Systematic Review of XAI Tools for AI-HCI Research

Ahmad Alaqsam
Lancaster University
a.alaqsam@lancaster.ac.uk

Corina Sas
Lancaster University
c.sas@lancaster.ac.uk

The explainability of machine learning black-box models is key for designing and adopting AI technologies by end users. XAI tools such as SHAP or LIME have been purposely developed to support such explainability but their exploration in the HCI community has been limited. This paper reports a systematic review of 142 papers targeting design, use or evaluation of XAI tools with the aim to investigate their different types, users, application domains, input and output data sets, and their user interfaces. Findings indicate a broad range of XAI tools but extensive use of a few, a prevalence of AI experts as users rather than evaluators of these tools. We discuss our findings arguing for the need to move beyond the design of novel XAI tools towards increasing their use and comparative evaluation. We also argue for the need for HCI-grounded user interface design for XAI tools and advance an initial design space for AI-HCI research integrating AI affordances with the application domains of XAI tools mapped to key HCI research areas.

Explainable AI. XAI tools. Users. Application domains. XAI interfaces. Input data. Output data

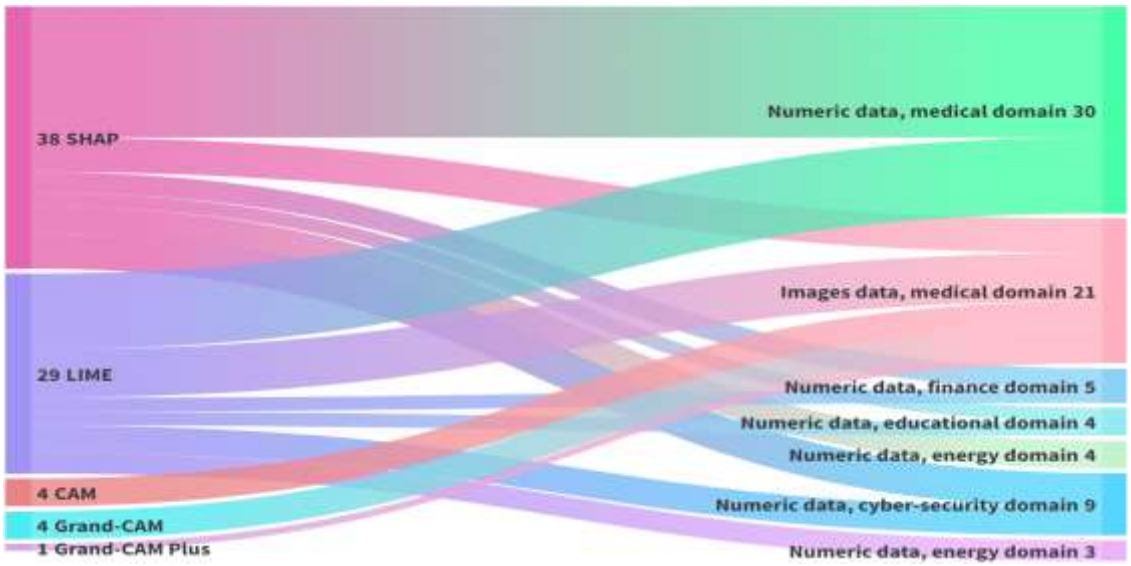

Figure 1: *Alluvial diagram showing the most common XAI tools (left), data type and application domain (right), with number of papers mentioning them from the total of 142 reviewed papers.*

1. INTRODUCTION

Despite the potential for innovation of AI-based technologies (Clement et al., 2023; Dosovitskiy et al., 2020; K. Muhammad et al., 2021; Smith et al., 2022), the complexity of their black-box machine learning models raises significant issues for both expert and non-expert users (Nauta et al., 2023; Saeed & Omlin, 2021). Such models use input data to make predictions, albeit models' outcomes need to be explained in accessible terms, understandable to people, because the reasoning, functioning and causality of these models remains hidden (Keleko et al., 2023; Schoonderwoerd et al., 2021). Explainable AI (XAI) tools such as SHAP (Albahri et al., 2023) or LIME (Dieber & Kirrane, 2020) have emerged to make machine learning models and their predictions easy to understand. However, despite their potential value, most XAI tools have been developed by AI experts, mostly for AI experts (Laato et al., 2021;

Ras et al., 2018) which raises challenges for non-expert users due to their different understanding of such tools and their outputs (Bertini et al., 2022).

The growing interest in XAI tools is reflected in several recent systematic reviews conducted mostly in AI research area which have focused on the types of XAI (Albahri et al., 2023; Chromik & Butz, 2021; Chromik & Schuessler, 2020; Nazaret al., 2021; Suh et al., 2023; Vieira & Digiampietri, 2022; Vilone & Longo, 2021; Weberet al., 2023), or types of provided explanations (Albahri et al., 2023; Laato et al., 2021; Vieira & Digiampietri, 2022; Vilone & Longo, 2021). Such reviews also explored the different users of XAI tools, their AI expertise (Chromik & Schuessler, 2020; Laato et al., 2021; Nazar et al., 2021; Vieira & Digiampietri, 2022; Vilone & Longo, 2021; Weber et al., 2023), and the provision of user interface for XAI tools (Chromik & Butz, 2021; Chromik & Schuessler, 2020; Laato et al., 2021; Nazar et al., 2021; Weber et al., 2023).

These reviews however have focused less on the input and output of XAI tools such as data sets, user interface and interaction. Previous findings also argue for the value of future research on such tools for understandable, interactive, evaluated and ethically informed user interfaces (Nazar et al., 2021), grounded in design principles (Holzinger et al., 2020). Thus, scholars have suggested the importance of the HCI lens (Chromik & Butz, 2021) to the design (D. Wanget et al., 2019) and exploration of XAI tools in order to increase their value for both experts and end-users (Bistarelli et al., 2022; Sarp et al., 2023).

Within the emerging AI-HCI research, scholars have argued for the value of human-centered XAI approaches (Ehsan et al., 2022) better tailored to the needs of different user groups, the focus on HCI affordances and challenges of these approaches, or the exploration of the theoretical underpinning for instance through the lens of social transparency theory (Ehsan et al., 2021). Previous work has also pointed out additional challenges of XAI tools regarding limited interdisciplinary underpinning (Anjomshoae et al., 2019; Langley, Meadows et al., 2017; Xu, 2019), evaluation (Šarčević et al., 2022; Sarp et al., 2021), mental models (Chromik & Butz, 2021; Hoffman et al., 2023; Laato et al., 2021).

Such work has also highlighted ethical challenges pertaining to issues around users' trust and fairness (Brdnik, 2023; Ehsan et al., 2023), privacy (Antoniadi et al., 2021; Longo et al., 2020; Villaronga et al., 2018), security (Barredo et al., 2020; Liang et al., 2021; Tjoa & Guan, 2021), or safety (Barredo et al., 2020; Naiseh et al., 2020).To address this gap, we report a systematic review of 142 papers on XAI tools, focusing on the following research questions:

- What is the broad range of XAI tools and their main focus?
- What are the main input and output of XAI tools?
- Which are HCI opportunities with respect to XAI tools?

Our contribution is three-fold. First, we present more nuance findings on XAI tools focused on their application domains, input/output data, and user interface; second, we advance a design space for AI-HCI research, and third, we articulate some new research and design opportunities in this space.

2. METHOD

We searched ACM, IEEE Xplore databases, and SCOPUS using keyword "XAI tools". The search was conducted in spring 2023 to identify papers published in the previous 10 years (2013-2023). The keyword is the standard term used to describe explainable AI tools. The ten-year window was chosen to ensure that the breadth of work in this space is covered, including early papers on XAI tools. Previous systematic reviews in HCI also tend to employ a ten-year window (Bach et al., 2024; Sanches et al., 2019). The systematic literature review was aligned with the PRISMA standards (Moher et al., 2010).

This search returned 429 initial papers, from which 56 duplicates were removed. The remaining 373 papers were screened by reading their title, and abstracts leading to the exclusion of 154 papers which did not focus on XAI tools but rather on AI algorithms more broadly, or their development. The remaining 219 papers were fully read to ensure their eligibility which led to 77 papers being excluded because although they explored XAI topics such as recommendations or principles they did not focus on designing, using, or evaluating of XAI tools.

This selection process led to a set of 142 papers including 86 journal papers and 56 conference papers (see Prisma diagram in Fig 2). These papers met the following inclusion criteria: 1) mention at least one XAI tool, 2) such XAI tool(s) should be designed, used, or evaluated in the paper. Papers were considered non-eligible if they refer to XAI topics more broadly, albeit without using, designing or evaluating them.

The final set of 142 papers were analysed through hybrid coding (Fereday & Muir-Cochrane, 2006). This included both deductive and inductive codes. The former was informed by previous work, such as taxonomies of XAI tools (Love et al., 2022; Nauta et al., 2023), types of XAI tools such as SHAP or LIME (Albahri et al., 2023; Vieira & Digiampietri, 2022), types of explanation (Laato et al., 2021; Vilone & Longo, 2021) such as ante-hoc explanations, or post-hoc explanations (Barredo et al., 2020; Schwalbe & Finzel, 2023; Singh et al., 2020; Speith,

2022). Additional deductive codes included explanation models such as those based on feature importance (Malandri et al., 2022; Moscato et al., 2021; Wellawatte et al., 2022), the focus of XAI tool, i.e., use for prediction, tool design, or tool evaluation (Fouladgar et al., 2022), and AI expertise of users of XAI tools (Chromik & Schuessler, 2020).

Beside deductive codes, the inductive ones emerged from the reviewed papers and included types and privacy of input data sets, type of output data, application domains, as well as features pertaining to XAI tools' interface and user interaction. Table 1 shows these codes and their frequencies.

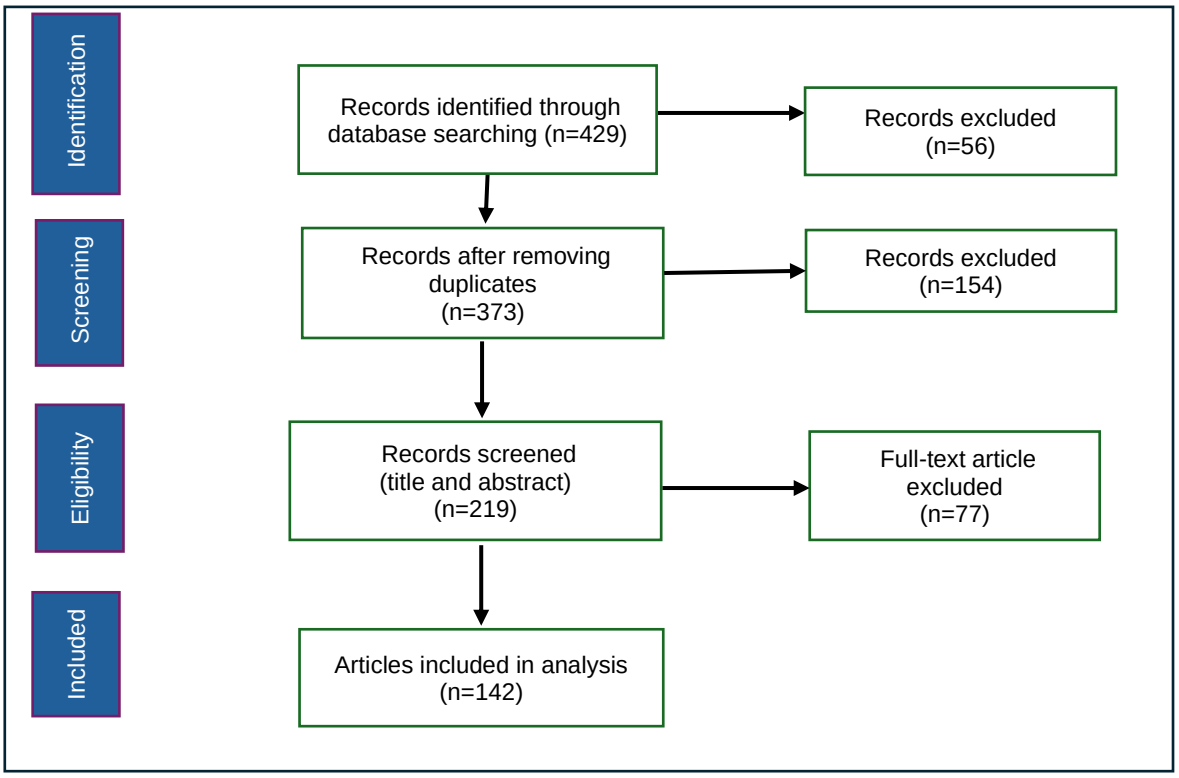

Figure 2: *PRISMA diagram that shows the procedure of papers' selection.*

Table 1: *Main codes and their frequencies.*

Code	Frequency		Code	Frequency		Code	Frequency	
XAI tool			*Explanation type*			*Output data type*		
SHAP	*71*	*(30.3%)*	*Post-hoc*	*138*	*97.2 %)*	*Bar charts*	*111*	*(78.3%)*
LIME	*58*	*(24.8%)*	*Ante-hoc*	*4*	*(2.8 %)*	*Heatmaps*	*22*	*(9.4%)*
Grad-CAM	*16*	*(6.8%)*				*Images*	*15*	*(6.4%)*
CAM	*7*	*(3%)*	*Explanation model*			*Confusion matrix*	*8*	*(3.4%)*
Grad-Cam ++	*4*	*(1.7%)*	*Feature importance*	*125*	*(88%)*	*Text*	*3*	*(1.3%)*
ELIS	*4*	*(1.7%)*	*Visual explanation*	*7*	*(4.9%)*	*Tree charts*	*3*	*(1.3%)*
SmoothGrad	*3*	*(1.3%)*	*Example based*	*4*	*(2.8%)*			
Anchors	*2*	*(0.9%)*	*Text recognition*	*3*	*(2.11%)*			
LRP	*2*	*(0.9%)*	*White-box*	*3*	*(2.11%)*	*XAI tool interface*		
DeepShap	*2*	*(0.9%)*				*Not provided*	*123*	*(86.6%)*
IG	*2*	*(0.9%)*	*Input data type*			*Non-interactive*	*10*	*(7%)*
Others	*63*	*(26.9%)*	*Numeric*	*93*	*(65.5%)*	*Interactive*	*9*	*(6.3%)*
			Image	*36*	*(25.4%)*			
Focus on XAI			*Text*	*4*	*(2.8%)*			
Use	*88*	*(62%)*	*Video*	*2*	*(1.4%)*	*Domains*		
Design	*38*	*(26.8%)*				*Medical*	*60*	*(42.3%)*
Evaluate	*16*	*(11.3%)*	*Input data privacy*			*Financial*	*8*	*(5.6%)*
			Public data set	*73*	*(51.4%)*	*Educational*	*6*	*(4.2%)*
Users			*Private data set*	*35*	*(24.7%)*	*Energy*	*6*	*(4.2%)*
AI experts	*124*	*(87%)*	*Not specified*	*34*	*(23.9%)*	*Cyber security*	*6*	*(4.2%)*
Non AI experts	*18*	*(12.7%)*						

3. FINDINGS

We now report the key findings regarding each of the main codes highlighting the breadth of XAI tools and most used, their focus, application domains, and users, different explanation types and models, and input/ output data and tools' interfaces.

3.1. Overview of XAI tools, their focus and users

An important finding is the recent exponential growth of research on XAI tools, with over 4, and 8 times more papers exploring them in 2021 and 2022 compared to 2020. Outcomes also indicate a broad range of XAI tools, namely 74, although most of them (63) have been mentioned in one paper only. As shown in XAI tool code in Table 1, the remaining 11 tools were mentioned in more than one paper, with SHAP, LIME and CAM related tools being by far the mostly used ones, i.e., over 30%, 24%, and 10% respectively. Together, these tools have been used in 112 out of 142 papers.

SHAP (Lundberg & Lee, 2017) is a post-hoc, model-agnostic XAI tool (Angelov et al., 2021) that uses game theory to explain the importance of specific features or predictor variables from a given set which contributes to target variable, or in other words, impacts on model's prediction.

LIME (Dieber & Kirrane, 2020) is also a post-hoc, model-agnostic XAI tool applicable to all machine learning models (Love et al., 2022), and which involves adding random noise to features to identify the local ones predicting the model (Angelov et al., 2021).

CAM (Zhou et al., 2016) is a post-hoc, model-specific XAI tool built on convolutional neural networks to identify important features and generates heatmap highlighting them, with variants (Chattopadhay et al., 2018; Selvaraju et al., 2017) of this tool being developed to improve the explainability (Angelov et al., 2021).

Our findings indicate that the reviewed papers have a three-fold focus namely to use, design, or evaluate XAI tools. An interesting outcome is that most papers merely use the tools for understanding machine learning predictions (62%) (see Focus on XAI tool code in Table 1). Also important, new tools are being designed and developed as described in over 25% of papers, but only 11% of papers focus on evaluating XAI tools. In particular, SHAP, LIME and Grad-CAM are also the most evaluated tools.

With respect to users of XAI tools, these are by large AI experts (87% of papers) with almost a quarter of these papers involving developers of XAI tools as users. In contrast, non-expert users are mentioned in less than 13% of papers and none of the top five most used tools have been evaluated with non-AI experts.

3.2. Application domains

An interesting outcome is the well-defined range of application domains for which XAI tools tend to be used (Figure 1). We found five such domains: medical, financial, educational, energy and cyber security, with the medical being by far the largest (42.3%) (see Application domain code Table 1). We found that numeric data is more applied than other types. For instance, 30 out of 51 papers in medical domain focused on explaining numeric data while the rest focused on image data. With respect to specific XAI tools, papers targeting applications in medical domain used each of the top five most common tools, albeit with an overemphasis on SHAP and LIME. The other four domains used most SHAP and LIME, but made limited use of CAM related tools, which prioritize images as input data. We now look briefly at these domains, illustrating their focus of XAI tools being used, designed, or evaluated.

For medical domain, 40 out of 60 papers focus on using XAI tools to support the explanations of machine learning results for diagnosis. For example, COVID-19 from X-ray images 23, or brain tumour from brain images (D. Wang et al., 2019), utilized wearable sensor-based gait analysis to identify Osteopenia and Sarcopenia (Kim et al., 2022) for prediction of stroke by utilizing dataset included four-channel recordings of stroke patients (Islam, Hussain, Rahman, Park, & Hossain, 2022), or limb rehabilitation outcomes following stroke from a local dataset included criteria were confirmed diagnosis of first-ever stroke. (Gandolfi et al., 2023); for osteoporosis risk screening from the national health and nutrition examination survey datasets, or classification for example of spinal posture from images (Dindorf et al., 2021), or of stress from biodata captured through physiological data during baseline, stress, recovery, and cycling sessions.

About a third from papers targeting medical domain (18) describe the design and development of new XAI tools for diagnosis of skin lesions (Lucieri et al., 2022), prediction of heart attack or for supporting clinical decisions more broadly (Panigutti et al., 2023) as well as doctors' understanding of medical data analytics (Kapcia et al., 2021) such as eye scans or brain scans (Antoniadi et al., 2021; Heberle et al., 2023; Henriksen et al., 2022).

Papers targeting financial domain focused on using XAI tools to support the explanations of machine learning results for prediction for example of stock market performance (Mandeep et al., 2022) or of repaying loans on P2P platforms (Moscato et al., 2021). Scholars also focused on developed new XAI tools for predicting income, or credit risk (Shree et al., 2022), stock performance (Carta et al., 2022), or for understanding financial data analytics (Bistarelli et al., 2022; Leung et al., 2021).

In education domain XAI tools were used to support the explanations of machine learning to predict and understand students' dropouts (Nagy & Molontay, 2023), to understand students' performance in virtual learning environment (Adnan et al., 2022), or to enhance students' understanding of ML models' techniques (C. Wang et al., 2021).

In the energy domain the focus has been less on designing or evaluating XAI tools, but mostly on using them to support the explanations of machine learning to understand forecasting of photovoltaic power generation (Kuzlu et al., 2020; Sarp et al., 2021), electrical load (Šarčević et al., 2022) as well as to assess energy performance (Galli et al., 2022; Laato et al., 2021; Moreno-Sanchez, 2020; Roy et al., 2023).

In cybersecurity application domain, XAI tools were used to support the explanations of machine learning to understand cyber vulnerability assessment (Alperin et al., 2020) website fingerprinting attacks (Gulmezoglu, 2022), or intrusion detection (Ehsan et al., 2023; Heimerl et al., 2019; Šarčević et al., 2022).

3.3. Explanation type and model

Outcomes on the type of explanation provided by the XAI tools show the overemphasis on post-hoc explanation (97%) of papers rather than on ante-hoc explanation (4 papers, less than 3%). This is less surprising, given that ante-hoc explanation is built in transparent or white-box machine learning models, while post-hoc ones are needed for black-box models. We agree that the spread and use of XAI post-hoc methods is more due to its natural which makes it usable with several AI models, however, it should not lead to pay less attention to XAI anti-hoc method that makes AI applications explainable and clarifies lots of the ambiguity without the need of external post-hoc tools.

Our findings show also how the reviewed papers and their XAI tools reflect the five explanation models of their results and predictions such as feature importance, visual, example-based, text recognition and white-box explanation (Moreno-Sanchez, 2020; Skuppin et al., 2022; Q.Wang et al., 2023). Most of the papers leverage feature importance to explain tools' prediction (88% of papers) (see Explanation model code in Table 1). Feature importance explanation highlights the most significant aspects of input data impacting on prediction. In contrast, fewer papers leverage other explanation models, such as visual explanations (less than 5%) highlighting specific aspects in input images that impact on model's results (Heberle et al., 2023), example-based explanation (less than 3%) which leverages example from the training data (Konradi et al., 2022), while text recognition and white-box explanations are limitedly used (each less than 2%).

3.4. Input data: type and privacy

All papers refer to input data sets used for the prediction models whose outcomes are subsequently used as input to the XAI tools, albeit the type of input data varies. The most common type of input data is numeric (65% of papers), followed by images (25%). In contrast, video format was used by only 2 papers. One paper explained 2164 video clips by applying LIME and Anchors tools (Jayakumar & Skandhakumar, 2022), while the second used FEES tool to explain 50 clips (Konradi et al., 2022). (See input data type code in Table 1).

Most common numerical input data were from medical datasets and tend to consist of records or logs such as heart rate (Chalabianloo et al., 2022). Among image-based input data, common ones include those for medical diagnosis or object identification for instance for autonomous vehicles. Less common types of input data are text and videos, with less than 3% of papers employing them. Among the few papers employing text-based input data sets they may use textual explanations as recorded made by doctors for diagnosis purposes (Albahri et al., 2023) or by students to predict their learning outcomes (Adadi & Berrada, 2018). With regard to video as input data, one paper explored 2164 video clips to explain the predictions of a deep fake detector model built on top of the EfficientNet architecture (Jayakumar & Skandhakumar, 2022).

With regard to the privacy of input data sets, over 50% of papers use publicly available data sets, about 25% use private data sets, and the remaining 25% do not specify (see Input data privacy code in Table 1).

3.5. Interface of XAI tools and output data

A key finding is that most of reviewed papers focus on XAI tools provide their explanations in different formats. The XAI tools also provide their output data in a range of formats, such as diagrammatic, visual, or text. Diagrams include mostly bar charts (78.3% of papers), followed by heatmaps (9.4%), confusion matrix (3.4%), and tree charts (1.3%) (see Output data type code in Table 1). Visual outputs consist of images such as Brain images to identify brain tumour diagnosis (Kumar et al., 2021) (6.4%), and a limited number of papers output the results of XAI tools in text format, such as (Qian et al., 2021) that illustrates XAI publications and distil their content. Common XAI tools such as SHAP and LIME use graphs and heatmaps to explain AI model results, while CAM, Grad-CAM and Grad-CAM++ deal with images.

An important outcome is however the noticeable absence of XAI tools' user interface (86.65% of papers) (See XAI tool interface code in Table 1). From the limited number of papers whose XAI tools have user interfaces (13.3%), about half of such

interfaces are non-interactive (7%). No such interfaces are available for the most common XAI tools, but for less used ones such as (Malandri et al., 2022) and (Mercorio et al., 2020). Unfortunately, these new XAI tools tend to be research prototypes, and only a few are available online (Malandri al., 2022; Mercorio et al., 2020; Oduor et al., 2020).

For example, (C. Wang & An, 2021) provide a webpage that allows a user to add a picture then by utilising CNN model highlighting which areas of the picture help the system to decide what is the picture. The main purpose of this webpage is to explain to new designers how the ML model recognises an image. However, it has a main problem with its accuracy and functionality. Although it works with some example, it could not recognise many pictures correctly. Also, it would not be classified as an XAI tool, because it does not provide explanations for the model results.

Another example is (Kadir et al., 2023) that is a tool for understanding an ML image classification model's behaviour in relation to its explanation. It allows users to select the percentage of picture features to show how recognizable the picture would be for the ML model. However, they need to improve the visibility of system status so users informed about what is going on. Moreover, the explanation of why the model provides such results is still messing.

4. DISCUSSION

In this section we highlight the key findings from our systematic review, and their novelty with regard to each of the three research questions.

4.1. Beyond designing novel XAI tools: Increase use and comparative evaluation of existing tools

With respect to the first question on the range of XAI tools and their main focus, our outcomes confirm previous ones on the most common SHAP, LIME and CAM-related tools (Albahri et al., 2023; Chromik & Butz, 2021; Chromik & Schuessler, 2020; Nazar et al., 2021; Suh et al., 2023; Vieira & Digiampietri, 2022; Vilone & Longo, 2021; Weber et al., 2023). We extend this with additional insights into the rapid growth of this research area where many new XAI tools are available or being developed, yet limitedly used: 63 of the 74 identified XAI tools are described in only one paper. In terms of focus, most emphasis is on using the tools, followed by developing them, with limited work focused on their evaluation.

We strengthen the previous calls for evaluating XAI tools (Hoffman et al., 2023; Holzinger et al., 2020; Weber et al., 2023) with the urge for substantial involvement of non-AI experts and end users to design, but more importantly to use and evaluate these tools. For this, we can leverage the five application domains which we identified in which these tools are most commonly used. The importance of the domain in which machine learning models are used for understanding their outcomes has been previously suggested (Keleko et al., 2023).

Our findings indicate that these five domains in which XAI tools are used also align with strong HCI research interest in health, money, education, sustainability, privacy and trust. At the intersection of these domains with HCI interests we call for the exploration of design space that leverage AI affordances such as identification, classification, prediction, or forecasting. For example, in the context of HCI interest on mental health (Sanches et al., 2019). AI technologies can be leveraged for detection, diagnosis, or recommendation of interventions (Thieme et al., 2020) and interactive XAI tools to explain or support users to tailor these outcomes.

However, evaluating XAI tools is crucial to ensure that they meet usability, effectiveness, and user satisfaction (Lopes et al., 2022). HCI evaluation methods for XAI should consider several factors. Firstly, they should assess the clarity of the explanations provided by the XAI tools. This involves examining how well users understand the AI system's decision-making process and whether the explanations align with users' mental models (Hoffman et al., 2023). Secondly, the evaluation methods should gauge the impact of the XAI tools on users' trust and acceptance of AI systems. This includes measuring users' confidence in the tool's explanations, their willingness to rely on the system's recommendations, and their perception of the tool's transparency (Ehsan et al., 2023).

Additionally, the evaluation methods should consider the usability and user experience aspects of the XAI tools, taking into account factors such as ease of use, efficiency, and satisfaction (Lopes et al., 2022). By developing and employing appropriate evaluation methods, HCI researchers and practitioners can contribute to the iterative improvement of XAI tools, ensuring that they effectively support users in understanding and interacting with AI systems (Laato et al., 2022).

4.2. Towards user interfaces for XAI tools

The second research question focuses on the less explored input and output data sets of XAI tools. Our findings indicate the prevalence of numerical and

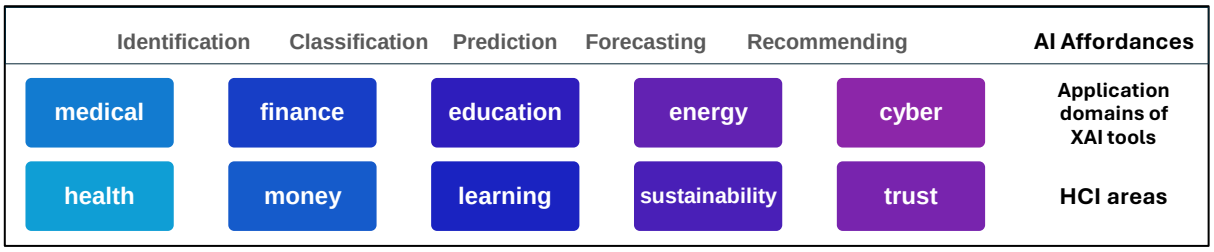

Figure 3. *Towards a design space for AI-HCI research integrating AI affordances with the application domains of XAI tools mapped to key HCI research areas.*

image-based as input data sets and limited use of text and multimodal data. The latter is interesting and opens up research opportunities for better integration of machine learning models with XAI tools to leverage text-based data, which in turn holds potential for increased explainability as semantic (Poli et al., 2021). The XAI tools present their output mostly through charts like in the case of many papers used SHAP and LIME. We argue for the value of multimodal output data of XAI tool as complementary output formats may be easier to understand.

Another key outcome is the limited user interfaces for XAI tools and their interactivity, despite the argue need for such interfaces (Chromik & Butz, 2021; Chromik & Schuessler, 2020; Laato et al., 2021; Nazar et al., 2021; Weber et al., 2023). Some recently designed XAI tools aim to provide such interfaces (Malandri et al., 2022), (Mercorio et al., 2020). These outcomes open up opportunities for exploring the design principles of user interfaces for XAI tools (Chromik & Schuessler, 2020; Laato et al., 2021) such as human-centred XAI approaches (Ehsan et al., 2022), which previous research has also argued for (Chromik & Butz, 2021; Laato et al., 2021).

For instance, XAI user interface should present explanations in a concise and accessible manner, utilizing clear language, visualizations, and interactive elements to convey complex information effectively. It should provide users with the ability to navigate through different levels of granularity in explanations and explore alternative scenarios (Mucha et al., 2021). The UI should also offer customization features, enabling users to tailor the presentation of explanations to their specific needs and preferences (Baniecki et al., 2023). Additionally, the UI should be responsive and provide real-time feedback, allowing users to observe how their interactions influence the AI system's outputs. Overall, a well-designed XAI UI should foster transparency, promote trust, engagement, and support decision-making (Antoniadi et al., 2021).

A key outcome is the limited use of HCI knowledge for the design and evaluation of XAI tools. We argue for the value of leveraging HCI methods and expertise, to ensure that the XAI tools and their interfaces can be accessed and adopted by users with limited AI expertise, such as end users or domain experts, i.e., clinicians. Here, we can think of developing tailored heuristics to support expert evaluation of XAI tools. Articulating such heuristics will be much-needed starting point for identifying the limitations of such tools, and subsequently, user-centred design approaches could be leveraged to design novel XAI interfaces that meet users' needs for explanations (Naiseh et al., 2024; Oliveira et al., 2023).

4.3. Towards a design space for AI-HCI research

To address the third research question on the HCI opportunities with respect to XAI tools, we advance an initial design space for AI-HCI research integrating AI affordances with the application domains of XAI tools mapped to key HCI research areas (Fig 3).

The application domains of medicine, finance, energy, education, or cybersecurity could benefit from bespoke interfaces for XAI tools. For instance, XAI tools have been used to explain ML prediction of heart failure survival, (Moreno-Sanchez, 2020) predict stock market trends with explanations (Mandeep et al., 2022), predict the earliest possible interpretation of students' performance in the virtual learning environment (Adnan et al., 2022), or explain intrusion detection-based convolutional neural networks (Younisse et al., 2022), albeit without XAI interfaces which could lower the entry barrier of domain experts with limited AI knowledge. Furthermore, the exploration of this space is extending research using AI models with descriptions of these models inputted into XAI tools.

For instance, (Chalabianloo et al., 2022) employed several machine learning models on a range of biodata to detect and classify stress. They used SHAP XAI tool to explain the models and their outcomes. Such approaches can be further extended through the use of multiple XAI tools as previously suggested for mental health apps (Alotaibi & Sas, 2024), so that their comparative advantages and shortcomings can be better understood by both experts and non experts.

5. LIMITATION AND FUTURE RESEARCH

As a systematic reviews, this study was limited to papers returned by our search terms which may be seen as narrow. Future work could extend it to include others such as “explainable AI tool”, or “explainable artificial intelligence tool”. Similar to other systematic reviews targeting academic work, ours does not include grey literature which may have been relevant (Bach et al., 2024).

With regard to future work, another important direction is the exploration of the design space for AI-HCI. This is an interdisciplinary research agenda which could leverage HCI design and evaluation methods, with design guidelines for visualization and user mental models, alongside XAI expertise (Hoffman et al., 2023; Muhammad, et al., 2023; Yüksel et al., 2023).

6. CONCLUSION

The potential of AI technologies relies on improved explainability of machine learning black-box models. Despite their significant growth, the exploration of XAI tools in HCI has been limited. We conducted a systematic review of 142 papers focused on XAI tools. Findings indicate a broad range of XAI tools but extensive use of a few, and prevalence of AI experts as users rather than evaluators of these tools. Findings also show five main application domains of XAI tools, their emphasis on input data in numerical or image format, and of output data in charts format, as well as limited provision of user interfaces for these tools. We discuss our findings arguing for increased use and comparative evaluation of exiting tools, and the design of HCI grounded user interfaces for XAI tools. We also advanced a design space for AI-HCI research.

7. ACKNOWLEDGMENT

This work was supported by Tabuk University, Saudi Arabia and the Saudi Arabian Cultural Bureau, London, and by Digital Health: Innovative engineering technologies to improve the understanding and management of fatigue funded by EPSRC. Grant reference EP/W003228/1.

8. REFERENCES

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*. https://doi.org/10.1109/ACCESS.2018.2870052

Adnan, M., Uddin, M. I., Khan, E., Alharithi, F. S., Amin, S., & Alzahrani, A. A. (2022). Earliest Possible Global and Local Interpretation of Students’ Performance in Virtual Learning Environment by Leveraging Explainable AI. *IEEE Access*, *10*(December), 129843–129864. https://doi.org/10.1109/ACCESS.2022.3227072

Albahri, A. S., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., … Deveci, M. (2023). A Systematic Review of Trustworthy and Explainable Artificial Intelligence in Healthcare: Assessment of Quality, Bias Risk, and Data Fusion. *Information Fusion*, *96*(January), 156–191. https://doi.org/10.1016/j.inffus.2023.03.008

Alonso, J. M., & Casalino, G. (2019). Explainable Artificial Intelligence for Human-Centric Data Analysis in Virtual Learning Environments. *Communications in Computer and Information Science*, *1091*(September), 125–138. https://doi.org/10.1007/978-3-030-31284-8_10

Alotaibi, A., & Sas, C. (2023). Review of AI-Based Mental Health Apps. in Proceedings of British HCI Conference 2023, 13 pages, DOI: 10.14236/ewic/BCSHCI2023.27

Alperin, K. B., Wollaber, A. B., & Gomez, S. R. (2020). Improving Interpretability for Cyber Vulnerability Assessment Using Focus and Context Visualizations. *2020 IEEE Symposium on Visualization for Cyber Security, VizSec 2020*, 30–39. https://doi.org/10.1109/VizSec51108.2020.00011

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(5). https://doi.org/10.1002/widm.1424

Anjomshoae, S., Calvaresi, D., Najjar, A., & Främling, K. (2019). Explainable agents and robots: Results from a systematic literature review. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, *2* (Aamas), 1078–1088. https://doi.org/10.5555/3306127.3331806

Antoniadi, A. M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. A., & Mooney, C. (2021). Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: A systematic review. *Applied Sciences (Switzerland)*, *11*(11), 1–23. https://doi.org/10.3390/app11115088

Bach, T. A., Khan, A., Hallock, H., Beltrão, G., & Sousa, S. (2024). A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective. *International Journal of Human-Computer Interaction*, *40*(5), 1251–1266. https://doi.org/10.1080/10447318.2022.2138826

Baniecki, H., Parzych, D., & Biecek, P. (2023). The grammar of interactive explanatory model

analysis. *Data Mining and Knowledge Discovery*, (January). https://doi.org/10.1007/s10618-023-00924-w

Barredo, A., Díaz-Rodríguez, N., Del, J., Bennetot, A., Tabik, S., Barbado, A., … Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58* (October 2019), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bertini, F., Dal Palù, A., Fabiano, F., & Iotti, E. (2022). CARING for xAI. *CEUR Workshop Proceedings*, *3204*, 47–60.

Bistarelli, S., Mancinelli, A., Santini, F., & Taticchi, C. (2022). Arg-XAI: a Tool for Explaining Machine Learning Results. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, *2022-Octob*, 205–212. https://doi.org/10.1109/ICTAI56018.2022.00037

Brdnik, S. (2023). GUI Design Patterns for Improving the HCI in Explainable Artificial Intelligence. *International Conference on Intelligent User Interfaces, Proceedings IUI*, 240–242. https://doi.org/10.1145/3581754.3584114

Carta, S., Consoli, S., Podda, A. S., Reforgiato Recupero, D., & Stanciu, M. M. (2022). An eXplainable Artificial Intelligence tool for statistical arbitrage. *Software Impacts*, *14*(June), 100354. https://doi.org/10.1016/j.simpa.2022.100354

Chalabianloo, N., Can, Y. S., Umair, M., Sas, C., & Ersoy, C. (2022). Application level performance evaluation of wearable devices for stress classification with explainable AI. *Pervasive and Mobile Computing*, *87*, 101703. https://doi.org/10.1016/j.pmcj.2022.101703

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*, *2018-January*, 839–847. https://doi.org/10.1109/WACV.2018.00097

Chromik, M., & Butz, A. (2021). Human-XAI Interaction: A Review and Design Principles for Explanation User Interfaces. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12933 LNCS*, 619–640. https://doi.org/10.1007/978-3-030-85616-8_36

Chromik, M., & Schuessler, M. (2020). A taxonomy for human subject evaluation of black-box explanations in XAI. *CEUR Workshop Proceedings*, *2582*.

Clement, T., Kemmerzell, N., Abdelaal, M., & Amberg, M. (2023). XAIR: A Systematic Metareview of Explainable AI (XAI) Aligned to the Software Development Process. *Machine Learning and Knowledge Extraction*, *5*(1), 78–108. https://doi.org/10.3390/make5010006

Dieber, J., & Kirrane, S. (2020). Why model why? Assessing the strengths and limitations of LIME. *ArXiv, abs/2012.00093*.

Dindorf, C., Konradi, J., Wolf, C., Taetz, B., Bleser, G., Huthwelker, J., … Fröhlich, M. (2021). Classification and automated interpretation of spinal posture data using a pathology-independent classifier and explainable artificial intelligence (Xai). *Sensors*, *21*(18), 1–16. https://doi.org/10.3390/s21186323

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … Houlsby, N. (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*.

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding explainability: Towards social transparency in ai systems. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3411764.3445188

Ehsan, U., Wintersberger, P., Liao, Q. V., Watkins, E. A., Manger, C., Daumé, H., … Riedl, M. O. (2022). Human-Centered Explainable AI (HCXAI): Beyond Opening the Black-Box of AI. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3491101.3503727

Ehsan, U., Wintersberger, P., Watkins, E. A., Ramos, G., Weisz, J. D., Riener, A., & Riedl, M. O. (2023). *Human-Centered Explainable AI (HCXAI): Coming of Age*. https://doi.org/10.1145/3544549.3573832

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods*, *5*(1), 80–92. https://doi.org/10.1177/160940690600500107

Fouladgar, N., Alirezaie, M., & Framling, K. (2022). Metrics and Evaluations of Time Series Explanations: An Application in Affect Computing. *IEEE Access*, *10*, 23995–24009. https://doi.org/10.1109/ACCESS.2022.3155115

Galli, A., Piscitelli, M. S., Moscato, V., & Capozzoli, A. (2022). Bridging the gap between complexity and interpretability of a data analytics-based process for benchmarking energy performance of buildings. *Expert Systems with Applications*, *206*(June 2021), 117649. https://doi.org/10.1016/j.eswa.2022.117649

Gandolfi, M., Galazzo, I. B., Pavan, R. G., Cruciani, F., Vale, N., Picelli, A., … Menegaz, G. (2023). eXplainable AI Allows Predicting Upper Limb Rehabilitation Outcomes in Sub-Acute Stroke Patients. *IEEE Journal of Biomedical and Health Informatics*, *27*(1), 263–273. https://doi.org/10.1109/JBHI.2022.3220179

Gulmezoglu, B. (2022). XAI-Based Microarchitectural Side-Channel Analysis for Website Fingerprinting Attacks and Defenses. *IEEE Transactions on Dependable and Secure Computing*, *19*(6), 4039–4051. https://doi.org/10.1109/TDSC.2021.3117145

Heberle, H., Zhao, L., Schmidt, S., Wolf, T., & Heinrich, J. (2023). XSMILES: interactive visualization for molecules, SMILES and XAI attribution scores. *Journal of Cheminformatics*, *15*(1), 1–12. https://doi.org/10.1186/s13321-022-00673-w

Heimerl, A., Baur, T., Lingenfelser, F., Wagner, J., & Andre, E. (2019). NOVA - A tool for eXplainable Cooperative Machine Learning. *2019 8th International Conference on Affective Computing and Intelligent Interaction, ACII 2019*, (Xcml). https://doi.org/10.1109/ACII.2019.8925519

Henriksen, E., Halden, U., Kuzlu, M., & Cali, U. (2022). Electrical Load Forecasting Utilizing an Explainable Artificial Intelligence (XAI) Tool on Norwegian Residential Buildings. *SEST 2022 - 5th International Conference on Smart Energy Systems and Technologies*, 1–6. https://doi.org/10.1109/SEST53650.2022.9898500

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science*, *5*. https://doi.org/10.3389/fcomp.2023.1096257

Holzinger, A., Carrington, A., & Müller, H. (2020). Measuring the Quality of Explanations: The System Causability Scale (SCS): Comparing Human and Machine Explanations. *KI - Kunstliche Intelligenz*, *34*(2), 193–198. https://doi.org/10.1007/s13218-020-00636-z

Islam, M. S., Hussain, I., Rahman, M. M., Park, S. J., & Hossain, M. A. (2022). Explainable Artificial Intelligence Model for Stroke Prediction Using EEG Signal. *Sensors*, *22*(24). https://doi.org/10.3390/s22249859

Jayakumar, K., & Skandhakumar, N. (2022). A Visually Interpretable Forensic Deepfake Detection Tool Using Anchors. *7th International Conference on Information Technology Research: Digital Resilience and Reinvention, ICITR 2022 - Proceedings*. https://doi.org/10.1109/ICITR57877.2022.9993294

Kadir, M. A., Mohamed Selim, A., Barz, M., & Sonntag, D. (2023). A User Interface for Explaining Machine Learning Model Explanations. *International Conference on Intelligent User Interfaces, Proceedings IUI*, 59–63. https://doi.org/10.1145/3581754.3584131

Kapcia, M., Eshkiki, H., Duell, J., Fan, X., Zhou, S., & Mora, B. (2021). ExMed: An AI Tool for Experimenting Explainable AI Techniques on Medical Data Analytics. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, *2021-Novem*, 841–845. https://doi.org/10.1109/ICTAI52525.2021.00134

Keleko, A. T., Kamsu-Foguem, B., Ngouna, R. H., & Tongne, A. (2023). Health condition monitoring of a complex hydraulic system using Deep Neural Network and DeepSHAP explainable XAI. *Advances in Engineering Software*, *175*(January 2022), 103339. https://doi.org/10.1016/j.advengsoft.2022.103339

Kim, J. K., Bae, M. N., Lee, K., Kim, J. C., & Hong, S. G. (2022). Explainable Artificial Intelligence and Wearable Sensor-Based Gait Analysis to Identify Patients with Osteopenia and Sarcopenia in Daily Life. *Biosensors*, *12*(3). https://doi.org/10.3390/bios12030167

Konradi, J., Zajber, M., Betz, U., Drees, P., Gerken, A., & Meine, H. (2022). AI-Based Detection of Aspiration for Video-Endoscopy with Visual Aids in Meaningful Frames to Interpret the Model Outcome. *Sensors*, *22*(23). https://doi.org/10.3390/s22239468

Kumar, A., Manikandan, R., Kose, U., Gupta, D., & Satapathy, S. C. (2021). Doctor's dilemma: Evaluating an explainable subtractive spatial lightweight convolutional neural network for brain tumor diagnosis. *ACM Transactions on Multimedia Computing, Communications and Applications*, *17*(3s). https://doi.org/10.1145/3457187

Kuzlu, M., Cali, U., Sharma, V., & Güler, Ö. (2020). Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools. *IEEE Access*, *8*, 187814–187823. https://doi.org/10.1109/ACCESS.2020.3031477

Laato, S., Tiainen, M., Najmul Islam, A. K. M., & Mäntymäki, M. (2021). How to explain AI systems to end users: a systematic literature review and research agenda. *Internet Research*, *32*(7), 1–31. https://doi.org/10.1108/INTR-08-2021-0600

Langley, P., Meadows, B., Sridharan, M., & Choi, D. (2017). Explainable Agency for Intelligent Autonomous Systems. *Proceedings of the AAAI*

Conference on Artificial Intelligence, *31*(2), 4762–4763. https://doi.org/10.1609/aaai.v31i2.19108

Leung, C. K., Pazdor, A. G. M., & Souza, J. (2021). Explainable Artificial Intelligence for Data Science on Customer Churn. *2021 IEEE 8th International Conference on Data Science and Advanced Analytics, DSAA 2021*. https://doi.org/10.1109/DSAA53316.2021.9564166

Liang, Y., Li, S., Yan, C., Li, M., & Jiang, C. (2021). Explaining the black-box model: A survey of local interpretation methods for deep neural networks. *Neurocomputing*, *419*, 168–182. https://doi.org/10.1016/j.neucom.2020.08.011

Longo, L., Goebel, R., Lecue, F., Kieseberg, P., & Holzinger, A. (2020). Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12279 LNCS*, 1–16. https://doi.org/10.1007/978-3-030-57321-8_1

Lopes, P., Silva, E., Braga, C., Oliveira, T., & Rosado, L. (2022). XAI Systems Evaluation: A Review of Human and Computer-Centred Methods. *Applied Sciences (Switzerland)*, *12*(19). https://doi.org/10.3390/app12199423

Love, P. E. D., Fang, W., Matthews, J., Porter, S., Luo, H., & Ding, L. (2022). Explainable Artificial Intelligence: Precepts, Methods, and Opportunities for Research in Construction. *ArXiv Preprint ArXiv:2211.06579*, 1–58.

Lucieri, A., Bajwa, M. N., Braun, S. A., Malik, M. I., Dengel, A., & Ahmed, S. (2022). ExAID: A multimodal explanation framework for computer-aided diagnosis of skin lesions. *Computer Methods and Programs in Biomedicine*, *215*. https://doi.org/10.1016/j.cmpb.2022.106620

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *2017-Decem*(Section 2), 4766–4775.

Malandri, L., Mercorio, F., Mezzanzanica, M., & Nobani, N. (2022). ConvXAI: a System for Multimodal Interaction with Any Black-box Explainer. In *Cognitive Computation*. Springer US. https://doi.org/10.1007/s12559-022-10067-7

Mandeep, Agarwal, A., Bhatia, A., Malhi, A., Kaler, P., & Pannu, H. S. (2022). Machine Learning Based Explainable Financial Forecasting. *2022 4th International Conference on Computer Communication and the Internet, ICCCI 2022*, 34–38. https://doi.org/10.1109/ICCCI55554.2022.9850272

Mercorio, F., Mezzanzanica, M., & Seveso, A. (2020). eXDiL: A Tool for Classifying and eXplaining Hospital Discharge Letters. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12279 LNCS*(DiL), 159–172. https://doi.org/10.1007/978-3-030-57321-8_9

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2010). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *International Journal of Surgery*, *8*(5), 336–341. https://doi.org/10.1016/j.ijsu.2010.02.007

Moreno-Sanchez, P. A. (2020). Development of an Explainable Prediction Model of Heart Failure Survival by Using Ensemble Trees. *Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020*, 4902–4910. https://doi.org/10.1109/BigData50022.2020.9378460

Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications*, *165*(May 2020), 113986. https://doi.org/10.1016/j.eswa.2020.113986

Mucha, H., Robert, S., Breitschwerdt, R., & Fellmann, M. (2021). Interfaces for Explanations in Human-AI Interaction: Proposing a Design Evaluation Approach. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3411763.3451759

Muhammad, A. P., Knauss, E., & Bärgman, J. (2023). Human factors in developing automated vehicles: A requirements engineering perspective. *Journal of Systems and Software*, *205*, 111810. https://doi.org/10.1016/j.jss.2023.111810

Muhammad, K., Ullah, A., Lloret, J., Ser, J. Del, & De Albuquerque, V. H. C. (2021). Deep Learning for Safe Autonomous Driving: Current Challenges and Future Directions. *IEEE Transactions on Intelligent Transportation Systems*, *22*(7). https://doi.org/10.1109/TITS.2020.3032227

Nagy, M., & Molontay, R. (2023). Interpretable Dropout Prediction: Towards XAI-Based Personalized Intervention. *International Journal of Artificial Intelligence in Education*, (0123456789). https://doi.org/10.1007/s40593-023-00331-8

Naiseh, M., Jiang, N., Ma, J., & Ali, R. (2020). Explainable Recommendations in Intelligent Systems: Delivery Methods, Modalities and Risks. *Lecture Notes in Business Information Processing*, *385 LNBIP*(March), 212–228. https://doi.org/10.1007/978-3-030-50316-1_13

Naiseh, M., Simkute, A., Zieni, B., Jiang, N., & Ali, R. (2024). C-XAI: A Conceptual Framework for

Designing XAI tools that Support Trust Calibration. *Journal of Responsible Technology*, 100076. https://doi.org/10.1016/j.jrt.2024.100076

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., … Seifert, C. (2023). From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys*. https://doi.org/10.1145/3583558

Nazar, M., Alam, M. M., Yafi, E., & Su'Ud, M. M. (2021). A Systematic Review of Human-Computer Interaction and Explainable Artificial Intelligence in Healthcare with Artificial Intelligence Techniques. *IEEE Access*, *9*, 153316–153348. https://doi.org/10.1109/ACCESS.2021.3127881

Oduor, E., Qian, K., Li, Y., & Popa, L. (2020). XAIT: An interactivewebsite for explainable ai for text. *International Conference on Intelligent User Interfaces, Proceedings IUI*, 120–121. https://doi.org/10.1145/3379336.3381468

Oliveira, E., Braga, C., Sampaio, A., Oliveira, T., Soares, F., & Rosado, L. (2023). Designing XAI-based Computer-aided Diagnostic Systems: Operationalising User Research Methods. *CEUR Workshop Proceedings*, *3359*, 25–36.

Panigutti, C., Beretta, A., Fadda, D., Giannotti, F., Pedreschi, D., Perotti, A., & Rinzivillo, S. (2023). Co-design of human-centered, explainable AI for clinical decision support. *ACM Transactions on Interactive Intelligent Systems*. https://doi.org/10.1145/3587271

Poli, J.-P., Ouerdane, W., & Pierrard, R. (2021). Generation of Textual Explanations in XAI: the Case of Semantic Annotation. *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–6. IEEE. https://doi.org/10.1109/FUZZ45933.2021.9494589

Qian, K., Danilevsky, M., Katsis, Y., Kawas, B., Oduor, E., Popa, L., & Li, Y. (2021). XNLP: A living survey for XAI research in natural language processing. *International Conference on Intelligent User Interfaces, Proceedings IUI*, 78–80. https://doi.org/10.1145/3397482.3450728

Ras, G., van Gerven, M., & Haselager, P. (2018). *Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges*. https://doi.org/10.1007/978-3-319-98131-4_2

Roy, I., Feng, B., Roychowdhury, S., Ravi, S. K., Umretiya, R. V., Reynolds, C., … Hoffman, A. (2023). Understanding oxidation of Fe-Cr-Al alloys through explainable artificial intelligence. *MRS Communications*, *13*(1), 82–88. https://doi.org/10.1557/s43579-022-00315-0

Saeed, W., & Omlin, C. (2021). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, *263*(Dl). https://doi.org/10.1016/j.knosys.2023.110273

Sanches, P., Janson, A., Karpashevich, P., Nadal, C., Qu, C., Daudén Roquet, C., … Sas, C. (2019). HCI and Affective Health. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–17. New York, NY, USA: ACM. https://doi.org/10.1145/3290605.3300475

Šarčević, A., Pintar, D., Vranić, M., & Krajna, A. (2022). Cybersecurity Knowledge Extraction Using XAI. *Applied Sciences (Switzerland)*, *12*(17). https://doi.org/10.3390/app12178669

Sarp, S., Catak, F. O., Kuzlu, M., Cali, U., Kusetogullari, H., Zhao, Y., … Guler, O. (2023). An XAI approach for COVID-19 detection using transfer learning with X-ray images. *Heliyon*, *9*(4), e15137. https://doi.org/10.1016/j.heliyon.2023.e15137

Sarp, S., Kuzlu, M., Cali, U., Elma, O., & Guler, O. (2021). An Interpretable Solar Photovoltaic Power Generation Forecasting Approach Using An Explainable Artificial Intelligence Tool. *2021 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, (ii), 1–5. IEEE. https://doi.org/10.1109/ISGT49243.2021.9372263

Sarp, S., Kuzlu, M., Wilson, E., Cali, U., & Guler, O. (2021). The enlightening role of explainable artificial intelligence in chronic wound classification. *Electronics (Switzerland)*, *10*(12). https://doi.org/10.3390/electronics10121406

Schoonderwoerd, T. A. J., Jorritsma, W., Neerincx, M. A., & van den Bosch, K. (2021). Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. *International Journal of Human Computer Studies*, *154*, 102684. https://doi.org/10.1016/j.ijhcs.2021.102684

Schwalbe, G., & Finzel, B. (2023). A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*. https://doi.org/10.1007/s10618-022-00867-8

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, *128*(2), 618–626. https://doi.org/10.1109/ICCV.2017.74

Shree, S., Chandrasekaran, J., Lei, Y., Kacker, R. N., & Kuhn, D. R. (2022). DeltaExplainer: A Software Debugging Approach to Generating Counterfactual Explanations. *Proceedings - 4th IEEE International Conference on Artificial*

Intelligence Testing, AITest 2022, 103–110. https://doi.org/10.1109/AITest55621.2022.00023

Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020). *Explainable Deep Learning Models in Medical Image Analysis*. 1–19. https://doi.org/10.3390/jimaging6060052

Skuppin, N., Hoffmann, E. J., Shi, Y., & Zhu, X. X. (2022). *EXPLAINABILITY ANALYSIS OF CNN IN DETECTION OF VOLCANIC DEFORMATION SIGNAL*. 5844–5847.

Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., … Catanzaro, B. (2022). *Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model*.

Speith, T. (2022). A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. *In 2022 ACM Conference on Fairness, Accountability, AndTransparency (FAccT'22)*. https://doi.org/https://doi.org/10.1145/3531146.3534639 1

Suh, B., Yu, H., Kim, H., Lee, S., Kong, S., Kim, J. W., & Choi, J. (2023). Interpretable Deep-Learning Approaches for Osteoporosis Risk Screening and Individualized Feature Analysis Using Large Population-Based Data: Model Development and Performance Evaluation. *Journal of Medical Internet Research*, *25*. https://doi.org/10.2196/40179

Thieme, A., Belgrave, D., & Doherty, G. (2020). Machine Learning in Mental Health. *ACM Transactions on Computer-Human Interaction*, *27*(5), 1–53. https://doi.org/10.1145/3398069

Tjoa, E., & Guan, C. (2021). A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314

Vieira, C. P., & Digiampietri, L. A. (2022). Machine Learning post-hoc interpretability: a systematic mapping study. *ACM International Conference Proceeding Series*, *Par F18047*. https://doi.org/10.1145/3535511.3535512

Villaronga, E. F., Kieseberg, P., & Li, T. (2018). Humans forget, machines remember: Artificial intelligence and the Right to Be Forgotten. *Computer Law and Security Review*, *34*(2), 304–313. https://doi.org/10.1016/j.clsr.2017.08.007

Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, *76*(April), 89–106. https://doi.org/10.1016/j.inffus.2021.05.009

Wang, C., & An, P. (2021). A Mobile Tool that Helps Nonexperts Make Sense of Pretrained CNN by Interacting with Their Daily Surroundings. *Extended Abstracts of MobileHCI 2021 - ACM International Conference on Mobile Human-Computer Interaction: Mobile Apart, Mobile Together*. https://doi.org/10.1145/3447527.3474873

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. *Conference on Human Factors in Computing Systems - Proceedings*, (February). https://doi.org/10.1145/3290605.3300831

Wang, Q., Huang, K., Chandak, P., Zitnik, M., & Gehlenborg, N. (2023). Extending the Nested Model for User-Centric XAI: A Design Study on GNN-based Drug Repurposing. *IEEE Transactions on Visualization and Computer Graphics*, *29*(1), 1266–1276. https://doi.org/10.1109/TVCG.2022.3209435

Weber, P., Carl, K. V., & Hinz, O. (2023). Applications of Explainable Artificial Intelligence in Finance—a systematic review of Finance, Information Systems, and Computer Science literature. In *Management Review Quarterly*. Springer International Publishing. https://doi.org/10.1007/s11301-023-00320-0

Wellawatte, G. P., Gandhi, H. A., Seshadri, A., & White, A. D. (2022). A Perspective on Explanations of Molecular Prediction Models. *Chemrxiv*. https://doi.org/10.1021/acs.jctc.2c01235

Xu, W. (2019). Toward human-centered AI: A perspective from human-computer interaction. *ACM*, *26*(4), 42–46. https://doi.org/10.1145/3328485

Younisse, R., Ahmad, A., & Abu Al-Haija, Q. (2022). Explaining Intrusion Detection-Based Convolutional Neural Networks Using Shapley Additive Explanations (SHAP). *Big Data and Cognitive Computing*, *6*(4). https://doi.org/10.3390/bdcc6040126

Yüksel, N., Börklü, H. R., Sezer, H. K., & Canyurt, O. E. (2023). Review of artificial intelligence applications in engineering design perspective. *Engineering Applications of Artificial Intelligence*, *118*(April 2022), 105697. https://doi.org/10.1016/j.engappai.2022.105697

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning Deep Features for Discriminative Localization. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-December*, 2921–2929. https://doi.org/10.1109/CVPR.2016.319