0	●	●	●	Low	Low	●
	[65]	Healthcare	Imgs	1	●	●	●	Low	Low	●
	[52]	Healthcare	Imgs	1	●	●	●	Low	Low	●
	[53]	Healthcare	Imgs	1	●	●	●	Low	Low	●
	[54]	Healthcare	Imgs	1	●	●	●	Low	Low	●
	[50]	Agnostic	Imgs	1	●	●	●	Low	Low	●
Natural language	[44]	Agnostic	Struc	2	●	●	●	Low	Low	✓



Sort nodes: Name (A to Z)

Color edges: Integrated Gradients

Color nodes: GNNExplainer

Range: (0.08) - (0.87)

(a) Networks [50]

(b) Visual Analytics [42]

(c) Interactive Analytics [83]

Your chances of survival for age = 25 are 0.104
But keeping everything else constant
chances for different values of age are...

(d) Dialogue [37]

Actual Class: Mucosa

Classification

Mucosa	30.50 %
Tumor	11.85 %
Con./Ad.	11.53 %
Inflamm.	11.53 %
Muscle	11.53 %
Mucus	11.53 %
Necrosis	11.53 %

Model VGG16

Class Mucosa
Layer block3_pool
Moran 0.4520481894080813

(e) Image Visualisation [53]

Fig. 6. Examples of interfaces for each cluster in our taxonomy.

In the explainable interfaces we surveyed, diverse output modalities have been identified. Primarily, Feature Contribution explanations are prevalent, employing modalities such as bar plots, line plots, and textual representations. Partial Dependence explanations commonly utilize line plots to illustrate the impact of individual input features on model predictions. Counterfactual explanations, contrasting predicted outcomes, are often presented through textual descriptions, numerical values, or visual aids like charts with arrows denoting feature value changes. Rule-based methods predominantly communicate through natural language. Neural Network Concepts are represented via human-interpretable patterns, often visualized by images of high activation regions or synthetic images. Feature Context is conveyed through charts like histograms, scatter plots, and box plots, along with algorithmic tools like t-SNE for dimensionality reduction. Model Performance metrics are commonly expressed using charts, particularly bar plots derived from confusion matrices. Charts are the most common output modality in the interfaces developed so far.

RQ2: What user interaction styles and methods are used in explainable interfaces?

Explainable interfaces can employ four primary user interaction styles: exploration, control, contestability, and support. Exploration, the most prevalent style, facilitates users in navigating interface information, often incorporating direct manipulation through actions like clicking, hovering, and drag-and-drop. Control enables users to directly influence predictive models, typically achieved through indirect manipulation involving sliders, drop-down menus, and buttons. Contestability involves allowing users to dispute outputs and demanding active mechanisms for interpreting and resolving user disagreement. However, while recognized as crucial, contestability has not been implemented in any of the interfaces we surveyed. Finally, the supportive interaction style integrates explainability into experts' workflows, permitting them to analyse data using their domain knowledge. These interaction styles contribute to shifting agency to users and accommodating diverse preferences and needs.

RQ3: What properties of explainable interfaces promote interactivity and accessibility in medical domain tasks?

The effectiveness of explainable interfaces in medical domain tasks hinges on several properties that enhance interactivity and accessibility. Interfaces incorporating exploration styles empower users to navigate information from diverse perspectives. Control-oriented interfaces allow direct user adjustments to predictive models, accommodating domain-specific knowledge integration and retraining options. Contestability, though underrepresented, is crucial for resolving user disagreements, requiring active interaction mechanisms. Supportive interaction styles integrate explanations seamlessly into experts' workflows, reducing biases and facilitating as-needed predictions. Dynamic explanation detail, what-if analysis, prototypes, sub-population creation, and model architecture understanding are exploration themes that empower users in medical tasks, promoting detailed investigation, scenario testing, instance similarity exploration, and data contextualization. Furthermore, explanation customization allows tailoring explanations to user preferences or requirements, enhancing accessibility and user-centred adaptability.

The interplay between visual and interactive complexity in explainable interfaces within the medical domain is evident. While interfaces offering multiple exploration types tend to exhibit heightened visual and interaction complexities, they provide users with diverse avenues for in-depth understanding. Incorporating multi-modal explanations, encompassing textual, graphical, and interactive elements, contributes to a richer user experience. However, the inherent challenge lies in balancing complexity and interactivity. As interfaces become more intricate to accommodate various exploration styles and modalities, the potential trade-off emerges, necessitating careful design considerations.

7. Future research directions

Completing this literature review has given us a comprehensive understanding of the present hurdles and prospects associated with developing XAI interfaces in health care. As a result, we have pinpointed several research prospects that can be explored in future studies within this field.

Exploring decision support tools for medical interfaces

A promising avenue for future research lies in exploring decision-support tools tailored specifically for medical interfaces. These tools should extend beyond mere explanation provision and actively assist healthcare professionals in making well-informed decisions. Interfaces should aim to integrate explainable algorithms into medical experts' existing workflows and investigate how different design considerations affect biases in human decision-making.

Investigating contestability in medical decision support

Future research in the field of XAI interfaces should delve into the concept of contestability in medical decision support. Contestability allows users to challenge and question the decisions made by the AI system. In the medical domain, where decisions can have profound consequences, developing interfaces that provide explanations and facilitate user contestability is crucial. Research endeavours should focus on designing interaction mechanisms that enable users to express disagreement, receive alternative predictions, and actively participate in refining the decision-making process. Understanding the user's perspective and incorporating contestability features can enhance AI systems' overall trustworthiness and acceptance in medical decision support.

Counterfactuals for image classification

A critical research direction involves the development of medical interfaces with a focus on image classification tasks and integrating counterfactual explanations. Addressing the complexities associated with image-based diagnostics, these interfaces should provide actionable insights to medical experts, particularly radiologists. The incorporation of counterfactuals can aid in enhancing the interpretability of image classification models. Recent literature started exploring image counterfactuals through natural language concepts [83] and image generation methods [84], but these have not yet been studied regarding interaction design in user interfaces.

Leveraging large language models in XAI interfaces

In recent years, the integration of large language models (LLMs) has emerged as a promising avenue in the development of XAI interfaces in radiology. LLMs can potentially revolutionize how medical experts interact with AI-driven systems. Natural language dialogue shows potential in addressing the trade-off between interface complexity and user exploration, which we found in our survey.

However, it is imperative to acknowledge the limitations and challenges associated with current LLMs. One major concern is their limited precision in technical fields. As reported by Kasneci et al. [85], LLMs may struggle to use precise technical language, essential for maintaining the accuracy and reliability of AI-assisted diagnoses and treatment plans. Furthermore, LLMs can inadvertently introduce biases and provide incorrect information, raising concerns about their reliability in the healthcare domain [86].

Research is already being done on tailoring the language of LLMs to more specific medical domains [87,88]. Medical education is a research direction being discussed as a potential application of LLMs [89,90], and we believe that the design considerations we discuss in this review are also important in this domain. Although not directly related to the XAI field, there is also research on how users can contest outputs given by these models [91,92].

Future research should focus on the synergistic combination of LLMs with other XAI techniques to harness their potential benefits while mitigating concerns related to bias and inaccuracy. Collaboration between LLMs and XAI methods holds great potential in creating balanced systems where the strengths of LLMs in natural language understanding and generation can be harnessed while maintaining the integrity of medical information.

8. Conclusion

Our comprehensive analysis of 42 interfaces in explainable artificial intelligence (XAI) has illuminated the multifaceted landscape of information presentation, user interaction styles, and exploration methods. The findings underscore the diverse goals of user interfaces, ranging from feature contribution and partial dependence explanations to counterfactuals, rule-based methods, neural network concepts, feature context, and model performance. We identified five clusters of interfaces: Visual Analytics, Interactive Analytics, Networks, Image Visualization, and Natural Language Dialogue.

The exploration interaction style proved to be the dominant mode of user engagement, with users predominantly relying on dynamic explanation details, what-if analysis, prototypes, sub-population creation, understanding model architecture, and explanation customization. We found an underlying trade-off between interface complexity and the ability to offer users these different exploration themes. Our investigation also revealed certain limitations and gaps in the current landscape of XAI interfaces, particularly in the medical domain. The lack of robust contestability features, limited incorporation of control or supportive interaction styles, and a scarcity of explainable interfaces with natural language dialogue pose challenges in achieving comprehensive user engagement.

Looking forward, we propose several research prospects to address these challenges and enhance the effectiveness of XAI interfaces. Future studies should explore decision support tools tailored for medical interfaces, investigate contestability in medical decision support, focus on counterfactual explanations for image classification tasks, and leverage LLMs in conjunction with other XAI techniques to overcome limitations and biases associated with current LLMs.

By delving into these research directions, we anticipate a significant advancement in developing XAI interfaces, making them more applicable, trustworthy, and accessible, especially in critical domains like healthcare. As technology evolves, the synergy between human expertise and AI capabilities will play a pivotal role in shaping the future of explainable artificial intelligence, fostering a deeper understanding and collaboration between users and intelligent systems.

CRediT authorship contribution statement

João Cálem: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Catarina Moreira:** Writing – review & editing, Validation, Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Joaquim Jorge:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by national funds through *Fundação para a Ciência e a Tecnologia* (FCT), Portugal, under grants 2022.09212.PTDC (XAVIER Project), 2023.03097.BD and project UIDB/50021/2020 (DOI:10.54499/UIDB/50021/2020).

References

[1] D.J. Mollura, M.P. Culp, M.P. Lungren (Eds.), Radiology in Global Health, Springer International Publishing, 2019, http://dx.doi.org/10.1007/978-3-319-98485-8.

[2] X. Liu, L. Faes, A.U. Kale, S.K. Wagner, D.J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, J.R. Ledsam, M.K. Schmid, K. Balaskas, E.J. Topol, L.M. Bachmann, P.A. Keane, A.K. Denniston, A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis, Lancet Digit. Health 1 (2019) e271–e297, http://dx.doi.org/10.1016/s2589-7500(19)30123-2.

[3] P. Rajpurkar, E. Chen, O. Banerjee, E.J. Topol, AI in health and medicine, Nat. Med. 28 (2022) 31–38, http://dx.doi.org/10.1038/s41591-021-01614-0.

[4] A.J. Larrazabal, N. Nieto, V. Peterson, D.H. Milone, E. Ferrante, Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis, Proc. Natl. Acad. Sci. 117 (2020) 12592–12594, http://dx.doi.org/10.1073/pnas.1919012117.

[5] H. Zhang, A.X. Lu, M. Abdalla, M. McDermott, M. Ghassemi, Hurtful words: Quantifying biases in clinical contextual word embeddings, 2020, http://dx.doi.org/10.48550/ARXIV.2003.11515, URL: https://arxiv.org/abs/2003.11515.

[6] J.R. Zech, M.A. Badgeley, M. Liu, A.B. Costa, J.J. Titano, E.K. Oermann, Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study, PLoS Med. 15 (2018) e1002683.

[7] U. Mahmood, R. Shrestha, D.D. Bates, L. Mannelli, G. Corrias, Y.E. Erdi, C. Kanan, Detecting spurious correlations with sanity tests for artificial intelligence guided radiology systems, Front. Digit. Health 3 (2021) 671015.

[8] R. Geirhos, J.H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, F.A. Wichmann, Shortcut learning in deep neural networks, Nat. Mach. Intell. 2 (2020) 665–673.

[9] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 9 (2019) e1312.

[10] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38, http://dx.doi.org/10.1016/j.artint.2018.07.007.

[11] C. Moreira, Y.L. Chou, M. Velmurugan, C. Ouyang, R. Sindhgatta, P. Bruza, Linda-bn: An interpretable probabilistic approach for demystifying black-box predictive models, Decis. Support Syst. 150 (2021) 113561.

[12] M. Chromik, A. Butz, Human-XAI Interaction: A Review and Design Principles for Explanation User Interfaces, Springer International Publishing, 2021, pp. 619–640, http://dx.doi.org/10.1007/978-3-030-85616-8_36.

[13] A. Bhattacharya, J. Ooge, G. Stiglic, K. Verbert, Directive explanations for monitoring the risk of diabetes onset: Introducing directive data-centric explanations and combinations to support what-if explorations, in: Proceedings of the 28th International Conference on Intelligent User Interfaces, 2023, pp. 204–219, http://dx.doi.org/10.1145/3581641.3584075, arXiv:2302.10671.

[14] E.W. Prince, T.C. Hankinson, C. Görg, The iterative design process of an explainable AI application for non-invasive diagnosis of CNS tumors: A user-centered approach, in: 2023 Workshop on Visual Analytics in Healthcare, VAHC, IEEE, Melbourne, Australia, 2023, pp. 7–13, http://dx.doi.org/10.1109/VAHC60858.2023.00008.

[15] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, J. Glanville, J.M. Grimshaw, A. Hróbjartsson, M.M. Lalu, T. Li, E.W. Loder, E. Mayo-Wilson, S. McDonald, L.A. McGuinness, L.A. Stewart, J. Thomas, A.C. Tricco, V.A. Welch, P. Whiting, D. Moher, The prisma 2020 statement: an updated guideline for reporting systematic reviews, BMJ (2021) n71, http://dx.doi.org/10.1136/bmj.n71.

[16] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (xai), IEEE Access 6 (2018) 52138–52160, http://dx.doi.org/10.1109/access.2018.2870052.

[17] L. Alzubaidi, A. Al-Sabaawi, J. Bai, A. Dukhan, A.H. Alkenani, A. Al-Asadi, H.A. Alwzwazy, M. Manoufali, M.A. Fadhel, A. Albahri, et al., Towards risk-free trustworthy artificial intelligence: Significance and requirements, Int. J. Intell. Syst. 2023 (2023) 4459198, http://dx.doi.org/10.1155/2023/4459198.

[18] D. Collaris, J.J. van Wijk, Machine learning interpretability through contribution-value plots, in: Proceedings of the 13th International Symposium on Visual Information Communication and Interaction, Association for Computing Machinery, New York, NY, USA, 2020, pp. pp. 1–5, http://dx.doi.org/10.1145/3430036.3430067.

[19] 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) 1–45, http://dx.doi.org/10.1145/3387166.

[20] A. Chatzimparmpas, R.M. Martins, I. Jusufi, A. Kerren, A survey of surveys on the use of visualization for interpreting machine learning models, Inf. Vis. 19 (2020) 207–233, http://dx.doi.org/10.1177/1473871620904671.

[21] G. Alicioglu, B. Sun, A survey of visual analytics for explainable artificial intelligence methods, Comput. Graph. 102 (2022) 502–520, http://dx.doi.org/10.1016/j.cag.2021.09.002.

[22] M.R. Islam, M.U. Ahmed, S. Barua, S. Begum, A systematic review of explainable artificial intelligence in terms of different application domains and tasks, Appl. Sci. 12 (2022) 1353, http://dx.doi.org/10.3390/app12031353.

[23] S. Ali, F. Akhlaq, A.S. Imran, Z. Kastrati, S.M. Daudpota, M. Moosa, The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review, Comput. Biol. Med. 166 (2023) 107555, http://dx.doi.org/10.1016/j.compbiomed.2023.107555.

[24] J. Petch, S. Di, W. Nelson, Opening the black box: The promise and limitations of explainable machine learning in cardiology, Can. J. Cardiol. 38 (2022) 204–213, http://dx.doi.org/10.1016/j.cjca.2021.09.004.

[25] A.S. Albahri, A.M. Duhaim, M.A. Fadhel, A. Alnoor, N.S. Baqer, L. Alzubaidi, O.S. Albahri, A.H. Alamoodi, J. Bai, A. Salhi, J. Santamaría, C. Ouyang, A. Gupta, Y. Gu, M. Deveci, A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion, Inf. Fusion 96 (2023) 156–191, http://dx.doi.org/10.1016/j.inffus.2023.03.008.

[26] M. Nazar, M.M. Alam, E. Yafi, M.M. Su'ud, A systematic review of human–computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques, IEEE Access 9 (2021) 153316–153348, http://dx.doi.org/10.1109/access.2021.3127881.

[27] M. Szymanski, M. Millecamp, K. Verbert, Visual, textual or hybrid: the effect of user expertise on different explanations, in: 26th International Conference on Intelligent User Interfaces, ACM, 2021, pp. 109–119, http://dx.doi.org/10.1145/3397481.3450662.

[28] N.O. Bernsen, Defining a taxonomy of output modalities from an hci perspective, Comput. Stand. Interfaces 18 (1997) 537–553, http://dx.doi.org/10.1016/s0920-5489(97)00018-4.

[29] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 4768–4777.

[30] M.T. Ribeiro, S. Singh, C. Guestrin, Why should i trust you?: Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, 2016, pp. 1135–1144, http://dx.doi.org/10.1145/2939672.2939778.

[31] M. Kahng, P.Y. Andrews, A. Kalro, D.H. Chau, ActiVis: Visual exploration of industry-scale deep neural network models, IEEE Trans. Vis. Comput. Graph. Trans. Vis. Comput. Graph. 24 (2018) 88–97, http://dx.doi.org/10.1109/tvcg.2017.2744718.

[32] X. Zhao, Y. Wu, D.L. Lee, W. Cui, iForest: Interpreting random forests via visual analytics, IEEE Trans. Vis. Comput. Graph. Trans. Vis. Comput. Graph. 25 (2019) 407–416, http://dx.doi.org/10.1109/tvcg.2018.2864475.

[33] R. Li, C. Yin, S. Yang, B. Qian, P. Zhang, Marrying medical domain knowledge with deep learning on electronic health records: A deep visual analytics approach, IEEE Trans. Vis. Comput. Graph. Trans. Vis. Comput. Graph. 22 (2020) e20645, http://dx.doi.org/10.2196/20645.

[34] T. Spinner, U. Schlegel, H. Schafer, M. El-Assady, explAIner: A visual analytics framework for interactive and explainable machine learning, IEEE Trans. Vis. Comput. Graph. Trans. Vis. Comput. Graph. (2019) http://dx.doi.org/10.1109/tvcg.2019.2934629, 1–1.

[35] J. Huang, A. Mishra, B.C. Kwon, C. Bryan, ConceptExplainer: Interactive explanation for deep neural networks from a concept perspective, IEEE Trans. Vis. Comput. Graph. Trans. Vis. Comput. Graph. 83 (2023) 1–841, http://dx.doi.org/10.1109/tvcg.2022.3209384.

[36] B.C. Kwon, M.J. Choi, J.T. Kim, E. Choi, Y.B. Kim, S. Kwon, J. Sun, J. Choo, RetainVis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records, IEEE Trans. Vis. Comput. Graph. Trans. Vis. Comput. Graph. 25 (2019) 299–309, http://dx.doi.org/10.1109/tvcg.2018.2865027.

[37] A. Vyas, P. Calyam, An interactive graphical visualization approach to CNNs and RNNs, in: 2020 IEEE Applied Imagery Pattern Recognition Workshop, AIPR, IEEE, 2020, pp. 1–7, http://dx.doi.org/10.1109/aipr50011.2020.9425299.

[38] F. Hohman, A. Srinivasan, S.M. Drucker, TeleGam: Combining visualization and verbalization for interpretable machine learning, in: 2019 IEEE Visualization Conference, VIS, IEEE, 2019, pp. 151–155, http://dx.doi.org/10.1109/visual.2019.8933695, URL: https://ieeexplore.ieee.org/abstract/document/8933695.

[39] J. Yuan, G.Y.Y. Chan, B. Barr, K. Overton, K. Rees, L.G. Nonato, E. Bertini, C.T. Silva, Subplex: A visual analytics approach to understand local model explanations at the subpopulation level, IEEE Trans. Vis. Comput. Graph. Trans. Vis. Comput. Graph. 42 (2022) 24–36, http://dx.doi.org/10.1109/mcg.2022.3199727.

[40] M. Naiseh, D. Al-Thani, N. Jiang, R. Ali, How the different explanation classes impact trust calibration: The case of clinical decision support systems, Int. J. Hum.-Comput. Stud. 169 (2023) 102941, http://dx.doi.org/10.1016/j.ijhcs.2022.102941.

[41] T. Wünn, D. Sent, L.W.P. Peute, S. Leijnen, Trust in artificial intelligence: Exploring the influence of model presentation and model interaction on trust in a medical setting, in: S. Nowaczyk, P. Biecek, N.C. Chung, M. Vallati, P. Skruch, J. Jaworek-Korjakowska, S. Parkinson, A. Nikitas, M. Atzmüller, T. Kliegr, U. Schmid, S. Bobek, N. Lavrac, M. Peeters, R. van Dierendonck, S. Robben, E. Mercier-Laurent, G. Kayakutlu, M.L. Owoc, K. Mason, A. Wahid, P. Bruno, F. Calimeri, F. Cauteruccio, G. Terracina, D. Wolter, J.L. Leidner, M. Kohlhase, V. Dimitrova (Eds.), Artificial Intelligence. ECAI 2023 International Workshops, Springer Nature, Switzerland, Cham, 2024, pp. 76–86, http://dx.doi.org/10.1007/978-3-031-50485-3_6.

[42] J. Wang, K. Mueller, DOMINO : Visual causal reasoning with time-dependent phenomena, IEEE Trans. Vis. Comput. Graphics 29 (2023) 5342–5356, http://dx.doi.org/10.1109/TVCG.2022.3207929, arXiv:2303.06556.

[43] Y. Ouyang, Y. Wu, H. Wang, C. Zhang, F. Cheng, C. Jiang, L. Jin, Y. Cao, Q. Li, Leveraging historical medical records as a proxy via multimodal modeling and visualization to enrich medical diagnostic learning, 2023, arXiv:2307.12199.

[44] M. Kuźba, P. Biecek, What would you ask the machine learning model? Identification of user needs for model explanations based on human-model conversations, in: ECML PKDD 2020 Workshops, Springer International Publishing, 2020, pp. 447–459, http://dx.doi.org/10.1007/978-3-030-65965-3_30.

[45] H. Park, N. Das, R. Duggal, A.P. Wright, O. Shaikh, F. Hohman, D.H.P. Chau, NeuroCartography: Scalable automatic visual summarization of concepts in deep neural networks, IEEE Trans. Vis. Comput. Graph. Trans. Vis. Comput. Graph. 28 (2022) 813–823, http://dx.doi.org/10.1109/tvcg.2021.3114858.

[46] F. Hohman, H. Park, C. Robinson, D.H.P. Chau, Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations, IEEE Trans. Vis. Comput. Graph. Trans. Vis. Comput. Graph. 26 (2020) 1096–1106, http://dx.doi.org/10.1109/tvcg.2019.2934659.

[47] M.N. Hoque, K. Mueller, Outcome-explorer: A causality guided interactive visual interface for interpretable algorithmic decision making, IEEE Trans. Vis. Comput. Graph. Trans. Vis. Comput. Graph. 28 (2022) 4728–4740, http://dx.doi.org/10.1109/tvcg.2021.3102051.

[48] J.M. Metsch, A. Saranti, A. Angerschmid, B. Pfeifer, V. Klemt, A. Holzinger, A.C. Hauschild, CLARUS: An interactive explainable AI platform for manual counterfactuals in graph neural networks, J. Biomed. Inform. 150 (2024) 104600, http://dx.doi.org/10.1016/j.jbi.2024.104600.

[49] A. Mohammed, C. Geppert, A. Hartmann, P. Kuritcyn, V. Bruns, U. Schmid, T. Wittenberg, M. Benz, B. Finzel, Explaining and evaluating deep tissue classification by visualizing activations of most relevant intermediate layers, Curr. Direct. Biomed. Eng. 8 (2022) 229–232, http://dx.doi.org/10.1515/cdbme-2022-1059.

[50] M. Zurowietz, T.W. Nattkemper, An interactive visualization for feature localization in deep neural networks, IEEE Trans. Vis. Comput. Graph. Trans. Vis. Comput. Graph. 3 (2020) 49, http://dx.doi.org/10.3389/frai.2020.00049.

[51] N.A. Hroub, A.N. Alsannaa, M. Alowaifeer, M. Alfarraj, E. Okafor, Explainable deep learning diagnostic system for prediction of lung disease from medical images, Comput. Biol. Med. 170 (2024) 108012, http://dx.doi.org/10.1016/j.compbiomed.2024.108012.

[52] N. Gorre, E. Carranza, J. Fuhrman, H. Li, R.K. Madduri, M. Giger, I. El Naqa, MIDRC CRP10 AI interface - an integrated tool for exploring, testing and visualization of AI models, Phys. Med. Biol. (2023) 68, http://dx.doi.org/10.1088/1361-6560/acb754.

[53] N. Sarkar, M. Kumagai, S. Meyr, S. Pothapragada, M. Unberath, G. Li, S.R. Ahmed, E.B. Smith, M.A. Davis, G.D. Khatri, A. Agrawal, Z.S. Delproposto, H. Chen, C.G. Caballero, D. Dreizin, An ASER AI ML expert panel formative user research study for an interpretable interactive splenic AAST grading graphical user interface prototype, Emerg. Radiol. 31 (2024) 167–178, http://dx.doi.org/10.1007/s10140-024-02202-8.

[54] S. Laguna, J.N. Heidenreich, J. Sun, N. Cetin, I. Al-Hazwani, U. Schlegel, F. Cheng, M. El-Assady, ExpLIMEable: A visual analytics approach for exploring LIME, in: 2023 Workshop on Visual Analytics in Healthcare, VAHC, 2023, pp. 27–33, http://dx.doi.org/10.1109/VAHC60858.2023.00011.

[55] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viegas, J. Wilson, The what-if tool: Interactive probing of machine learning models, IEEE Trans. Vis. Comput. Graph. Trans. Vis. Comput. Graph. 5 (2019) 6–65, http://dx.doi.org/10.1109/tvcg.2019.2934619.

[56] M. Velmurugan, C. Ouyang, R. Sindhgatta, C. Moreira, Through the looking glass: evaluating post hoc explanations using transparent models, Int. J. Data Sci. Anal. (2023) 1–21.

[57] Y.L. Chou, C. Moreira, P. Bruza, C. Ouyang, J. Jorge, Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications, Inf. Fusion 81 (2022) 59–83.

[58] J. Krause, A. Dasgupta, J. Swartz, Y. Aphinyanaphongs, E. Bertini, A workflow for visual diagnostics of binary classifiers using instance-level explanations, in: 2017 IEEE Conference on Visual Analytics Science and Technology, VAST, IEEE, 2017, pp. 162–172, http://dx.doi.org/10.1109/vast.2017.8585720.

[59] D. Wang, Q. Yang, A. Abdul, B.Y. Lim, Designing theory-driven user-centric explainable AI, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, ACM, 2019, pp. 1–15, http://dx.doi.org/10.1145/3290605.3300831.

[60] C. Hsieh, C. Moreira, C. Ouyang, Dice4el: interpreting process predictions using a milestone-aware counterfactual approach, in: 2021 3rd International Conference on Process Mining, ICPM, IEEE, 2021, pp. 88–95.

[61] O. Gomez, S. Holter, J. Yuan, E. Bertini, Vice: visual counterfactual explanations for machine learning models, in: Proceedings of the 25th International Conference on Intelligent User Interfaces, Association for Computing Machinery, New York, NY, USA, 2020, pp. 531–535, http://dx.doi.org/10.1145/3377325.3377536.

[62] O. Gomez, S. Holter, J. Yuan, E. Bertini, Advice: Aggregated visual counterfactual explanations for machine learning model validation, in: 2021 IEEE Visualization Conference, VIS, 2021, pp. 31–35, http://dx.doi.org/10.1109/VIS49827.2021.9623271, URL: https://ieeexplore.ieee.org/abstract/document/9623271.

[63] J. Yuan, E. Bertini, Context sight: model understanding and debugging via interpretable context, in: Proceedings of the Workshop on Human-in-the-Loop Data Analytics, Association for Computing Machinery, New York, NY, USA, 2022, http://dx.doi.org/10.1145/3546930.3547502.

[64] H.J. Schulz, Treevis.net: A tree visualization reference, IEEE Comput. Graph. Appl. 31 (2011) 11–15, http://dx.doi.org/10.1109/MCG.2011.103.

[65] K.A. Tarnowska, B.C. Dispoto, J. Conragan, Explainable ai-based clinical decision support system for hearing disorders, in: AMIA ... Annual Symposium Proceedings. AMIA Symposium 2021, 2021, p. 595.

[66] Y. Ming, H. Qu, E. Bertini, RuleMatrix: Visualizing and understanding classifiers with rules, IEEE Trans. Vis. Comput. Graph. Trans. Vis. Comput. Graph. 25 (2019) 342–352, http://dx.doi.org/10.1109/tvcg.2018.2864812.

[67] F. Cheng, D. Liu, F. Du, Y. Lin, A. Zytek, H. Li, H. Qu, K. Veeramachaneni, VBridge: Connecting the dots between features and data to explain healthcare models, IEEE Trans. Vis. Comput. Graph. Trans. Vis. Comput. Graph. 28 (2022) 378–388, http://dx.doi.org/10.1109/tvcg.2021.3114836.

[68] Q. Wang, S. L'Yi, N. Gehlenborg, DRAVA: Aligning human concepts with machine learning latent dimensions for the visual exploration of small multiples, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1–15, http://dx.doi.org/10.1145/3544548.3581127.

[69] N. Choudhury, S. Ara, A survey on case-based reasoning in medicine, Int. J. Adv. Comput. Sci. Appl. 7 (2016) 136–144, http://dx.doi.org/10.14569/ijacsa.2016.070820.

[70] L. van der Maaten, G. Hinton, Visualizing data using t-sne, J. Mach. Learn. Res. 9 (2008) 2579–2605, URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

[71] Y. Li, T. Fujiwara, Y.K. Choi, K.K. Kim, K.L. Ma, A visual analytics system for multi-model comparison on clinical data predictions, IEEE Trans. Vis. Comput. Graph. Trans. Vis. Comput. Graph. 4 (2020) 122–131, http://dx.doi.org/10.1016/j.visinf.2020.04.005.

[72] Q. Wang, K. Huang, P. Chandak, M. Zitnik, N. Gehlenborg, Extending the nested model for user-centric XAI: A design study on GNN-based drug repurposing, IEEE Trans. Vis. Comput. Graph. 29 (2023) 1266–1276, http://dx.doi.org/10.1109/TVCG.2022.3209435.

[73] B. Shneiderman, Designing the user interface strategies for effective human–computer interaction, ACM SIGBIO Newslett. 9 (1987) 6, http://dx.doi.org/10.1145/25065.950626.

[74] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, Fernández-leal, Human-in-the-loop machine learning: a state of the art, Artif. Intell. Rev. 56 (2022) 3005–3054, http://dx.doi.org/10.1007/s10462-022-10246-w.

[75] K. Vaccaro, C. Sandvig, K. Karahalios, ‘‘At the end of the day facebook does what itwants’’: How users experience contesting algorithmic content moderation, Proc. ACM Hum.-Comput. Interacti. 4 (2020) 1–22, http://dx.doi.org/10.1145/3415238.

[76] H. Lyons, T. Miller, E. Velloso, Algorithmic decisions, desire for control, and the preference for human review over algorithmic review, in: 2023 ACM Conference on Fairness, Accountability, and Transparency, ACM, 2023, pp. 764–774, http://dx.doi.org/10.1145/3593013.3594041.

[77] C. Panigutti, A. Beretta, F. Giannotti, D. Pedreschi, Understanding the impact of explanations on advice-taking: a user study for ai-based clinical decision support systems, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2022, http://dx.doi.org/10.1145/3491102.3502104.

[78] A.J. Barda, C.M. Horvat, H. Hochheiser, A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare, IEEE Trans. Vis. Comput. Graph. Trans. Vis. Comput. Graph. 20 (2020) 1–16, http://dx.doi.org/10.1186/s12911-020-01276-x.

[79] C. Hur, J. Wi, Y. Kim, Facilitating the development of deep learning models with visual analytics for electronic health records, IEEE Trans. Vis. Comput. Graph. Trans. Vis. Comput. Graph. 17 (2020) 8303, http://dx.doi.org/10.3390/ijerph17228303.

[80] M. Chromik, M. Eiband, F. Buchner, A. Krüger, A. Butz, I think i get your point, AI! The illusion of explanatory depth in explainable AI, in: 26th International Conference on Intelligent User Interfaces, ACM, 2021, pp. 307–317, http://dx.doi.org/10.1145/3397481.3450644.

[81] J. Brooke, et al., Sus-a quick and dirty usability scale, Usability Eval. Ind. 189 (1996) 4–7.

[82] A. Holzinger, A. Carrington, H. Müller, Measuring the quality of explanations: the system causability scale (scs) comparing human and machine explanations, KI-Künstliche Intell. 34 (2020) 193–198.

[83] S. Kim, J. Oh, S. Lee, S. Yu, J. Do, T. Taghavi, Grounding counterfactual explanation of image classifiers to textual concept space, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, Vancouver, BC, Canada, 2023, pp. 10942–10950, http://dx.doi.org/10.1109/CVPR52729.2023.01053.

[84] A.J. DeGrave, Z.R. Cai, J.D. Janizek, R. Daneshjou, S.I. Lee, Auditing the inference processes of medical-image classifiers by leveraging generative AI and the expertise of physicians, Nat. Biomed. Eng. (2023) 1–13, http://dx.doi.org/10.1038/s41551-023-01160-9, URL: https://www.nature.com/articles/s41551-023-01160-9.

[85] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, G. Kasneci, ChatGPT for good? On opportunities and challenges of large language models for education, Learn. Indiv. Differ. 103 (2023) 102274, http://dx.doi.org/10.1016/j.lindif.2023.102274.

[86] P.P. Ray, ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope, Internet Things Cyber-Phys. Syst. 3 (2023) 121–154, http://dx.doi.org/10.1016/j.iotcps.2023.04.003.

[87] Z. Liu, A. Zhong, Y. Li, L. Yang, C. Ju, Z. Wu, C. Ma, P. Shu, C. Chen, S. Kim, et al., Tailoring large language models to radiology: A preliminary approach to llm adaptation for a highly specialized domain, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2023, pp. 464–473.

[88] L. Yunxiang, L. Zihan, Z. Kai, D. Ruilong, Z. You, Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge, 2023, arXiv preprint arXiv:2303.14070.

[89] M. Sallam, N. Salim, M. Barakat, A. Al-Tammemi, ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations, Narra J. 3 (2023) e103, http://dx.doi.org/10.52225/narra.v3i1.103.

[90] H. Lee, The Rise of ChatGPT: Exploring its Potential in Medical Education, Anatomical Sciences Education, 2023, http://dx.doi.org/10.1002/ase.2270.

[91] S. Petridis, B. Wedin, J. Wexler, A. Donsbach, M. Pushkarna, N. Goyal, C.J. Cai, M. Terry, ConstitutionMaker: Interactively critiquing large language models by converting feedback into principles, 2023, http://dx.doi.org/10.48550/arXiv.2310.15428, arXiv:2310.15428.

[92] Y. Kim, J. Lee, S. Kim, J. Park, J. Kim, Understanding users' dissatisfaction with ChatGPT responses: Types, resolving tactics, and the effect of knowledge level, in: Proceedings of the 29th International Conference on Intelligent User Interfaces, 2024, pp. 385–404, http://dx.doi.org/10.1145/3640543.3645148, arXiv:2311.07434.