Assertiveness-based Agent Communication for a Personalized Medicine on Medical Imaging Diagnosis

Francisco Maria Calisto
Institute for Systems and Robotics
Lisbon, Portugal
francisco.calisto@tecnico.ulisboa.pt

João Fernandes
Instituto Superior Técnico
Lisbon, Portugal
joao.g.m.fernandes@tecnico.ulisboa.pt

Margarida Morais
Instituto Superior Técnico
Lisbon, Portugal
margarida.p.morais@tecnico.ulisboa.pt

Carlos Santiago
Institute for Systems and Robotics
Lisbon, Portugal
carlos.santiago@tecnico.ulisboa.pt

João Maria Abrantes
Centro Hospitalar de Trás-Os-Montes
e Alto Douro
Vila Real, Portugal
jmvabrantes@chtmad.min-saude.pt

Nuno Nunes
Interactive Technologies Institute
Lisbon, Portugal
nunojnunes@tecnico.ulisboa.pt

Jacinto C. Nascimento
Institute for Systems and Robotics
Lisbon, Portugal
jan@isr.tecnico.ulisboa.pt

Figure 1: A sample of "Non-Assertive" (*e.g.*, suggesting "it looks like") *vs.* "Assertive" (*e.g.*, imposing "must") communications.

ABSTRACT

Intelligent agents are showing increasing promise for clinical decision-making in a variety of healthcare settings. While a substantial body of work has contributed to the best strategies to convey these agents' decisions to clinicians, few have considered the impact of personalizing and customizing these communications on the clinicians' performance and receptiveness. This raises the question of how intelligent agents should adapt their tone in accordance with their target audience. We designed two approaches to communicate the decisions of an intelligent agent for breast cancer diagnosis with different tones: a suggestive (non-assertive) tone and an imposing (assertive) one. We used an intelligent agent to inform about: (1) number of detected findings; (2) cancer severity on each breast and per medical imaging modality; (3) visual scale representing severity estimates; (4) the sensitivity and specificity of the agent; and (5) clinical arguments of the patient, such as pathological co-variables. Our results demonstrate that assertiveness plays an important role in how this communication is perceived and its benefits. We show that personalizing assertiveness according to the professional experience of each clinician can reduce medical errors and increase satisfaction, bringing a novel perspective to the design of adaptive communication between intelligent agents and clinicians.

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3580682

CCS CONCEPTS

• **Human-centered computing** → *Human computer interaction (HCI)*; • **Computing methodologies** → **Intelligent agents**; • **Applied computing** → **Health informatics**.

KEYWORDS

Clinical Decision Support System, Healthcare, Breast Cancer

ACM Reference Format:
Francisco Maria Calisto, João Fernandes, Margarida Morais, Carlos Santiago, João Maria Abrantes, Nuno Nunes, and Jacinto C. Nascimento. 2023. Assertiveness-based Agent Communication for a Personalized Medicine on Medical Imaging Diagnosis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3544548.3580682

1 INTRODUCTION

Artificial Intelligence (AI) systems are showing an increasing promise for numerous healthcare applications. Recently, the advantages of Deep Learning (DL) are spawning AI systems with human-like performance in several clinical domains [26, 45, 88, 107]. However, these applications are not designed to capture the variability of personal or subpopulation level of clinicians (*e.g.*, interns, juniors, middles, and seniors) [116]. In fact, recent works are highlighting how AI and the advancement of technologies together are empowering the aim of personalized and precision medicine [46, 111, 120]. Given the need of personalizing and customizing the AI recommendations, an important question in the design of AI systems is how they should communicate (Figure 1), considering the professional experience of the clinician.

As the means to achieve a more persuasive and reliable intelligent agent, we need to analyze and collect data regarding the clinician's behavior [83]. Communication is essential to increase the reliability of an intelligent agent [81] providing a diagnosis to a clinician. One way to achieve that is by aligning the levels of assertiveness [79] of the agent with the years of experience of the clinician. Besides that, providing an explanation of 'how' and 'why' the AI assistant achieved a certain output increases trust on the system, solving a problem known as the "black-box" problem [25, 80].

In this paper, we present our study for applying *BreastScreening-AI* (Section A.4) in two conditions, where clinicians will interact with conventional and assertiveness-based intelligent agents [79, 81]. The assistant will act as a second reader, resulting in improvements in diagnostic performance, by reducing false-positives and false-negatives (*i.e.*, Over-Diagnosis *vs* Under-Diagnosis), as well as in efficiency and efficacy in the clinical workflow (Section A.5). While considerable work has focused on improving the accuracy of AI algorithms, comparatively less work focused on improving adoption and usability of interactive assistance techniques. This paper contributes broadly to the latter by examining what clinicians need when using AI-powered assistance, the practices they adopt while using diagnostic tools, and how these tools affect the end-user attitudes towards the underlying AI algorithms.

Here, we detail a within-subject study with 52 clinicians who interacted with both conventional and assertiveness-based agents, diagnosing a total of 35 patients from a dataset of 289 patients, out of which 34% have benign abnormalities, 31% have malignant abnormalities, and the other 35% are healthy patients (Section A.1). Two different tones were used to communicate the AI recommendations in our assertiveness-based agent, from a more suggestive to a more imposing tone. Moreover, the assertiveness-based agent explained '*how*' and '*why*' the AI algorithms achieved a particular diagnostic by providing human-interpretable clinical arguments for the achieved outputs.

While we used real AI outputs and clinical arguments curated from human clinicians, the AI models were trained for classification and segmentation purposes through different architectures [44]. Specifically, through a DenseNet model [49] for a multimodal diagnosis of MammoGraphy (MG) and UltraSound (US) images. Similarly, a 3D ResNet model [5] was trained to diagnose the Magnetic Resonance Imaging (MRI) volumes. Our findings suggest that explaining the AI outputs and clinical arguments by exploring how to adapt the communication through an assertiveness-based agent can benefit AI assistance of medical reasoning.

The novelty of this work is in the application of communication theories through the use of DL systems. Our goal was to explore and understand how assertiveness-based AI mediation is affecting different expertise levels in a high-stake decision-making domain, such as critical medical decisions dealing with the life of a patient. For that purpose, we updated the design of the *BreastScreening-AI* framework [26] in where we aim to study how does different expertise levels in clinical scenarios are influenced by an assertiveness-based communication. Not only we are contributing with knowledge of computational interaction approaches in the HCI filed (*i.e.*, assertiveness-based AI mediation), but also understanding on how to design interactive systems underpinned by computational principles to the CHI community.

In sum, the main contributions of this work are as follows:

(1) We present a novel approach for personalizing and customizing the AI-assisted medical reasoning, providing evidence that assertiveness-based agents can alter clinical workflows by effectively adapting the communication depending on the categories of medical professional experience.
(2) We demonstrate that while explaining the AI outputs can contribute to enhancing medical efficiency, its impact heavily depends on the communication tone (*i.e.*, more suggestive or imposing the AI recommendations) of the provided clinical arguments.
(3) We report our results demonstrating that these assertiveness-based agents can increase the utility of clinical information found and increase user trust in the AI recommendations, without a loss in diagnostic performance.
(4) We provide design considerations for adapting the communication in AI-assisted medical reasoning, laying a foundation for future implementations of intelligent agents better capable of personalizing and customizing explanations.

Across the following sections, we outline related works on the issues of guiding the Human-AI Interaction (HAII) topic, assisting clinical decision-making, going through some examples of Clinical Decision Support Systems (CDSS) present in the literature [41, 76], and ending on the effects of AI communication. We then introduce the design of our *Assertiveness-based BreastScreening-AI* assistant, followed by our research questions, hypotheses, and methods. Last, we report our quantitative and qualitative findings, as well as concluding with a discussion of design considerations.

2 RELATED WORK

Medical imaging systems allow the end-user to diagnose several modalities, such as MG, US, or MRI, from a seamless retrieval of medical imaging data [39]. Bringing those modalities together offers new possibilities for quantitative imaging and diagnoses, but also requires specialized data handling, post-processing, and novel visualization methods [51]. In the clinical domain, medical imaging tools can help experts make better decisions, *e.g.*, by identifying cancer prognostics among the available multi-modal data [50, 113]. In this document, we focus on understanding different aspects and expectations of a medical imaging CDSS integrated into the radiology workflow. In particular, our work demonstrates how an assertiveness-based interaction can improve the medical imaging diagnosis.

2.1 Human-AI Interaction

Intelligent agents need to provide users with, not only results, but also accounting for their behaviors during decision-making [121]. In the field of Human-Computer Interaction (HCI), the topic of eXplainable-AI (XAI) contains subjects that intersect cognitive psychology, learnability, and context awareness [98, 99]. Cognitive psychology is a subject focusing more on explanation theory [16]. For cognitive psychology, Lombrozo [62] found that cognitive explanations are strongly connected with causality reasoning. Learnability is an important part of usability [43]. Here, Abdul [2] summarized topics of learnibility related for designing a XAI system, such as hints, guidance, and visualizations.

Finally, explainable context awareness is a simplistic representation of the context to inform users what is obtained and which action will be done by the system [32]. Dey et al. [31] designed a tailored interface, providing visual and textual explanations following several context-aware rules. However, research in both HCI and AI communities is oftenly disconnected between the two fields [2, 121]. There is a research gap that is not crossing nor combining both fields to the interdisciplinary approach of accounting user's different behaviors during decision-making.

HAII incorporates human feedback in the model training process to create better Machine Learning (ML) models. In this document, we refer to the topic as HAII, that somehow is addressed by Amershi et al. [8] providing a set of design guidelines [22]. The work of Kocielnik et al. [55] is also addressing the study on the impact of several methods of expectation setting, and others studied the design for specific HAII scenarios [4]. While much of the mentioned prior work has employed handcrafted features [8, 55], we leverage the rich image data features automatically learned from DL algorithms.

Many researchers have argued that HAII would be improved if the AI systems could *explain their reasoning* [10, 53, 89]. In medicine, explaining predictions from AI models is particularly salient, where the uncovered patterns of the model can be more important than prediction performance [65]. Lundberg et al. [66] are demonstrating how to retain interpretability by developing a method to provide explanations of model predictions. Although these works are exploring how clinicians interact with AI recommendations and their perceptions of AI outcomes, they are not taking into account cognitive bias in decision-making. One of the most notorious cognitive differences is seen between people with different levels of expertise and knowledge [1, 95]. It is why we are studying how assertiveness-based agents are designed to adapt their communication tone based on expertise levels to reduce cognitive bias.

2.2 Assisting Clinical Decision-Making

Although the research in interaction with intelligent agents is recent [17], still this topic has seen new advances, *e.g.*, chat-bots and other agents [74]. Recent advances in medical technologies that promote the generation of data have continued to drive interaction research in the clinical domain [9, 64]. Moreover, the new interest of the medical community to support AI research projects and the available public *datasets*, are encouraging researchers to work in both fields [58]. Therefore, we bring together both HCI and AI communities to leverage the high-stakes of clinical decision-making.

The introduction of technology for assisting clinical decision-making is fraught with challenges and unintended consequences, such as critical decisions dealing with patient safety, clinician fatigue, and increased medical errors [6, 96, 102]. Moreover, clinicians find AI systems challenging to use because they may have limited technical skills for the adoption of these novel technologies, where these technologies are not customized to behavioral aspects of clinicians [23]. In fact, the AI outcomes are challenging to understand and communicate to clinicians, as these systems often have poorly designed interfaces [110], without taking into account differences of clinician's characteristics during decision-making. For instance, the reasoning of a novice clinician vary from an expert [33].

There is a lack of large-scale deployment of these systems in healthcare [109, 118, 126], making it difficult to understand how these systems are perceived and used by their intended users in real-world settings. HCI has proposed and conceptualized several approaches of human-AI relationships, such as interactive Machine Learning (iML) [38], Human-In-The-Loop (HITL) [48, 117], human-AI symbiosis [52], and human-AI collaboration [118]. However, these approaches mainly use human input to improve the prediction accuracy, model efficiency, and interpretability of AI to the unwanted added burdens on healthcare professionals [110, 122]. Wang et al. [118] studied the perception and usage of AI systems for assisting clinical decision-making, but the work is not accounting the potential differences between behavioral reasoning. Similarly, the work of Panigutti et al. [80] is only considering accurate algorithmic suggestions without considering the clinician's professional medical experience. In our work, we are covering this literature gap by studying how to personalize and customize algorithmic suggestions to different levels of professional medical experience.

2.3 Clinical Decision Support Systems

In medical applications, DL systems have also been the major contributor to the success of several CDSS applications [37]. Such CDSS applications can detect and learn patterns or make predictions to assist clinicians, such as pathologists, or radiologists, among others, in high-stakes clinical decision-making [110]. For instance, on the diagnosis of skin cancer [36], the segmentation of cardiac MRI [73], or breast cancer detection [69], there are a variety of works where DL systems were introduced for clinical purposes. Their outstanding performance in identifying meaningful patterns within the available data, recently used to help humans learn new biomarkers of specific diseases [119]. Because of that, several works are arguing that these models can see beyond what a trained radiologist sees in medical images [70, 71, 85]. Although some works are already contemplating the idea of a CDSS that predicts and explains some AI outcomes [26, 69], they ignore the need of adapting the communication tone for a personalized and customized medicine.

Most of the best performing CDSS rely on ML algorithms that learn specific tasks from training data [24, 101, 125]. The field recently gained enormous interest, primarily due to the practical successes of DL [72]. The rapid and widespread development of DL methods supports a wide range of image analysis tasks in breast cancer diagnosis, including classification, detection, and segmentation [59, 115]. These methods rely on large annotated datasets to learn essential and discriminative image features for each specific task, with performances matching and even surpassing humans [36]. However, past works are highlighting several obstacles in going from research and development environments to hospital or real clinical settings for these set of applications [12, 13].

The lack of utility to clinicians and logistical hurdles that slow or block deployment are frequent obstacles in real clinical settings [34, 75]. Even systems with widespread adoption, such as systems to aid radiologists during breast cancer diagnosis, are requiring them to do more work [56]. Instead of reducing the radiologist workload or ease the clinical workflow, these systems are generally not improving the radiologist's diagnostic accuracy [29, 56].

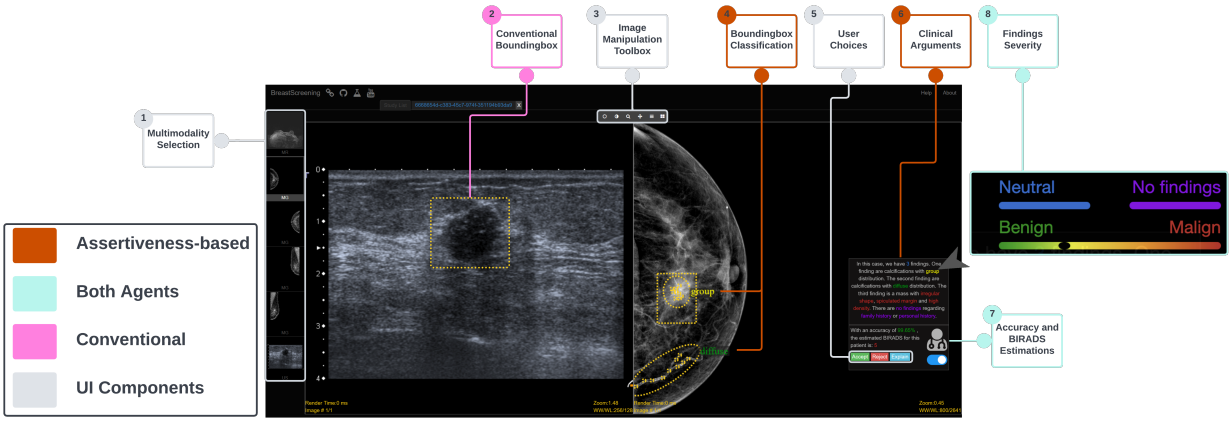

Figure 2: Interface for conventional and assertiveness-based AI agents for medical imaging analysis. Attributes are associated with numbers in each testing condition. When a clinician hovers the mouse over each variable (*e.g.*, accuracy, BIRADS, or any other clinical argument of attribute 6), the AI agent will pop up a window to inform the severity of each finding. The colors are ranging from benign (green) to malign (red). The number of findings is a (blue) neutral color, conclusions taken from the interviews, as one highly severe finding can be more critical than having several findings. During the interviews, clinicians choose the purple color for the family and personal history variables of the patient.

The work of Beede et al. [12] studies the introduction of CDSS set of tools, with the primary focus on AI systems that use clinical data. However, they are not examining the use of real patient AI predictions. Our work covers this gap by, not only incorporating real granular patient information from the AI outputs, but also studying how to personalize and customize a CDSS for the clinical setting.

Across the HCI literature [20, 40, 82, 100], or more precisely, from the CHI community [12, 19], human-centered evaluation of interactive, DL systems, is an open area of research within clinical environments. Cai et al. [19] created interactive techniques, leading to an increased diagnostic utility and user trust in the predictions from a DL system, used by clinicians in a lab setting. Correspondingly, Cai et al. [20] examined in the lab setting what information clinicians found to be important when being introduced to AI assistants, before integrating these assistants into routine prostate cancer screening practice. While these works bring us closer to understand the clinician's needs as they interact with DL-based systems, they do not account for the heterogeneous behavioral nature of decision-making.

2.4 Effects of AI Communication

Trust is critical in communication, especially, in clinical environments, where clinicians are exposed to critical scenarios that affect life decisions [7]. From the HCI literature [15, 61, 124], we know that the development of trust is influenced by the positive motivational attribution between the communication entity and the user. The work of Hohenstein et al. [47] is showing that a successful collaboration between human and AI occurs when ambiguity and uncertainty in terms of perceptions are reduced through trust [47]. While communicating AI predictions and explanations is shaping the design of recent works [31, 65], we do not know how assertiveness-based AI mediation is affecting novice or expert clinicians.

To avoid unexpected clinical consequences, we need to understand the effects of AI communication on human interactions. In fact, the direct effects of communication are suggesting that clinicians' level of trust in an AI system directly affects their perception of the outcomes [47]. Panigutti et al. [80] are arguing that higher levels of trust will cause the clinician to have a positive attitude, resulting in high satisfaction and positive perceptions of performance with respect to the interaction outcome. Moderation via adapting the communication suggests that trust will influence how a clinician interprets and evaluates information relevant to attitude and behavior. Attribution theory tell us that when behavior is consistent with explanations, humans will attribute causality to self characteristics and needs [62]. On the other hand, when behavior is inconsistent with prior expectations, where there is missing information or ambiguity, external cues will determine behavior [47]. As an example, a novice clinician asking for help and receiving a suggestive, *i.e.*, non-assertive communication. When in a real human-human interaction, would receive an assertive recommendation from an expert advisor. The novelty of our work is in the application of assertive communication theories in a deep learning system and clinical scenario that considers its use.

3 ASSERTIVENESS-BASED SYSTEM

In this study, we explore how human-AI interactions are affected by the ability of an AI agent to not only incorporate granular patient information from the AI outputs but also exploring how to adapt the communication tone (*i.e.*, more assertive or suggestive) depending on the medical experience (*i.e.*, novice or expert) of the clinician. Specifically, we compare the AI outputs (Figure 2) that explain to clinicians some clinical arguments with more granular information about the patient regarding the lesion details, to a conventional agent that only provides numeric estimates (*e.g.*, BIRADS and accuracy) of the classification.

Our assertiveness-based agent uses recommendations for classifying and segmenting: (1) the number of detected findings; (2) the patient severity of each breast and per medical imaging modality; (3) a visual scale representing the benign or malign estimates; (4) providing visualization of the sensitivity and specificity outcomes of the models; and (5) with clinical arguments of the patient, such as pathological co-variables. To compare the assertiveness-based agent to the conventional agent, we inform participants that the recommendations are generated by our AI models, so that they can also provide some feedback concerning the model performance.

Figure 2 illustrates how the two AI agents were integrated into an existing medical workflow for the classification of medical imaging data on the support of breast cancer diagnosis. Both agents are recommending classification and segmentation based on the DenseNet model [49] for MG and US, as well as based on the ResNet model [112] for MRI. The two AI agents provided the severity classification of the patient via BIRADS [105], the accuracy of the model for that classification, and the segmentation of the lesion to explain the regions that derived from that classification.

Clinicians can open the patient by selecting the ID in a list of patients. When the patient is open (attribute 1 of Figure 2), clinicians can select each respective image of the breast, by dragging-drop each image to the view-ports. From here, clinicians can manipulate the image through the toolbox (attribute 3 of Figure 2). In the end, the idea is to *accept* or *reject* the final recommendation of the AI agent (attribute 5 of Figure 2), while clinicians can also ask for an *explanation* to support their final decision-making. The key difference between the two AI agents was in how they show communication with clinicians. A typical output from an AI model includes not only the predicted classification of the BIRADS but also a likelihood distribution over all possible classification choices.

In our study, both AI agents were designed to communicate this type of *quantitative* confidence in two ways (Figure 3): (1) for the conventional condition, clinicians could simply see the suggested numeric representation of the BIRADS and the respective accuracy of the model; (2) for the assertiveness-based condition, the agent was communicating the clinical arguments along with the communication of the BIRADS and accuracy, but this time by descriptive information. While our conventional agent employed this baseline numeric representation of confidence, our assertiveness-based agent is communicating the *quantitative* confidence based on a descriptive sentence of the clinical arguments. Specifically, the image view-port was augmented with an additional bounding box or circle ellipse **highlighting the lesion characteristics** that were likely to explain the final BIRADS classification. Note that these suggestions did not dictate the order in which imaging modalities are presented to clinicians. Indeed, clinicians can remain to decide freely what modalities and clinical arguments are reviewed first.

The patient's detailed augmentation for medical imaging on breast cancer diagnosis was extended with an **assertiveness-based explanation**. With our system, we are listing human-interpretable clinical arguments for classification and segmentation recommendations that would adapt their communication depending on the personalized and customized demographic characteristics of the clinician. These clinical arguments correspond to the classification outputs of an AI model and were trained on data (Section 5.2) from real-world clinical cases, as described next.

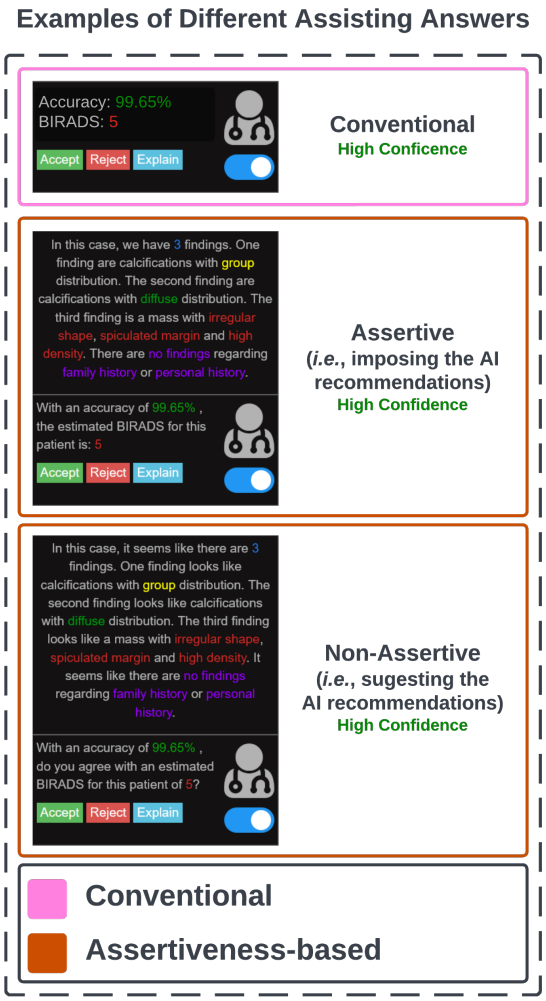

Figure 3: Example of representative use cases for the different testing trials. From top to the bottom, the first agent is representing a conventional example (pink), while the other two are representing assertiveness-based examples (brown), from assertive to non-assertive communication.

4 RESEARCH QUESTIONS & HYPOTHESES

The final purpose of our research is twofold. Through assertiveness-based agents, we first aim to understand how personalized and customized communication could affect medical assessments in terms of the efficiency and efficacy of the clinical workflow. Secondly, we aim to understand how clinicians perceive assertiveness-based agents differently. Thus, our work addresses two primary research questions concerning the impact of assertiveness-based agents on efficiency and efficacy (RQ1), as well as the perception (RQ2) of clinicians.

Specifically, we consider the following research questions and related hypotheses:

- **RQ1.** How does an assertiveness-based agent affect medical assessments?
  - **H1.1.** Efficiency of clinicians in terms of time performance per each diagnosed patient will be higher with an assertiveness-based agent.
  - **H1.2.** Classification accuracy of clinicians will not suffer with an assertiveness-based agent.
  - **H1.3.** Through assertiveness-based communication, accuracy differences between novice and expert clinicians will depend on the tone of the personalized explanations.
- **RQ2.** How is an assertiveness-based agent perceived by clinicians?
  - **H2.1.** Clinicians will have a preference for an assertiveness-based agent.
  - **H2.2.** Clinicians will consider an assertiveness-based agent more trustworthy.
  - **H2.3.** Personalized highlights and explanations will not increase clinicians' workload nor decrease usability.
  - **H2.4.** Novice and expert clinicians will perceive reliability and capability differently, depending on the levels of assertiveness.

In real-world clinical settings, time of clinicians is a limited and expensive resource that should be reallocated efficiently. We take the position that while clinicians should make their clinical assessments with care, AI agents can help on the diagnosis efficiency. Our assertiveness-based agent is designed to summarize the recommendations, and to provide clinical arguments explaining the underlying classification and segmentation of the AI models.

For RQ1, our protection is that personalized and customized explanations can inform clinical decision-making. Therefore, it will increase clinicians' classification accuracy without reducing their time performance over diagnosis. Especially, we envision that adapting the communication of the personalized explanations is crucial for successfully informing decision-making of clinicians.

The HCI research community has settled that poor user perception can be a barrier to adoption of technology regardless of performance [77, 93, 94]. In fact, the user perceptions of preference and trust are essential in predicting technology adoption [94]. Hence, it is important to investigate clinicians' perception, as we do in RQ2. Beyond the primary outcome in terms of reliability in AI-assisted clinical assessments, our goal is also to understand how assertiveness-based agents are perceived by clinicians.

5 METHODS

As a way to improve clinical decision-making, the goal of our study is to attain a deeper understanding of personalized and customized mechanisms, exploring how to adapt the communication of agents depending on the medical experience of clinicians. Our study draws from 52 semi-structured interviews and user testing with clinicians for breast cancer detection via medical imaging diagnosis. To accomplish this goal, we conducted a two-condition study, within-subjects, counterbalanced experiment, in which each subject participated in three trials, providing us with rich information to analyze.

Interviews were conducted from March 2022 to June 2022. Research questions were fused from these interviews and observations, thanks to the support of user-centered activities, such as workshops, focus groups, affinity diagramming, data clustering, and prototype co-design, leading us to reported problems. Participants in these sessions include clinicians from different healthcare institutions, researchers from the ML field, and HCI researchers. Before the work commenced, the study was approved by the ethics committee of each clinical institution. Next, we describe the details of our controlled experiment including the task, dataset, information of participants, study procedures, and statistical analysis.

5.1 Task

We conducted our study in the field of imaging classification on breast cancer, a clinical domain with typically high False-Positive rates to over-diagnose a patient [54, 78]. In particular, we compared our conventional and assertiveness-based agents in the context of assisting trained medical personnel in the task of a breast cancer diagnosis. For the conventional condition, we used the *BreastScreening-AI* framework [26] publicly available (git.io/JMjDi) and built for the development of medical assistants. Aside from the already available functionalities of this framework, we developed two conditions for testing our hypothesis from a post-hoc analyses concerning how to personalize and customize the AI outcomes to clinicians. We raised several trials, from a more suggestive (non-assertive) to a more assertive tone of the AI recommendations, where clinicians with different levels of expertise are interacting with both trials and the conventional. In the end, all clinicians interacted with one conventional, one non-assertive, and one assertive agent. Each clinician diagnosed three patients with different severities, where the task was to *accept* or *reject* the AI recommendations.

There are two groups of clinicians with different medical professional experiences: (a) novice; and (b) expert. Patients are divided into three groups of breast severities: (i) low severity, representing the BIRADS of 1, meaning there are no findings for that patient and both breasts are healthy; (ii) medium severity, representing the BIRADS of 2 and 3, meaning there are some findings with higher probability of benign suspicious; and (iii) high severity, representing the BIRADS of 4 and 5, meaning there are findings with higher probability of malign suspicious. Usually, each patient has available three types of modalities (*i.e.*, MG, US, and MRI). For this task, clinicians need to read six imaging views approximately per each patient: (1) one CC-L; (2) one CC-R; (3) one MLO-L; (4) one MLO-R; (5) one US; and (6) one DCE-MRI volume with between 100 and 200 frames. Shortly, clinicians participated as readers while assessing each patient in terms of the likelihood and location of the malignancy.

During the task of responding to the lesion localization, the reader clinician provides the severity classification of the clinical argument for that suspicious attribute on the image. For each image, the clinician classifies the patient final severity assessment via BIRADS classification by default. Meaning that, although the patient has some clinical arguments (*e.g.*, group microcalcifications with diffuse distribution in Figure 3) pointing to a BIRADS of 2, if there is a suspicious irregular shape mass with spiculated margin, by default the clinician will consider the final BIRADS as a 5.

The task of diagnosing breast cancer involves reading heterogeneous appearances, ranging from obvious masses with spiculated margins to subtle asymmetric or faint microcalcifications [108]. This leads to difficulties for clinicians in achieving an accurate diagnosis and consistent interpretation of the patient. Because of that, clinicians apply rules from radiological guidelines to classify breast images based on visually inspecting the patterns on the image [28].

The classification of breast cancer via the BIRADS scale lends itself to a task for our study on AI agents in medical imaging assessments. Not only the task is a time-consuming and tedious procedure for clinicians, but also relies on non-trivial classification tasks. Indeed, the prior medical literature has established that the medical error of clinicians has between 50% and 30% chance of being a false-positive and about 10% chance of being a false-negative [103].

5.2 Dataset

In this paper, we used a total of 338 cases acquired in the first hospital (Section 5.3). From this set of cases (Section A.2), 289 were classified by the head of radiology. Each patient has several images concerning four X-ray mammography (two in CC and two MLO views), one ultrasound image to train the DenseNet model, and roughly DCE-MRI images in MRI to train the ResNet model. In the MRI volumes, we take several image slices per patient, where the lesion is present. This provides us roughly 2890 images (*i.e.*, 289 × (4 + 1 + 5)), that are used to train/test the AI models.

Traditional image processing and deep learning techniques require extensive pre-processing [128]. As a matter of fact, it is known that a cleaning dataset is welcome when training a deep neural network [87]. In our study, this stage is of the utmost importance, since the MG, US, and MR images contain quite different intensities and the images are of different sizes. Thus, before introducing the images to the DenseNet and ResNet models, we pre-process the data (Section A.6). Specifically, we perform data normalization, so that the images have the same intensity, regardless of the modality.

All the images are resized to 224x224 pixels. The images are normalized by subtracting their mean and dividing by their standard deviation. The above size is used since the DenseNet model is prepared to receive the input with this format.

5.3 Participants

We recruited 52 clinicians as participants for our study on a volunteer basis from a broad range of clinical environments, including different health institutions (public hospitals, cancer institutes, and private clinics). Our clinicians were recruited through the already established protocols under this study from 11 different clinical institutions: (1) Hospital Prof. Dr. Fernando Fonseca (HFF); (2) Instituto Português de Oncologia (IPO) de Lisboa (IPO-Lisboa); (3) Hospital de Santa Maria (HSM); (4) IPO de Coimbra (IPO-Coimbra); (5) Madeira Medical Center (MMC); (6) Serviços de Assistência Médico-Social do Sindicato dos Bancários do Sul e Ilhas (SAMS); (7) Hospital do Barreiro (HB); (8) Hospital de Santo António (HSA); (9) Champalimaud Foundation (CF); (10) Centro Hospitalar De Trás-Os-Montes E Alto Douro, E.P.E. (CHTMAD); and (11) Centro Hospitalar de Lisboa Ocidental. EPE (CHLO). All clinicians and clinical institutions gave prior permission to use their data for research purposes under this study.

From the demographic questionnaires (Section A.3), 55.77% of participants are expert clinicians, whereas 34.62% are seniors having more than 10 years of practical experience, and 21.15% are middle clinicians having more than 5 years but less than 10 years. Similarly, 44.23% of participants are novice clinicians, whereas 32.69% are juniors after taking the exam, having up to 5 years of clinical experience, and 11.54% are interns before the medical specialty exam. Each clinician was exposed to the three trials (*i.e.*, conventional, assertive, and non-assertive) in a counter-balanced manner.

5.4 Procedure

After providing an informed consent form for participation in the study, each clinician reported information concerning several self-characteristics. First, they reported their demographic characteristics (Section A.3). Second, they reported their professional backgrounds, such as clinical education (*i.e.*, radiology, surgery, nurse, technician, etc), areas of expertise, work sector, and medical experience. Finally, information about their experience while reading medical imaging data. Next, clinicians familiarized themselves for about 3 minutes with our user interface and with the basic functionalities common to both AI agents.

At this stage, each participant interacted with the assistant, *accepting* or *rejecting* the system suggestion in the two different conditions: (a) conventional; and (b) assertiveness-based. The set of patients was providing participants 289 patients, while all patients must have at least one of the three available modalities. Each participant open the set of three patients (*e.g.*, **P1**, **P2** or **P3**), chosen randomly, and examined it. During the examination, the participant interacts with the available functionalities of the system.

Clinicians performed the same task of diagnosing three patients twice, once with the conventional agent and another time with the assertiveness-based agent. For each task, clinicians were asked to read the suggested AI recommendations, where the task ends when clinicians *accept* or *reject* the proposed BIRADS. Additionally, clinicians could ask for a visual *explanation* inside the image during the task. The list of patients was fully classified by the AI models, and clinicians could revise the explanations (bounding boxes 2 and 4 of Figure 2) for the important regions to consider.

After each task, clinicians filled out a brief feedback questionnaire exploring their perception of each AI agent. The questionnaire included scales to measure three dimensions of trust [63] represented by perceived understanding, competence, and thoughtfulness, as well as cognitive workload by using NASA-TLX [92], and usability by using SUS [60]. With these measures, our purpose was to understand perceived diagnostic utility and decision-making support provided by the AI agents, and whether clinicians thought they would use the agents in practice. Upon completing all the tasks, we measure the preferences while using both conventional or assertiveness-based agents, and the different levels of assertiveness.

Clinicians compared both AI agents with respect to reliability, capability, and overall preference. To measure the levels of assertiveness rated between novice and expert clinicians, we measured the reliability and capability of the different (*i.e.*, from more suggestive to more imposing AI recommendations) communication tones of the assertiveness-based agent. Clinicians rated these items on a 7-point Likert scale.

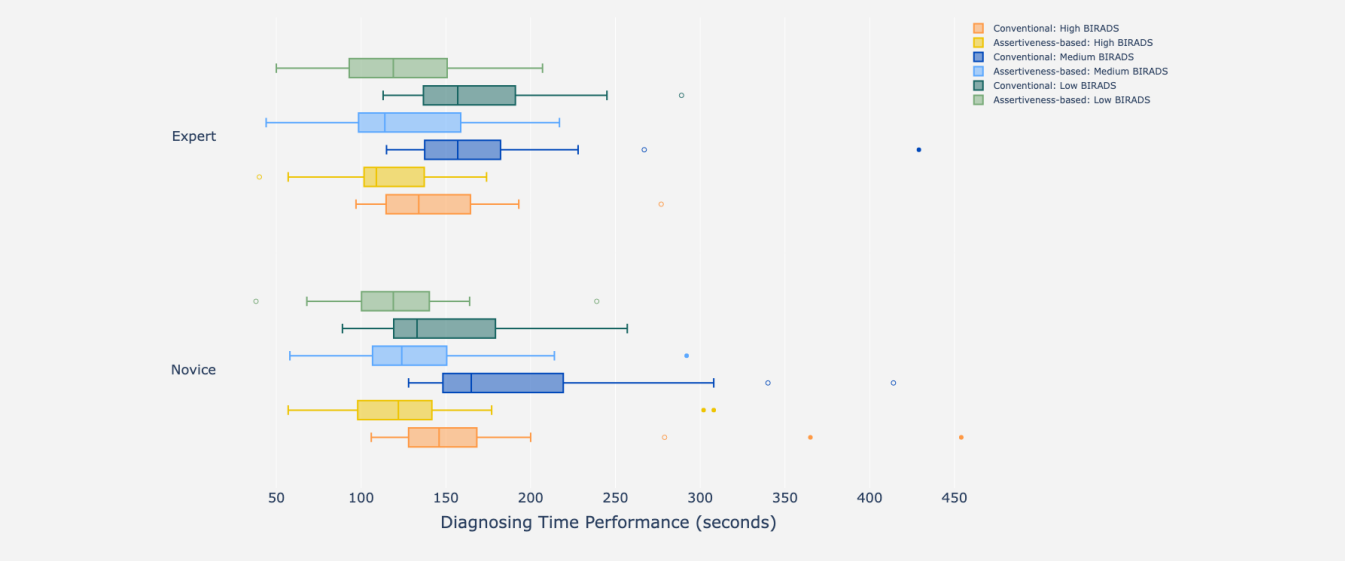

Figure 4: Diagnosing time performance in seconds of novice and expert clinicians to fully diagnose one patient. Different colors are representing different agent trials and breast severities of a patient. Clinicians' task was to read each patient and provide a final BIRADS classification by *accepting* or *rejecting* the AI recommendations.

5.5 Analysis

For **RQ1**, we investigated the impact of our assertiveness-based agent on clinicians' efficiency and efficacy in terms of time performance and accuracy in diagnosing patients with the support of the AI-suggested recommendations. Similar to the literature, we used the one-way ANOVA test [91, 127] to compare both AI agents with respect to the following outcome measures per clinician: (i) the time (in seconds) for diagnosing each patient (**H1.1.**); and (ii) accuracy rates via false-positives and false-negatives of clinician-provided classifications (**H1.2.**). For the efficacy differences between novice and expert clinicians during decision-making (**H1.3.**), we used the chi-squared test of independence [86] to assess the relationship between the expertise of clinicians and the assertiveness levels of the agents. Regarding human-AI accuracy, our dataset has post-biopsy verification, meaning that we could measure the real ground-truth of the patient.

For **RQ2**, we compared clinicians' perceptions of both conventional and assertiveness-based agents. A possible observed pattern in perceived preference (**H2.1.**) and trustworthiness (**H2.2.**) was examined using the ANOVA test and statistical significance ($p < 0.05$) for testing our hypothesis. Reported scores for cognitive workload and usability (**H2.3.**) were compared between the two AI agents using statistical significance ($p < 0.05$) for computing the likelihood of confidence. Last, we used the one-way ANOVA test of variance to test the levels of assertiveness for the provided clinical arguments between the two groups (*i.e.*, novice and expert clinicians) of medical professional experience. Specifically, we used this test to measure the perceived preferences of clinicians in terms of reliability and capability (**H2.4.**). From "Totally Non-Assertive" level, *i.e.*, more suggestive, to "Totally Assertive" level, *i.e.*, more imposing AI recommendations, we test the overall tendency between novice and expert clinicians of the communication tone.

Finally, we used the open coding comments and feedback from focus groups, workshops, and interviews. The purpose was to extract emerging themes from open-ended discussions during these sessions [14, 97]. We organized the responses of clinicians using affinity diagrams to cluster workflow clinical practices and main functional ideas of the agents in greater detail [30, 42]. Moreover, we used affinity diagramming to uncover clinicians' preferences and concerns based on the data gathered in a thematic (*e.g.*, card sorting) coding method. This information was then used to inform our conclusions about exploring how to personalize and customize the AI recommendations by adapting the communication tone.

Clinicians were asked to reflect on how they used to make their decisions, what information they need to be explained by the AI models, and why they need that. These qualitative analysis methodologies enable the identification of emerging themes in the data for revealing design considerations. As follows (Section 6.3), recurring themes are reported below as we detail them with provided feedback and comments from these sessions with clinicians.

6 RESULTS

To test our hypotheses, we used the `scipy` library from `python` to conduct a one-way ANOVA test with the level of medical professional experience as the main factor on the dependent variables [27]. The alpha level ($\alpha = 0.05$) was set for statistics, and the effect size was used to quantitatively measure the magnitude of the experimental comparison effect between variables [68, 123]. Briefly, we focus on statistically significant results and selectively report the results to address our hypotheses by following literature recommendations [18, 19, 127]. Next, we investigate the time performance (Figure 4), accuracy (Figure 5) and decision (Table 1) rates of clinicians, while addressing their preference choices (Figure 6), agreement comparisons (Table 2), reliability and capability (Figure 7).

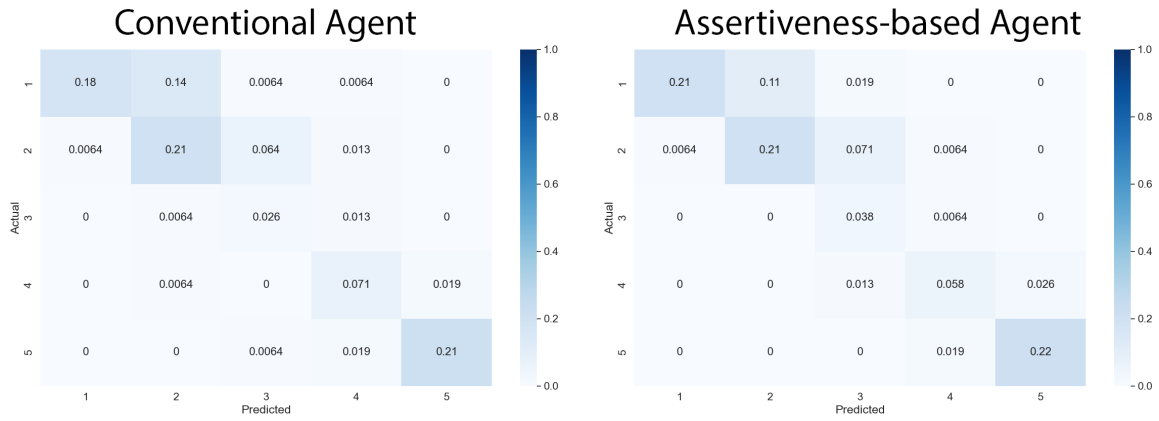

Figure 5: Accuracy rates using a confusion matrix. Comparison between the clinician's BIRADS classification (from 1 to 5) of a patient while using both conventional (left) and assertiveness-based (right) agents. Columns are representing the *Predicted* value (collaboration between the clinician and AI), and the rows are representing the *Actual* category (biopsy confirmed).

6.1 RQ1: How does an assertiveness-based intelligent agent affect medical assessments?

We hypothesized that using the assertiveness-based communication between clinicians and an intelligent agent, would alter clinicians' workflow and increase the time performance of clinicians (**H1.1.**) during patient diagnosis. On average, time performance of clinicians was significantly improved with the assertiveness-based agent (M = 124.02 seconds, SD = 44.60 seconds) than with conventional agent (M = 166.12 seconds, SD = 60.42 seconds), confirming our hypothesis (Figure 4). This difference was significant (F = 11.32, p = 0.005 < 0.05), indicating a large effect size (r = 0.49).

We also hypothesized (**H1.2.**) that using the assertiveness-based agent would not negatively affect accuracy of clinicians. Our results show (Figure 5) that there was no significant difference (F = 1.85, p = 0.37 > 0.05) in accuracy. Such results are providing support for our hypothesis (**H1.2.**) that clinicians' accuracy at classifying a patient was not negatively affected by being exposed to assertiveness-based explanations. This suggests that the assertiveness-based agent could be used in clinical settings without having a negative impact on the accuracy of patient diagnosis. To support that claim, Table 1 shows how often clinicians accept or reject the AI recommendations.

We further examined the potential impact of personalized explanations by customizing the agent communication differently between the two groups of professional medical experience, *i.e.*, novice and expert clinicians (**H1.3.**). We observed a significant association between the levels of assertiveness (*e.g.*, from non-assertive to assertive communication tone of the clinical arguments) and the medical professional experience while revising AI recommendations (χ^2 = 3.84, p = 0.001 < 0.05). In other words, the chance of a patient getting classified correctly by a novice was significantly higher ($\text{Accuracy}_{\text{novice}}$ = 91%) with the assertive agent (*i.e.*, imposing AI recommendations) than with the non-assertive (*i.e.*, more suggestive AI recommendations). On the contrary, the chance of correctly classifying the patient by an expert clinician was slightly higher ($\text{Accuracy}_{\text{expert}}$ = 78%) with the non-assertive agent. The odds of a patient getting classified correctly by a novice had 17.4% chance higher with the assertive agent, while expert clinicians had 4.4% chance higher with the non-assertive agent. These findings suggest that the level of assertiveness of the agent's communication may need to be tailored to the experience level of the clinician. Exploring the communication tone indicate that agents may need to be more assertive for novice clinicians, while suggestive tone may be more appropriate for expert clinicians.

Trials	Correct Accepts		Correct Rejects		Overall Corrects		Wrong Accepts		Wrong Rejects		Overall Mistakes	
	Novice	Expert	Novice	Expert	Novice	Expert	Novice	Expert	Novice	Expert	Novice	Expert
Conventional	68%	62%	1.7%	1.77%	69.70%	63.77%	3.7%	4.1%	26.6%	32.13%	30.30%	36.23%
Assertive	79.83%	63.13%	1.76%	2.63%	81.59%	65.76%	3.05%	4.75%	15.36%	29.49%	18.41%	34.25%
Non-Assertive	72.64%	64.72%	2.99%	1.69%	75.63%	66.41%	5.48%	3.91%	18.89%	29.68%	24.37%	33.59%

Table 1: Frequency of clinicians, showing how often they accept or reject the AI recommendations. These rates show how often clinicians switched to a different conclusion after interacting with each agent. The "Overall Corrects" denote the frequency of time clinicians correctly accept ("Correct Accepts") the recommendations of the AI agents and correctly reject ("Correct Rejects") by changing the wrong AI recommendation to the right diagnostic. On the other hand, "Overall Mistakes" is denoting the frequency of times clinicians wrongly accept ("Wrong Accepts") the AI recommendations, meaning that the AI agent was wrong, but they accept it, and wrongly reject ("Wrong Rejects"), meaning the AI agent was right, but clinicians changed to the wrong final diagnostic.

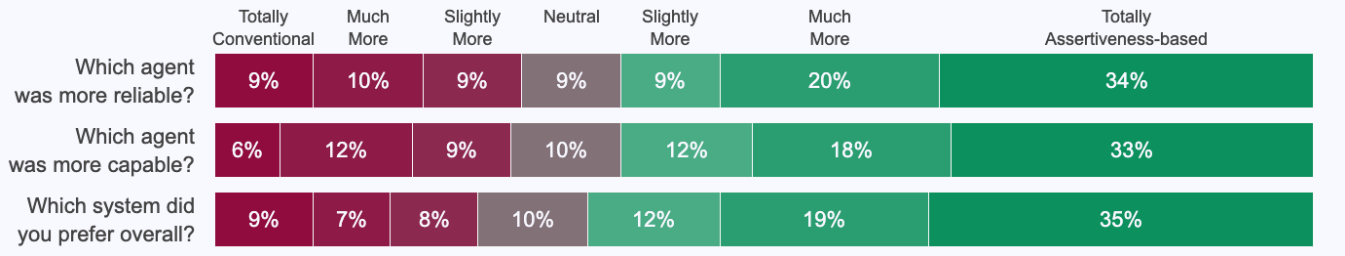

Figure 6: Preference choices of clinicians when comparing between both conventional and assertiveness-based agents within this study. Rates of clinicians are ranging from *Totally Conventional* **to** *Totally Assertiveness-based* **on perceived** *reliability*, *capability*, **and** *overall preference* **of each agent.**

Finally, Table 1 shows how often clinicians accept or reject the AI recommendations, as well as how often they are switching to a different conclusion. The highest overall correct rate was 81.59% in the assertive trial for novice clinicians and 66.41% in the non-assertive trial for expert clinicians. Moreover, experts are switching to less wrong decisions with the non-assertive agent (Total = 33.59%). On the other hand, novice are switching to less wrong decisions with the assertive agent (Total = 18.41%). Overall, clinicians did better decisions with assertiveness-based assistance while exploring how to adapt the communication tone. The results suggest that the assertiveness-based condition may have been more favorable to both novice and expert clinicians, with higher correct acceptance and correct reject rates compared to the conventional condition. The results highlight the importance of scenario design in evaluating the performance of clinicians, as the trials had a significant impact on the performance of both novice and expert clinicians.

6.2 RQ2: How is an assertiveness-based agent perceived by clinicians?

For RQ2, we explored clinicians' perception of both AI agents. Results for our hypothesis (**H2.1**) that clinicians would have a preference (Figure 6) for an assertiveness-based agent were statistically significant ($F = 8.35$, $p = 0.001 < 0.05$) between the groups of interns, juniors, middles, and seniors with a large effect size ($r = 0.41$). Out of the 52 participants who expressed a preference, 66% preferred the assertiveness-based agent and another 24% preferred the conventional agent, while 10% did not have a preference.

Additional to the significant differences of preference between both conventional and assertiveness-based agents, we also analyzed perceived trust of each agent (Table 2). Overall, there were only slightly differences ($F = 19.47$, $p = 0.06 > 0.05$) between conventional and assertiveness-based agents. Besides, there were no significant differences of understanding ($p = 0.14 > 0.05$) between both agents. While no significant differences in understanding could be detected between both agents, the assertiveness-based variant was considered to have greater competence ($p = 0.04 < 0.05$). Moreover, we observed a trend that clinicians had higher thoughtfulness in the assertiveness-based than in the conventional agent ($p = 0.001 < 0.05$). These results are providing partial support for our **H2.2.** hypothesis.

We also evaluated if there were no significant differences of workload and usability between both conventional and assertiveness-based agents. Specifically, there were no significant differences between the workload scores of the two AI agents on NASA-TLX ($p = 0.38 > 0.05$). Furthermore, we observed no significant differences between the usability scores on SUS ($p = 0.38 > 0.05$). Hence, providing support for our **H2.3.** hypothesis for workload and usability.

Finally, to assess how does clinicians perceive differently the levels of assertiveness (Figure 7), we compared the preferences in terms of reliability and capability from non-assertive (suggestive) to assertive (authoritative) communication of the clinical arguments. Here, we can denote that there are significant differences for reliability ($F = 31.36$, $p = 0.0001 < 0.05$) and capability ($F = 18.17$, $p = 0.0003 < 0.05$) between groups of novice and expert clinicians. In fact, novice clinicians perceived the assertive communication as more reliable (61%), although not mainly feeling the same for capability (48%). On the other hand, expert clinicians perceived the non-assertive communication as more reliable (69%) and capable (66%). Therefore, we can observe that the **H2.4.** hypothesis is supported by showing that novice and expert clinicians will perceive differently the provided clinical arguments depending on if the agent is imposing the AI recommendations or being more suggestive.

6.3 Qualitative Insights

We adopted a participatory approach for qualitatively analyzing our study results. Given our collected data (focus group sessions, participant opinions, and transcripts), we used emergent affinity diagrams to identify common themes in how participants intend to have their clinical arguments and visualization of the AI recommendations. Our qualitative analysis of participant responses to open-ended survey questions yielded insights on how assertiveness-based agents can affect clinicians' workflows and their mental model.

6.3.1 Altering Clinical Workflows. To validate the proposed design, we discussed with clinicians how could they use these set of personalized agent communications to perform diagnosis on a real clinical environment. The goal is to understand whether an assertiveness-based agent can (a) be compatible integrated into clinicians' workflow, and (b) provide added values to clinicians' diagnosis process. Next, we summarize the main opinions of clinicians between conventional and assertiveness-based assistance, as well as between suggestive (non-assertive) and more imposing (assertive) AI recommendations.

Questions	Conventional	Assertiveness-based
Overrall, I can trust in the agent recommendations.	86%	90%
I understand what the system is thinking.	91%	94%
The system seems competent.	82%	92%
The agent shows great thoughtfulness while dealing with the patient.	71%	75%

Table 2: Comparison for the percentage of agreement between conventional and assertiveness-based agents. The questions are following the three *dimensions of trust* represented by perceived *understanding*; *competence*; and *thoughtfulness*.

One major criticism to the traditional approach of representing the AI recommendations with numeric BIRADS classification, accuracy of the output and heatmap values is that it is not sufficient for clinicians to make sense of the decision-making reasoning behind the output classifications. In particular, when the output accuracy is lower than 80% confidence. Our qualitative findings suggest that in choosing between numeric representations of the AI output classifications and human-interpretable arguments while exploring how to adapt the communication, clinicians found the latter to be more effective during decision-making. This highlights the necessity for AI systems to be designed with a user-centered approach [129, 130], taking into account the preferences and needs of the end-users (in this case, clinicians) to ensure that they are usable and effective in real-world scenarios.

Specifically, a middle clinician (*i.e.*, expert clinician) reported:

"*When I was interacting with the AI agents, the first thing I did was to find classification conflicts between the final BIRADS of the patient, the BIRADS of each image, output accuracy, and clinical arguments. Something that I couldn't do so well in the first [conventional] agent, but could do better for the second [assertiveness-based] one.*" (C51)

In domains where clinicians' availability is rare, it can be exceedingly hard to obtain an immediate second reader every time a clinician needs the opinion of another human expert. Clinicians shared a positive attitude towards the use of an assertiveness-based agent to aid their decision-making, since we are exploring how these set of agents are adapting the communication depending on the level of professional medical experience. Something that is common when a senior (assertively) talks to an intern, while our agent is mimicking the same communication conditions. Such mimicking behavior must be designed to help diagnosis without incurring in learning misinformation that interrupts the main clinical workflow.

Here, a senior clinician (*i.e.*, expert clinician) is sustaining the above argument by reporting that:

"*Adapting the communication between more suggestive [non-assertive] or assertive tone of the clinical arguments can help the diagnosis workflow. I may prefer a suggestive agent, but an assertive agent will be more helpful for my interns concerning an educational purpose.*" (C15)

Categorically, clinicians were mostly stating that they could understand how such personalized communication could be beneficial to customize the interaction between humans and AI. In particular, clinicians valued the opportunity to choose the communication tone to explore the clinical arguments in a more detailed fashion. Our qualitative findings suggest that most clinicians (48/52) found the assertiveness-based agent to be more helpful and reliable. These findings also support our claim that customization could have a positive impact on the decision-making process and improve the overall effectiveness of AI-assisted systems.

As another example, a junior clinician (*i.e.*, novice clinician) reported that:

"*When the agent is talking to me in a more assertive way, I can feel more safe of my decision... and feeling more assurance of the right answer.*" (C17)

This analysis further highlights the effectiveness of the assertiveness-based agent in personalizing and customizing the communication of the agent, taking into account differences of medical professional experience. That is, 37 out of 52 clinicians in our study explicitly mentioned that their workflow differed between the two agents. Particularly, clinicians showed different confidence and trusting opinions depending on the levels of assertiveness.

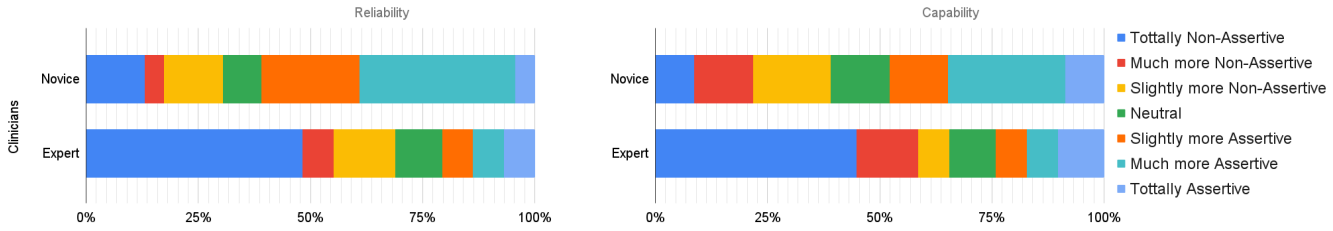

Figure 7: Ratings between novice and expert clinicians for perceived reliability and capability. Clinicians rated each agent, ranging from *Totally Non-Assertive* to *Totally Assertive* communication.

For instance, an intern clinician (*i.e.*, novice clinician) reported that:

"*It seems like, when I was interacting with the more suggestive [non-assertive] assistant, it has the same doubts on the communication tone as I am. The more assertive assistant gave me higher confidence in the decision.*" (C10)

On the other hand, a senior clinician (*i.e.*, expert clinician) reported the following:

"*In my opinion, I don't like the communication tone and the way assertive agents are reporting the clinical arguments. Imposing the AI recommendations feels like I need to follow the orders they give to me. I prefer a more suggestive agent, asking me if the clinical arguments are well classified or not.*" (C49)

To conclude, these levels of personalizing and customizing the agent communication (*e.g.*, from less assertive to more assertive) are important to take into account when designing systems for critical domains. Especially, for decision-making under these clinical workflows. We found that assertiveness-based not only enhanced decision-making, but also helped clinicians to develop a mental model of the AI agents, or probe for the likelihood of the diagnosis. Next, we describe in what manner our work is leveraging these insights.

6.3.2 Agent Mental Models. Both novice and expert clinicians have preconceived mental models about the levels of assertiveness in different ways. As an example, expert clinicians used the output results to disambiguate AI errors from their own errors, depending on the communication tone. It is, therefore, possible that this reasoning behavior is projected onto the AI agent to anticipate where the agent would likely make mistakes: "*It could be also important to adapt the communication tone of the clinical arguments depending on the AI confidence.*" From here, we could understand that expert clinicians expect that the assertiveness of a clinical argument can also be adapted depending on the accuracy of the AI output results for that particular variable. On the contrary, novice clinicians were more focused on the learning process and patient comparisons for educational purpose: "*For me, the most important thing was to look at the provided arguments and understand if they are right from what I learn or from similar cases. A junior like me must have an idea how the machine is thinking to mentally follow the same reasoning process from my side.*" Hence, it is important to provide a more '*storytelling-like*' view of the patient for novice clinicians, even though in a more assertive fashion.

Apart from preconceptions, we further observed that clinicians developed comparative mental models between conventional and assertiveness-based agents: "*The first AI [assertiveness-based] was outstanding... but in the second AI [conventional] I was frustrated with the lack of communication in comparison to the first one.*" Moreover, the interaction experience of clinicians with the different AI agents can also shape their reasoning while looking for AI recommendation mistakes: "*In the second assistant [assertiveness-based], I look for classification conflicts between the final BIRADS and the clinical arguments, while I couldn't do the same for the first assistant [conventional], taking me more time to see if there are some mistakes.*"

Our intelligent agents can leverage these insights by not only provide a personalized and customized communication with different perspectives between novice and expert clinicians, but also correction behavior while adjusting internal representations of specific levels of assertiveness. Yet, without hurting time performance of the diagnostic (Section 6.1), nor increasing the workload (Section 6.2). In sum, these observations support growing evidence that taking into account the communication of the AI outputs (*e.g.*, structure, order, and tone of the arguments) can alter the clinicians' perceptions of the mental models of assisting agents.

7 DISCUSSION

In this work, we studied how personalized and customized communication of an intelligent agent can aid clinicians in their decision-making during medical imaging diagnosis. We conducted a within-subject to investigate how clinicians perceive assertiveness-based agents differently. Our results are showing that assertiveness-based agents can alter clinicians' workflows by increasing the efficiency of clinicians while maintaining overall efficacy.

Overall, the classification accuracy was not affected by providing assertiveness-based communication. We observed a significant effect of differences between novice and expert clinicians, depending on the explanations tone. This promising insight motivates future directions in the development and validation of compliant agents, capable of providing relevant and customized explanations.

Participants were keener on following AI recommendations that adapt their communication tone than the one that did not. Although this effect may occur because of adding more explanations, it reflected in differences of behavioral decisions between novice and expert clinicians. We gain even more insight on the effect of tone from the feedback provided by the clinicians across our qualitative analysis. Our qualitative results are showing that participants appreciate the idea of adapting tone to probe the likelihood of the diagnosis. This finding might be in line with the previous research in psychology and decision support science [16, 62, 95], bringing new directions for the theoretical application of assertiveness-based communication in deep learning systems and clinical domains.

We studied how personalizing and customizing the AI recommendations according to the professional experience of each clinician can reduce medical errors and increase satisfaction. Specifically, we found significant differences in terms of accuracy, perceived reliability and capability between novice and expert clinicians, depending on the tone of the personalized explanations. This finding is relevant in the design of AI agents for the healthcare sector, while providing a contribution to the HCI field.

Clinicians' overall preferences and perceived trust were also increased for the assertiveness-based agent in comparison with the conventional. Results suggest a higher perceived understanding of the assertiveness-based agent compared to the conventional variant. On the same hand, the assertiveness-based agent showed to be perceived by clinicians as more competent and thoughtful. However, our results are indicating the existence of other latent variables. For instance, the demographic characteristics of clinicians with different levels of clinical experience could shape the implementation of intelligent agents to take into account the differences in clinicians' perception of AI systems generally.

In terms of results significance for the HCI community, our research focuses on the use of intelligent agents in medical imaging, which is an important and growing area within HCI. By developing personalized and customized explanations, we aim to improve the effectiveness of decision-making by human clinicians, which is a key concern in HCI. Our specific contribution to the CHI community is the design of a novel interactive approach for personalized and customized explanations on intelligent agents, underpinned by computational principles. Our approach combines machine learning with image processing techniques to generate explanations that are tailored to the individual clinician's expertise. To our knowledge, this is the first time that this approach has been proposed and evaluated in the context of medical imaging.

Concerning the broader implications of our work, we believe that it has relevance for both decision support research and AI communication research. Our approach to personalized and customized explanations has the potential to improve the accuracy and effectiveness of decision-making by humans in critical domains, which is an essential goal in decision-support research. At the same time, our work also contributes to the growing body of research on how to improve the communication between AI systems and human users, which is a key concern in AI communication research. Next, we are discussing the design implications and generalizability of our findings, by concluding the limitations of our study and directions for future work.

7.1 Design Implications

Our findings have different stages of design implications for the development of novel AI-assisted systems in this clinical domain. Presented findings are ranging from the combination of different knowledge classifiers of the clinical arguments, training these models with enriched information, to the design of user interfaces for embedded intelligent agents. In addition, it is important to conduct further research to explore the potential benefits and limitations of different knowledge representation methods and to evaluate the effectiveness of different design features in enhancing the performance of AI systems. Ultimately, our findings aim to inform the development of more human-centered medical AI systems that effectively support clinical decision-making and enhance patient outcomes. As follows, we will provide our recommendations to inform future work on human-centered medical AI systems.

7.1.1 Different Knowledge Combination. In a real clinical workflow, extra patient information is necessary for a proper breast cancer diagnosis. Providing the lesion details and relevance of the classification is an important functionality for better decision-making. From our interviews, we learn that such information is crucial for diagnostic speed and accuracy, as it informs the clinician on what to look for and where to find the lesions. To better match the AI with the mental model of clinicians and provide better guidance, as well as explanations, we should incorporate granular patient information from the model classifiers [67]. For instance, the AI might use different classifiers to provide information on the lesion contours, whereas other classifiers are focused on the lesion margins. This assumption takes us to another recommendation, how should we train the models with such mixed information, for proper integration into real clinical workflows.

7.1.2 Training Mixed Models. Our study suggests that clinical workflows and trust can be positively affected by endowing personalization of the agent communication. In fact, with the ability to, not only incorporate granular patient information from the mixed model classifiers, but also adapt the tone depending on the medical experience of the clinician. Implementation of such intelligent agents would require that DL models are equipped with the additional prediction of mixed clinical arguments (*e.g.*, lesion contours, margin, or cancer type of the patient) beyond diagnosis alone. This additional granular information about the patient could include its importance to the diagnostic, while also customizing the communication tone depending on the various demographic characteristics of clinicians. Such an idea could be integrated either into one fused training, or by developing multi separated models, one for each clinical variable.

7.1.3 Adapting Communication. In this work, we evaluate one specific way of personalizing and customizing the communication between agents and clinicians with different levels of medical experience. We did that by exploring how to adapt the communication tone depending on if the agent was communicating with a novice or an expert clinician. While our results suggest that this communication technique may be effective, we recommend that future work may explore different demographic characteristics of clinicians. For example, from different medical institutions (*e.g.*, public hospitals, private clinics, cancer centers, etc), or different medical fields (*e.g.*, family physicians, breast surgeons, etc.), where some behavioral decision-making of clinicians should differ. Besides, our qualitative results (Section 6.3.2) are showing that clinicians are also willing for adapting the communication tone of the clinical arguments depending on the AI confidence. For instance, if confidence is greater than 80%, then the system should display an Assertive recommendation (Figure 3, middle). Otherwise, it should display a Non-Assertive recommendation (Figure 3, bottom). As a research direction, we should explore how different performance actions of the intelligent agents will impact behavioral decision-making of clinicians.

7.1.4 Generalizability. Through this exploration, our study sheds light on the use of assertiveness-based agents in the specific domain of breast cancer based on medical imaging diagnosis. It is important to note that the results of our study should not be generalized to other medical domains without caution. Hence, we warn in generalizing the results of this study to other domains. However, we argue that similar communication techniques of personalized and customized explanations can be useful for various types of medical diagnosis because the challenges motivating our study are common across several other medical specialties.

Despite our focus on the breast cancer domain, this demographic characteristics of clinicians (*i.e.*, differences in behavioral decision-making between novice and expert) is transversal to other applications [57, 90, 106]. Such claim is making our approach useful beyond the specific domain of breast cancer diagnosis. For instance, lung cancer diagnosis requires that specialized radiologists visually inspect chest imaging data similar in nature to that used in our study [121]. In both of these fields, it may be valuable to personalize and customize the agent communication, depending on their background and expertise, leading to more accurate diagnoses and improved patient outcomes.

The potential of personalizing and customizing the agent communication depending on the levels of medical experience can also be addressed for other clinical domains. As another example, in skin cancer some works are trying to mimic the medical procedures, where clinicians rely on their past experience across similar cases to reach the final diagnosis [11, 35, 114]. Depending on the past experience of that clinician, the agent should adapt the provided information and communication to them, or other customizable techniques. Others are stating that AI-based systems must be improved with personalized medicine supporting diagnosis and treatment guidance [3, 50, 104]. These studies suggest that the recommendations we make for medical imaging diagnosis in this work have been considered independently and may be of merit beyond the development of assertiveness-based agents.

7.2 Limitations and Future Work

In this work, we conducted a within-subject experiment to investigate the use of assertiveness-based agents by clinicians in the particular medical domain of breast cancer diagnosis. We investigate this question through the design and study of Assertiveness-based BreastScreening-AI. More specifically, this tool was used to explore how an intelligent agent should adapt its communication tone depending on the professional experience (*i.e.*, novice vs expert) of the clinician.

Due to the short availability of clinicians and the remote nature of our study, it was challenging to control the tasks of each step in the experiment precisely. For example, participants varied how long they complete the task for the first patient, in comparison to the second and third patients. This lack in experimental control may have impacted the degree to which exposure to the first patient, while interacting for the first time with the assertiveness-based agent, affected how clinicians interacted with the latter.

Another limitation is related with the implications of liability when using AI in medical settings [110], where the legal framework for addressing these issues is still evolving. The use of AI in medical settings raises complex questions that are not yet fully understood or addressed by existing laws and regulations [118]. This can make it difficult to determine who might be liable in the event of an error or harm caused by an AI system. AI systems often operate in complex and dynamic environments, making it challenging to identify the specific factors that led to a particular outcome. Overall, the limitations concerning the implications of liability when using AI in medical settings highlight the need for further research and legal developments in this area. It is important for policymakers and other stakeholders to continue to explore these issues and work to address them in a way that ensures the safe and effective use of AI in medical settings.

In our study, the assertiveness-based agent was using specific AI outputs, curated and selected by us alongside choosing the most typical clinical setups in a real-world environment. While prior work has demonstrated the potential of predicting the likelihood of a clinician to trust on an AI recommendation from raw medical data [84], future work should focus on training DL models based on personalized and customized explanations to provide human-interpretable arguments for clinicians.

As another future direction, we will study the effects of the two different main features of the assertiveness-based agent (*i.e.*, explanations and tone), in more conditions separately (Section A.7). However, inferring useful information for adapting the DL model, presents new technical challenges for the AI community. For the HCI community, the challenge is making such inferences transparent, considering some behavioral characteristics of clinicians.

8 CONCLUSION

In this work, we provide a novel perspective on how to personalize and customize the explanations of intelligent agents to human clinicians. Our results from an experimental study with 52 clinicians comparing a conventional agent to an assertiveness-based agent suggest that the ability of a system to not only exploring how to adapt the communication tone (*i.e.*, more suggestive or more assertive), but also provide granular explanations of patient cases has merits for end users. From our results, the time performance was satisfactory, where clinicians took less 25% of the time to diagnose a patient with the assertiveness-based agent in comparison to the conventional agent. As we observed, the caparison between the conventional agent and the assertiveness-based agent was more effective also with the latter in achieving the proper diagnostic of the patient. Additionally, our results demonstrate that if explanations are adapted taking into account the medical experience of clinicians, accuracy chance of correctly diagnosing a patient is 91% higher for novice and 78% higher for expert clinicians. Last, clinicians are showing an increase of trust, preferring the assertiveness-based agent by being more reliable and capable, as this agent was revealing to be further understandable, competent and thoughtful. Our work has implications for the design of AI systems not only in the medical domains, but also in fields that are facing similar challenges, demanding a personalization of the human-AI interaction.

ACKNOWLEDGMENTS

This work was supported by Fundação para a Ciência e a Tecnologia within two funding sources. First, the work was supported through a project with the reference LARSyS - FCT Project 2022.04485.PTDC. Second, funding was also provided under the Research & Development Strategic Plan - 2013/2015 - with the reference number UID/EEA/50009/2013 and the grant number PD/BD/150629/2020. We would like to thank participants, as well as respective clinical institutions, for the generous time spent and clinical expertise. A special thank to Dr. Clara Aleluia and her radiology team of HFF for valuable insights and collaboration within this work. Another great thank to Dr. Cristina Ribeiro da Fonseca who, among others, is giving us crucial directions. Finally, we would like to thank Catarina Barata, Hugo Lencastre, Nádia Mourão, João Bernardo, Madalena Pedreira, Pedro Diogo, Mauro Machado, and Miguel Bastos for their development, support, feedback, and opinions. Unfortunately, we could not address all people to acknowledge within limit concerns. Therefore, further acknowledgments are provided inside the `ACKNOWLEDGMENTS.md` file in the `sa-uta11-results` repository (github.com/MIMBCD-UI/sa-uta11-results) for the purpose. Additionally, we are grateful for the invaluable assistance provided by our colleagues at Carnegie Mellon University. We are indebted to those who gave their time and expertise to evaluate our work.

REFERENCES

[1] Enas Mahrous Abdelaziz, Iman Abdelmotelb Diab, Marwa Mohamed Ahmed Ouda, Nadia Bassiouni Elsharkawy, and Fadia Ahmed Abdelkader. 2020. The effectiveness of assertiveness training program on psychological well-being and work engagement among novice psychiatric nurses. *Nursing Forum* 55, 3 (2020), 309–319. https://doi.org/10.1111/nuf.12430 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/nuf.12430

[2] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1â€“18. https://doi.org/10.1145/3173574.3174156

[3] Hugo J. W. L. Aerts, Patrick Grossmann, Yongqiang Tan, Geoffrey R. Oxnard, Naiyer Rizvi, Lawrence H. Schwartz, and Binsheng Zhao. 2016. Defining a Radiomic Response Phenotype: A Pilot Study using targeted therapy in NSCLC. *Scientific Reports* 6, 1 (20 Sep 2016), 33860. https://doi.org/10.1038/srep33860

[4] David W Aha and Alexandra Coman. 2017. The AI rebellion: Changing the narrative. In *Thirty-First AAAI Conference on Artificial Intelligence*. AAAI, San Francisco, California, USA, 4826–4830.

[5] Nader Aldoj, Steffen Lukas, Marc Dewey, and Tobias Penzkofer. 2020. Semi-automatic classification of prostate cancer on multi-parametric MR imaging using a multi-channel 3D convolutional neural network. *European Radiology* 30, 2 (01 Feb 2020), 1243–1253. https://doi.org/10.1007/s00330-019-06417-z

[6] Robert Alexander, Stephen Waite, Michael A. Bruno, Elizabeth A. Krupinski, Leonard Berlin, Stephen Macknik, and Susana Martinez-Conde. 2022. Mandating Limits on Workload, Duty, and Speed in Radiology. *Radiology* 304, 2 (2022), 274–282. https://doi.org/10.1148/radiol.212631 arXiv:https://doi.org/10.1148/radiol.212631 PMID: 35699581.

[7] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I. Madai, and the Precise4Q consortium. 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making* 20, 1 (30 Nov 2020), 310. https://doi.org/10.1186/s12911-020-01332-6

[8] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI â€™19)*. Association for Computing Machinery, New York, NY, USA, 1â€“13. https://doi.org/10.1145/3290605.3300233

[9] Francisco Azuaje. 2019. Artificial intelligence for precision oncology: beyond patient stratification. *NPJ precision oncology* 3, 1 (2019), 6.

[10] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[11] Catarina Barata and Carlos Santiago. 2021. Improving the Explainability of Skin Cancer Diagnosis Using CBIR. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert (Eds.). Springer International Publishing, Cham, 550–559.

[12] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1â€“12. https://doi.org/10.1145/3313831.3376718

[13] Wenya Linda Bi, Ahmed Hosny, Matthew B. Schabath, Maryellen L. Giger, Nicolai J. Birkbak, Alireza Mehrtash, Tavis Allison, Omar Arnaout, Christopher Abbosh, Ian F. Dunn, Raymond H. Mak, Rulla M. Tamimi, Clare M. Tempany, Charles Swanton, Udo Hoffmann, Lawrence H. Schwartz, Robert J. Gillies, Raymond Y. Huang, and Hugo J. W. L. Aerts. 2019. Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA: A Cancer Journal for Clinicians* 69, 2 (2019), 127–157. https://doi.org/10.3322/caac.21552 arXiv:https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21552

[14] Till Bieg, Cornelia Gerdenitsch, Isabel Schwaninger, Bettina Manuela Johanna Kern, and Christopher Frauenberger. 2022. Evaluating Active and Assisted Living technologies: Critical methodological reflections based on a longitudinal randomized controlled trial. *Computers in Human Behavior* 133 (2022), 107249. https://doi.org/10.1016/j.chb.2022.107249

[15] Tommy Bruzzese, Irena Gao, Griffin Dietz, Christina Ding, and Alyssa Romanos. 2020. Effect of Confidence Indicators on Trust in AI-Generated Profiles. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1â€“8. https://doi.org/10.1145/3334480.3382842

[16] Wesley Buckwalter. 2014. The Cognitive Science of Science: Explanation, Discovery, and Conceptual Change, by Paul Thagard. *Mind* 122, 488 (03 2014), 1201–1204. https://doi.org/10.1093/mind/fzu023 arXiv:https://academic.oup.com/mind/article-pdf/122/488/1201/2982379/fzu023.pdf

[17] Christopher Burr, Nello Cristianini, and James Ladyman. 2018. An Analysis of the Interaction Between Intelligent Software Agents and Human Users. *Minds and machines* 28, 4 (2018), 735–774.

[18] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The Effects of Example-Based Explanations in a Machine Learning Interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 258â€“262. https://doi.org/10.1145/3301275.3302289

[19] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1â€“14. https://doi.org/10.1145/3290605.3300234

[20] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 104 (nov 2019), 24 pages. https://doi.org/10.1145/3359206

[21] Francisco Maria Calisto and Jacinto C. Nascimento. 2020. Computational Method and System for Improved Identification of Breast Lesions. Retrieved April from https://patents.google.com/patent/WO2022071818A1 Application PCT/PT2021/050029 events of Patent No. 2022071818, Filed by Instituto Superior Técnico on September 30th., 2020 as Priority to PT116801 and PT116801A, Issued April 4th., 2022.

[22] Francisco M. Calisto, Alfredo Ferreira, Jacinto C. Nascimento, and Daniel Gonçalves. 2017. Towards Touch-Based Medical Image Diagnosis Annotation. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces* (Brighton, United Kingdom) *(ISS â€™17)*. Association for Computing Machinery, New York, NY, USA, 390â€“395. https://doi.org/10.1145/3132272.3134111

[23] Francisco Maria Calisto, Nuno Nunes, and Jacinto C. Nascimento. 2022. Modeling adoption of intelligent agents in medical imaging. *International Journal of Human-Computer Studies* 168 (2022), 102922. https://doi.org/10.1016/j.ijhcs.2022.102922

[24] Francisco Maria Calisto, Nuno Jardim Nunes, and Jacinto Carlos Nascimento. 2020. BreastScreening: On the Use of Multi-Modality in Medical Imaging Diagnosis. https://doi.org/10.13140/RG.2.2.28548.27523 arXiv:2004.03500 [cs.HC]

[25] Francisco Maria Calisto, Carlos Santiago, Nuno Nunes, and Jacinto C. Nascimento. 2021. Introduction of human-centric AI assistant to aid radiologists for multimodal breast image classification. *International Journal of Human-Computer Studies* 150 (2021), 102607. https://doi.org/10.1016/j.ijhcs.2021.102607

[26] Francisco Maria Calisto, Carlos Santiago, Nuno Nunes, and Jacinto C. Nascimento. 2022. BreastScreening-AI: Evaluating medical intelligent agents for human-AI interactions. *Artificial Intelligence in Medicine* 127 (2022), 102285. https://doi.org/10.1016/j.artmed.2022.102285

[27] Silvia Casale, Giulia Fioravanti, Sara Bocci Benucci, Andrea Falone, Valdo Ricca, and Francesco Rotella. 2022. A meta-analysis on the association between self-esteem and problematic smartphone use. *Computers in Human Behavior* 134 (2022), 107302. https://doi.org/10.1016/j.chb.2022.107302

[28] Jung Min Chang, Jessica W. T. Leung, Linda Moy, Su Min Ha, and Woo Kyung Moon. 2020. Axillary Nodal Evaluation in Breast Cancer: State of the Art. *Radiology* 295, 3 (2020), 500–515. https://doi.org/10.1148/radiol.2020192534 arXiv:https://doi.org/10.1148/radiol.2020192534 PMID: 32315268.

[29] Elodia B Cole, Zheng Zhang, Helga S Marques, R Edward Hendrick, Martin J Yaffe, and Etta D Pisano. 2014. Impact of computer-aided detection systems on radiologist accuracy with digital mammography. *AJR Am J Roentgenol* 203, 4 (Oct. 2014), 909–916.

[30] Inbal Deutsch, Hadas Erel, Michal Paz, Guy Hoffman, and Oren Zuckerman. 2019. Home robotic devices for older adults: Opportunities and concerns. *Computers in Human Behavior* 98 (2019), 122–133. https://doi.org/10.1016/j.chb.2019.04.002

[31] Anind K. Dey and Alan Newberger. 2009. Support for Context-Aware Intelligibility and Control. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 859â€“868. https://doi.org/10.1145/1518701.1518832

[32] Paul Dourish, Christopher Lawrence, Tuck Wah Leong, and Greg Wadley. 2020. On Being Iterated: The Affective Demands of Design Participation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1â€“11. https://doi.org/10.1145/3313831.3376545

[33] Amanda K. Edgar, Lucinda Ainge, Simon Backhouse, and James A. Armitage. 2022. A cohort study for the development and validation of a reflective inventory to quantify diagnostic reasoning skills in optometry practice. *BMC Medical Education* 22, 1 (11 Jul 2022), 536. https://doi.org/10.1186/s12909-022-03493-6

[34] Glyn Elwyn, Isabelle Scholl, Caroline Tietbohl, Mala Mann, Adrian GK Edwards, Catharine Clay, France Légaré, Trudy van der Weijden, Carmen L. Lewis, Richard M. Wexler, and Dominick L. Frosch. 2013. "Many miles to go …": a systematic review of the implementation of patient decision support interventions into routine clinical practice. *BMC Medical Informatics and Decision Making* 13, 2 (29 Nov 2013), S14. https://doi.org/10.1186/1472-6947-13-S2-S14

[35] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. Correction: Corrigendum: Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 546, 7660 (01 Jun 2017), 686–686. https://doi.org/10.1038/nature22985

[36] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (2017), 115.

[37] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. 2019. A guide to deep learning in healthcare. *Nature medicine* 25, 1 (2019), 24.

[38] Jerry Alan Fails and Dan R. Olsen. 2003. Interactive Machine Learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces* (Miami, Florida, USA) *(IUI â€™03)*. Association for Computing Machinery, New York, NY, USA, 39â€“45. https://doi.org/10.1145/604045.604056

[39] Farshid Faraji and Ron C Gaba. 2019. Radiologic Modalities and Response Assessment Schemes for Clinical and Preclinical Oncology Imaging. *Frontiers in Oncology* 9 (2019), 1–12.

[40] Geraldine Fitzpatrick and Gunnar Ellingsen. 2013. A Review of 25 Years of CSCW Research in Healthcare: Contributions, Challenges and Future Agendas. *Computer Supported Cooperative Work (CSCW)* 22, 4 (01 Aug 2013), 609–665. https://doi.org/10.1007/s10606-012-9168-0

[41] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. 2022. Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1362â€“1374. https://doi.org/10.1145/3531146.3533193

[42] Ariel Goldman, Cindy Espinosa, Shivani Patel, Francesca Cavuoti, Jade Chen, Alexandra Cheng, Sabrina Meng, Aditi Patil, Lydia B Chilton, and Sarah Morrison-Smith. 2022. QuAD: Deep-Learning Assisted Qualitative Data Analysis with Affinity Diagrams. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 419, 7 pages. https://doi.org/10.1145/3491101.3519863

[43] Tovi Grossman and George Fitzmaurice. 2010. ToolClips: An Investigation of Contextual Video Assistance for Functionality Understanding. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1515â€“1524. https://doi.org/10.1145/1753326.1753552

[44] Emrah Hancer and Abdulhamit Subasi. 2023. Chapter 13 - Diagnosis of breast cancer from histopathological images with deep learning architectures. In *Applications of Artificial Intelligence in Medical Imaging*, Abdulhamit Subasi (Ed.). Academic Press, San Diego, USA, 321–332. https://doi.org/10.1016/B978-0-443-18450-5.00002-5

[45] Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia, and Andrew Y. Ng. 2019. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine* 25, 1 (01 Jan 2019), 65–69. https://doi.org/10.1038/s41591-018-0268-3

[46] Dean Ho, Stephen R. Quake, Edward R.B. McCabe, Wee Joo Chng, Edward K. Chow, Xianting Ding, Bruce D. Gelb, Geoffrey S. Ginsburg, Jason Hassenstab, Chih-Ming Ho, William C. Mobley, Garry P. Nolan, Steven T. Rosen, Patrick Tan, Yun Yen, and Ali Zarrinpar. 2020. Enabling Technologies for Personalized and Precision Medicine. *Trends in Biotechnology* 38, 5 (2020), 497–518. https://doi.org/10.1016/j.tibtech.2019.12.021

[47] Jess Hohenstein and Malte Jung. 2020. AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior* 106 (2020), 106190. https://doi.org/10.1016/j.chb.2019.106190

[48] Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 3, 2 (2016), 119–131.

[49] Gao Huang, Zhuang Liu, Geoff Pleiss, Laurens Van Der Maaten, and Kilian Weinberger. 2019. Convolutional Networks with Dense Connectivity. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1, 1 (2019), 1–1. https://doi.org/10.1109/TPAMI.2019.2918284

[50] Abdalla Ibrahim, Martin ValliÃ¨res, Henry Woodruff, Sergey Primakov, Mohsen Beheshti, Simon Keek, Turkey Refaee, Sebastian Sanduleanu, Sean Walsh, Olivier Morin, Philippe Lambin, Roland Hustinx, and Felix M. Mottaghy. 2019. Radiomics Analysis for Clinical Decision Support in Nuclear Medicine. *Seminars in Nuclear Medicine* 49, 5 (2019), 438–449. https://doi.org/10.1053/j.semnuclmed.2019.06.005

[51] Takeo Igarashi, Naoyuki Shono, Taichi Kin, and Toki Saito. 2016. Interactive Volume Segmentation with Threshold Field Painting. In *Annual Symposium on User Interface Software and Technology (UIST)*. ACM, New York, NY, USA, 403–413. https://doi.org/10.1145/2984511.2984537

[52] Mohammad Hossein Jarrahi. 2018. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons* 61, 4 (2018), 577–586. https://doi.org/10.1016/j.bushor.2018.03.007

[53] Suzanne Kawamleh. 2022. Against explainability requirements for ethical artificial intelligence in health care. *AI and Ethics* 1, 1 (29 Aug 2022), 1–16. https://doi.org/10.1007/s43681-022-00212-1

[54] Hyo-Eun Kim, Hak Hee Kim, Boo-Kyung Han, Ki Hwan Kim, Kyunghwa Han, Hyeonseob Nam, Eun Hye Lee, and Eun-Kyung Kim. 2020. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *The Lancet Digital Health* 2, 3 (2020), e138–e148. https://doi.org/10.1016/S2589-7500(20)30003-0

[55] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, Article 411, 14 pages. https://doi.org/10.1145/3290605.3300641

[56] Ajay Kohli and Saurabh Jha. 2018. Why CAD Failed in Mammography. *Journal of the American College of Radiology* 15, 3, Part B (2018), 535–537. https://doi.org/10.1016/j.jacr.2017.12.029 Data Science: Big Data Machine Learning and Artificial Intelligence.

[57] Markus LandrÃ„, Audun Hetland, Rune Verpe Engeset, and Gerit Pfuhl. 2020. Avalanche decision-making frameworks: Factors and methods used by experts. *Cold Regions Science and Technology* 170 (2020), 102897. https://doi.org/10.1016/j.coldregions.2019.102897

[58] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data* 5 (2018), 180251.

[59] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436.

[60] James R. Lewis. 2018. The System Usability Scale: Past, Present, and Future. *International Journal of Humanâ€“Computer Interaction* 34, 7 (2018), 577–590. https://doi.org/10.1080/10447318.2018.1455307

[61] Yang Li, Ranjitha Kumar, Walter S. Lasecki, and Otmar Hilliges. 2020. Artificial Intelligence for HCI: A Modern Approach. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1â€“8. https://doi.org/10.1145/3334480.3375147

[62] Tania Lombrozo. 2010. Causalâ€“explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology* 61, 4 (2010), 303–332. https://doi.org/10.1016/j.cogpsych.2010.05.002

[63] Margaret L. Loper. 2020. *Dimensions of Trust in Cyber Physical Systems*. Springer International Publishing, Cham, 407–427. https://doi.org/10.1007/978-3-030-51909-4_16

[64] Daniel Simões Lopes, Pedro Duarte de Figueiredo Parreira, Soraia Figueiredo Paulo, Vitor Nunes, Paulo Amaral Rego, Manuel Cassiano Neves, Pedro Silva Rodrigues, and Joaquim Armando Jorge. 2017. On the Utility of 3D Hand Cursors to Explore Medical Volume Datasets with a Touchless Interface. *J. of Biomedical Informatics* 72, C (Aug. 2017), 140–149. https://doi.org/10.1016/j.jbi.2017.07.009

[65] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2, 1 (01 Jan 2020), 56–67. https://doi.org/10.1038/s42256-019-0138-9

[66] Scott M. Lundberg, Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Adams, David E. Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, and Su-In Lee. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering* 2, 10 (01 Oct 2018), 749–760. https://doi.org/10.1038/s41551-018-0304-0

[67] Luyang Luo, Hao Chen, Yongjie Xiao, Yanning Zhou, Xi Wang, Varut Vardhanabhuti, Mingxiang Wu, Chu Han, Zaiyi Liu, Xin Hao Benjamin Fang, Efstratios Tsougenis, Huangjing Lin, and Pheng-Ann Heng. 2022. Rethinking Annotation Granularity for Overcoming Shortcuts in Deep Learningâ€“based Radiograph Diagnosis: A Multicenter Study. *Radiology: Artificial Intelligence* 4, 5 (2022), e210299. https://doi.org/10.1148/ryai.210299 arXiv:https://doi.org/10.1148/ryai.210299 PMID: 35146431.

[68] Lech Madeyski and Barbara Kitchenham. 2018. Effect Sizes and Their Variance for AB/BA Crossover Design Studies. In *Proceedings of the 40th International Conference on Software Engineering* (Gothenburg, Sweden) *(ICSE '18)*. Association for Computing Machinery, New York, NY, USA, 420. https://doi.org/10.1145/3180155.3182556

[69] Gabriel Maicas, Andrew P. Bradley, Jacinto C. Nascimento, Ian Reid, and Gustavo Carneiro. 2019. Pre and post-hoc diagnosis and interpretation of malignancy from breast DCE-MRI. *Medical Image Analysis* 58 (2019), 101562. https://doi.org/10.1016/j.media.2019.101562

[70] Lena Maier-Hein, Matthias Eisenmann, Duygu Sarikaya, Keno MÃ¤rz, Toby Collins, Anand Malpani, Johannes Fallert, Hubertus Feussner, Stamatia Giannarou, Pietro Mascagni, Hirenkumar Nakawala, Adrian Park, Carla Pugh, Danail Stoyanov, Swaroop S. Vedula, Kevin Cleary, Gabor Fichtinger, Germain Forestier, Bernard Gibaud, Teodor Grantcharov, Makoto Hashizume, Doreen Heckmann-NÃ¶tzel, Hannes G. Kenngott, Ron Kikinis, Lars MÃ¼ndermann, Nassir Navab, Sinan Onogur, Tobias RoÃŸ, Raphael Sznitman, Russell H. Taylor, Minu D. Tizabi, Martin Wagner, Gregory D. Hager, Thomas Neumuth, Nicolas Padoy, Justin Collins, Ines Gockel, Jan Goedeke, Daniel A. Hashimoto, Luc Joyeux, Kyle Lam, Daniel R. Leff, Amin Madani, Hani J. Marcus, Ozanan Meireles, Alexander Seitel, Dogu Teber, Frank Ückert, Beat P. MÃ¼ller-Stich, Pierre Jannin, and Stefanie Speidel. 2022. Surgical data science â€“ from concepts toward clinical translation. *Medical Image Analysis* 76 (2022), 102306. https://doi.org/10.1016/j.media.2021.102306

[71] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg C Corrado, Ara Darzi, et al. 2020. International evaluation of an AI system for breast cancer screening. *Nature* 577, 7788 (2020), 89–94.

[72] Sofia Meacham, Georgia Isaac, Detlef Nauck, and Botond Virginas. 2019. Towards Explainable AI: Design and Development for Explanation of Machine Learning Predictions for a Patient Readmittance Medical Application. In *Intelligent Computing*, Kohei Arai, Rahul Bhatia, and Supriya Kapoor (Eds.). Springer International Publishing, Cham, 939–955.

[73] D. O. Medley, C. Santiago, and J. C. Nascimento. 2019. Segmenting The Left Ventricle In Cardiac In Cardiac MRI: From Handcrafted To Deep Region Based Descriptors. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, Venice, Italy, 644–648.

[74] Anthony Miller. 2019. The intrinsically linked future for human and Artificial Intelligence interaction. *Journal of Big Data* 6, 1 (2019), 38.

[75] Mark A. Musen, Blackford Middleton, and Robert A. Greenes. 2021. *Clinical Decision-Support Systems*. Springer International Publishing, Cham, 795–840. https://doi.org/10.1007/978-3-030-58721-5_24

[76] Mohammad Naiseh, Dena Al-Thani, Nan Jiang, and Raian Ali. 2023. How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies* 169 (2023), 102941. https://doi.org/10.1016/j.ijhcs.2022.102941

[77] Moses Namara and Bart P. Knijnenburg. 2021. The Differential Effect of Privacy-Related Trust on Groupware Application Adoption and Use during the COVID-19 Pandemic. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 405 (oct 2021), 34 pages. https://doi.org/10.1145/3479549

[78] Mei-Sing Ong and Kenneth D. Mandl. 2015. National Expenditure For False-Positive Mammograms And Breast Cancer Overdiagnoses Estimated At $4 Billion A Year. *Health Affairs* 34, 4 (2015), 576–583. https://doi.org/10.1377/hlthaff.2014.1087 arXiv:https://doi.org/10.1377/hlthaff.2014.1087

[79] António C Pacheco and Carlos Martinho. 2019. Alignment of Player and Non-Player Character Assertiveness Levels. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 15. AAAI, Georgia Institute of Technology, Atlanta, Georgia, USA, 181–187.

[80] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 2022. 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* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 568, 9 pages. https://doi.org/10.1145/3491102.3502104

[81] Raul Paradeda, Maria José Ferreira, Raquel Oliveira, Carlos Martinho, and Ana Paiva. 2019. The Role of Assertiveness in a Storytelling Game with Persuasive Robotic Non-Player Characters. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (Barcelona, Spain) *(CHI PLAY â€™19)*. Association for Computing Machinery, New York, NY, USA, 453â€“465. https://doi.org/10.1145/3311350.3347162

[82] Sun Young Park, Pei-Yi Kuo, Andrea Barbarin, Elizabeth Kaziunas, Astrid Chow, Karandeep Singh, Lauren Wilcox, and Walter S. Lasecki. 2019. Identifying Challenges and Opportunities in Human-AI Collaboration in Healthcare. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing* (Austin, TX, USA) *(CSCW '19)*. Association for Computing Machinery, New York, NY, USA, 506â€“510. https://doi.org/10.1145/3311957.3359433

[83] Corina Pelau, Dan-Cristian Dabija, and Irina Ene. 2021. What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior* 122 (2021), 106855. https://doi.org/10.1016/j.chb.2021.106855

[84] Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Bobby Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. 2019. Direct Uncertainty Prediction for Medical Second Opinions. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, Long Beach, CA, USA, 5281–5290. https://proceedings.mlr.press/v97/raghu19a.html

[85] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J. Topol. 2022. AI in health and medicine. *Nature Medicine* 28, 1 (01 Jan 2022), 31–38. https://doi.org/10.1038/s41591-021-01614-0

[86] Carolin Reichherzer, Andrew Cunningham, Tracey Coleman, Ruochen Cao, Kurt McManus, Dion Sheppard, Mark Kohler, Mark Billinghurst, and Bruce H Thomas. 2021. Bringing the Jury to the Scene of the Crime: Memory and Decision-Making in a Simulated Crime Scene. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 709, 12 pages. https://doi.org/10.1145/3411764.3445464

[87] Abtin Riasatian, Morteza Babaie, Danial Maleki, Shivam Kalra, Mojtaba Valipour, Sobhan Hemati, Manit Zaveri, Amir Safarpoor, Sobhan Shafiei, Mehdi Afshari, Maral Rasoolijaberi, Milad Sikaroudi, Mohd Adnan, Sultaan Shah, Charles Choi, Savvas Damaskinos, Clinton JV Campbell, Phedias Diamandis, Liron Pantanowitz, Hany Kashani, Ali Ghodsi, and H.R. Tizhoosh. 2021. Fine-Tuning and training of densenet for histopathology image representation using TCGA diagnostic slides. *Medical Image Analysis* 70 (2021), 102032. https://doi.org/10.1016/j.media.2021.102032

[88] Paisan Ruamviboonsuk, Jonathan Krause, Peranut Chotcomwongse, Rory Sayres, Rajiv Raman, Kasumi Widner, Bilson J. L. Campana, Sonia Phene, Kornwipa Hemarat, Mongkol Tadarati, Sukhum Silpa-Archa, Jirawut Limwattanayingyong, Chetan Rao, Oscar Kuruvilla, Jesse Jung, Jeffrey Tan, Surapong Orprayoon, Chawawat Kangwanwongpaisan, Ramase Sukumalpaiboon, Chainarong Luengchaichawang, Jitumporn Fuangkaew, Pipat Kongsap, Lamyong Chualinpha, Sarawuth Saree, Srirut Kawinpanitan, Korntip Mitvongsa, Siriporn Lawanasakol, Chaiyasit Thepchatri, Lalita Wongpichedchai, Greg S. Corrado, Lily Peng, and Dale R. Webster. 2019. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *npj Digital Medicine* 2, 1 (10 Apr 2019), 25. https://doi.org/10.1038/s41746-019-0099-8

[89] Cynthia Rudin. 2022. Why black box machine learning should be avoided for high-stakes decisions, in brief. *Nature Reviews Methods Primers* 2, 1 (27 Oct 2022), 81. https://doi.org/10.1038/s43586-022-00172-0

[90] Gabor Ruzsa, Csenge Szeverenyi, and Katalin Varga. 2020. Person- and job-specific factors of intuitive decision-making in clinical practice: results of a sample survey among Hungarian physicians and nurses. *Health Psychology and Behavioral Medicine* 8, 1 (2020), 152–184. https://doi.org/10.1080/21642850.2020.1741372 arXiv:https://doi.org/10.1080/21642850.2020.1741372

[91] Delaram Sadeghi, Afshin Shoeibi, Navid Ghassemi, Parisa Moridian, Ali Khadem, Roohallah Alizadehsani, Mohammad Teshnehlab, Juan M. Gorriz, Fahime Khozeimeh, Yu-Dong Zhang, Saeid Nahavandi, and U Rajendra Acharya. 2022. An overview of artificial intelligence techniques for diagnosis of Schizophrenia based on magnetic resonance imaging modalities: Methods, challenges, and future works. *Computers in Biology and Medicine* 146 (2022), 105554. https://doi.org/10.1016/j.compbiomed.2022.105554

[92] Sadiq Said, Malgorzata Gozdzik, Tadzio Raoul Roche, Julia Braun, Julian Rössler, Alexander Kaserer, Donat R Spahn, Christoph B Nöthiger, and David Werner Tscholl. 2020. Validation of the Raw National Aeronautics and Space Administration Task Load Index (NASA-TLX) Questionnaire to Assess Perceived Workload in Patient Monitoring Tasks: Pooled Analysis Study Using Mixed Models. *J Med Internet Res* 22, 9 (7 Sep 2020), e19472. https://doi.org/10.2196/19472

[93] Mike Schaekermann, Graeme Beaton, Elaheh Sanoubari, Andrew Lim, Kate Larson, and Edith Law. 2020. Ambiguity-Aware AI Assistants for Medical Data Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1â€“14. https://doi.org/10.1145/3313831.3376506

[94] Mariah L. Schrum, Glen Neville, Michael Johnson, Nina Moorman, Rohan Paleja, Karen M. Feigh, and Matthew C. Gombolay. 2021. Effects of Social Factors and Team Dynamics on Adoption of Collaborative Robot Autonomy. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '21)*. Association for Computing Machinery, New York, NY, USA, 149â€“157. https://doi.org/10.1145/3434073.3444649

[95] Tina Seidel, Katharina Schnitzler, Christian Kosel, Kathleen Stürmer, and Doris Holzberger. 2021. Student Characteristics in the Eyes of Teachers: Differences Between Novice and Expert Teachers in Judgment Accuracy, Observed Behavioral Cues, and Gaze. *Educational Psychology Review* 33, 1 (01 Mar 2021), 69–89. https://doi.org/10.1007/s10648-020-09532-2

[96] Manish Sharma, Madhuri Madasu, Sree Sudha Kota, Surabhi Bajpai, Yibin Shao, Srinivas Pasupuleti, and Michael Oâ€™Connor. 2022. Using reader disagreement index as a tool for monitoring impact on read quality due to reader fatigue in central reviewers. In *Medical Imaging 2022: Image Perception, Observer Performance, and Technology Assessment*, Claudia R. Mello-Thoms and Sian Taylor-Phillips (Eds.), Vol. 12035. International Society for Optics and Photonics, SPIE, San Diego, California, United States, 120350J. https://doi.org/10.1117/12.2613082

[97] Yuya Shibuya, Andrea Hamm, and Teresa Cerratto Pargman. 2022. Mapping HCI research methods for studying social media interaction: A systematic literature review. *Computers in Human Behavior* 129 (2022), 107131. https://doi.org/10.1016/j.chb.2021.107131

[98] Ben Shneiderman. 2016. The dangers of faulty, biased, or malicious algorithms requires independent oversight. *Proceedings of the National Academy of Sciences* 113, 48 (2016), 13538–13540. https://doi.org/10.1073/pnas.1618211113

[99] Ben Shneiderman. 2022. Commentary: extraordinary excitement empowering enhancing everyone. *Humanâ€“Computer Interaction* 37, 3 (2022), 243–245. https://doi.org/10.1080/07370024.2021.1977128

[100] Ben Shneiderman. 2022. Human-Centered AI: Ensuring Human Control While Increasing Automation. In *Proceedings of the 5th Workshop on Human Factors in Hypertext* (Barcelona, Spain) *(HUMAN '22)*. Association for Computing Machinery, New York, NY, USA, Article 1, 2 pages. https://doi.org/10.1145/3538882.3542790

[101] Edward H. Shortliffe and Martin J. SepÃºlveda. 2018. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA* 320, 21 (12 2018), 2199–2200. https://doi.org/10.1001/jama.2018.17163

[102] Dean F Sittig, Priti Lakhani, and Hardeep Singh. 2022. Applying requisite imagination to safeguard electronic health record transitions. *Journal of the American Medical Informatics Association* 29, 5 (01 2022), 1014–1018. https://doi.org/10.1093/jamia/ocab291 arXiv:https://academic.oup.com/jamia/article-pdf/29/5/1014/43372323/ocab291.pdf

[103] Rebecca Sivarajah, Mary L Dinh, and Alison Chetlen. 2021. Errors in Breast Imaging: How to Reduce Errors and Promote a Safety Environment. *Journal of Breast Imaging* 3, 2 (01 2021), 221–230. https://doi.org/10.1093/jbi/wbaa118 arXiv:https://academic.oup.com/jbi/article-pdf/3/2/221/36648802/wbaa118.pdf

[104] Martina Sollini, Francesco Bartoli, Andrea Marciano, Roberta Zanca, Riemer H. J. A. Slart, and Paola A. Erba. 2020. Artificial intelligence and hybrid imaging: the best match for personalized medicine in oncology. *European Journal of Hybrid Imaging* 4, 1 (09 Dec 2020), 24. https://doi.org/10.1186/s41824-020-00094-8

[105] D.A. Spak, J.S. Plaxco, L. Santiago, M.J. Dryden, and B.E. Dogan. 2017. BI-RADSÂ® fifth edition: A summary of changes. *Diagnostic and Interventional Imaging* 98, 3 (2017), 179–190. https://doi.org/10.1016/j.diii.2017.01.001

[106] Rebekka Stahnke and Sigrid BlÃ¶meke. 2021. Novice and expert teachersâ€™ situation-specific skills regarding classroom management: What do they perceive, interpret and suggest? *Teaching and Teacher Education* 98 (2021), 103243. https://doi.org/10.1016/j.tate.2020.103243

[107] Jens B. Stephansen, Alexander N. Olesen, Mads Olsen, Aditya Ambati, Eileen B. Leary, Hyatt E. Moore, Oscar Carrillo, Ling Lin, Fang Han, Han Yan, Yun L. Sun, Yves Dauvilliers, Sabine Scholz, Lucie Barateau, Birgit Hogl, Ambra Stefani, Seung Chul Hong, Tae Won Kim, Fabio Pizza, Giuseppe Plazzi, Stefano Vandi, Elena Antelmi, Dimitri Perrin, Samuel T. Kuna, Paula K. Schweitzer, Clete Kushida, Paul E. Peppard, Helge B. D. Sorensen, Poul Jennum, and Emmanuel Mignot. 2018. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature Communications* 9, 1 (06 Dec 2018), 5229. https://doi.org/10.1038/s41467-018-07229-3

[108] Li Sturesdotter, Malte Sandsveden, Kristin Johnson, Anna-Maria Larsson, Sophia Zackrisson, and Hanna Sartor. 2020. Mammographic tumour appearance is related to clinicopathological factors and surrogate molecular breast cancer subtype. *Scientific Reports* 10, 1 (30 Nov 2020), 20814. https://doi.org/10.1038/s41598-020-77053-7

[109] Xin Su, Li An, Zhen Cheng, and Yajuan Weng. 2023. Cloudâ€“edge collaboration-based bi-level optimal scheduling for intelligent healthcare systems. *Future Generation Computer Systems* 141 (2023), 28–39. https://doi.org/10.1016/j.future.2022.11.005

[110] Zhaoyuan Su, Lu He, Sunit P Jariwala, Kai Zheng, and Yunan Chen. 2022. "What is Your Envisioned Future?": Toward Human-AI Enrichment in Data Work of Asthma Care. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 267 (nov 2022), 28 pages. https://doi.org/10.1145/3555157

[111] Murugan Subramanian, Anne Wojtusciszyn, Lucie Favre, Sabri Boughorbel, Jingxuan Shan, Khaled B. Letaief, Nelly Pitteloud, and Lotfi Chouchane. 2020. Precision medicine in the era of artificial intelligence: implications in chronic disease management. *Journal of Translational Medicine* 18, 1 (09 Dec 2020), 472. https://doi.org/10.1186/s12967-020-02658-5

[112] Muhammed Talo, Ozal Yildirim, Ulas Baran Baloglu, Galip Aydin, and U Rajendra Acharya. 2019. Convolutional neural networks for multi-class brain disease detection using MRI images. *Computerized Medical Imaging and Graphics* 78 (2019), 101673. https://doi.org/10.1016/j.compmedimag.2019.101673

[113] Tao Tan, Alejandro Rodriguez-Ruiz, Tianyu Zhang, Lin Xu, Regina G. H. Beets-Tan, Yingzhao Shen, Nico Karssemeijer, Jun Xu, Ritse M. Mann, and Lingyun Bao. 2023. Multi-modal artificial intelligence for the combination of automated 3D breast ultrasound and mammograms in a population of women with predominantly dense breasts. *Insights into Imaging* 14, 1 (16 Jan 2023), 10. https://doi.org/10.1186/s13244-022-01352-y

[114] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, John Paoli, Susana Puig, Cliff Rosendahl, H. Peter Soyer, Iris Zalaudek, and Harald Kittler. 2020. Human–computer collaboration for skin cancer recognition. *Nature Medicine* 26, 8 (01 Aug 2020), 1229–1234. https://doi.org/10.1038/s41591-020-0942-0

[115] Nusrat Mohi ud din, Rayees Ahmad Dar, Muzafar Rasool, and Assif Assad. 2022. Breast cancer detection using deep learning: Datasets, methods, and challenges ahead. *Computers in Biology and Medicine* 149 (2022), 106073. https://doi.org/10.1016/j.compbiomed.2022.106073

[116] Mohammed Uddin, Yujiang Wang, and Marc Woodbury-Smith. 2019. Artificial intelligence for precision medicine in neurodevelopmental disorders. *npj Digital Medicine* 2, 1 (21 Nov 2019), 112. https://doi.org/10.1038/s41746-019-0191-0

[117] Almar van der Stappen and Mathias Funk. 2021. Towards Guidelines for Designing Human-in-the-Loop Machine Training Interfaces. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 514â€“519. https://doi.org/10.1145/3397481.3450668

[118] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI Doctor" in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 697, 18 pages. https://doi.org/10.1145/3411764.3445432

[119] Shuo Wang, Zhenyu Liu, Yu Rong, Bin Zhou, Yan Bai, Wei Wei, Meiyun Wang, Yingkun Guo, and Jie Tian. 2019. Deep learning provides a new computed tomography-based prognostic biomarker for recurrence prediction in high-grade serous ovarian cancer. *Radiotherapy and Oncology* 132 (2019), 171–177.

[120] Gordon Wetzstein, Aydogan Ozcan, Sylvain Gigan, Shanhui Fan, Dirk Englund, Marin Soljačić, Cornelia Denz, David A. B. Miller, and Demetri Psaltis. 2020. Inference in artificial intelligence with deep optics and photonics. *Nature* 588, 7836 (01 Dec 2020), 39–47. https://doi.org/10.1038/s41586-020-2973-6

[121] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang 'Anthony' Chen. 2020. CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1â€“13. https://doi.org/10.1145/3313831.3376807

[122] Doris Xin, Litian Ma, Jialin Liu, Stephen Macke, Shuchen Song, and Aditya Parameswaran. 2018. Accelerating Human-in-the-Loop Machine Learning: Challenges and Opportunities. In *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning* (Houston, TX, USA) *(DEEM'18)*. Association for Computing Machinery, New York, NY, USA, Article 9, 4 pages. https://doi.org/10.1145/3209889.3209897

[123] Soner Yigit and Mehmet Mendes. 2018. Which Effect Size Measure is Appropriate for One-Way and Two-Way ANOVA Models? : A Monte Carlo Simulation Study. *REVSTAT-Statistical Journal* 16, 3 (Jul. 2018), 295â€“313. https://doi.org/10.57805/revstat.v16i3.244

[124] Alexey Zagalsky, Dov Te'eni, Inbal Yahav, David G. Schwartz, Gahl Silverman, Daniel Cohen, Yossi Mann, and Dafna Lewinsky. 2021. The Design of Reciprocal Learning Between Human and Artificial Intelligence. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 443 (oct 2021), 36 pages. https://doi.org/10.1145/3479587

[125] Sojib Bin Zaman, Nisal De Silva, Tian Yu Goh, Roger G Evans, Rajkumari Singh, Rajesh Singh, Akash Singh, Parul Singh, and Amanda G Thrift. 2023. Design and development of a clinical decision support system for community health workers to support early detection and management of non-communicable disease. *BMJ Innovations* 9, 1 (2023), 49–56. https://doi.org/10.1136/bmjinnov-2022-000952

[126] Marco Zappatore, Antonella Longo, Angelo Martella, Beniamino Di Martino, Antonio Esposito, and Serena Angela Gracco. 2023. Semantic models for IoT sensing to infer environmentâ€“wellness relationships. *Future Generation Computer Systems* 140 (2023), 1–17. https://doi.org/10.1016/j.future.2022.10.005

[127] Qiaoning Zhang, Matthew L Lee, and Scott Carter. 2022. You Complete Me: Human-AI Teams and Complementary Expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 114, 28 pages. https://doi.org/10.1145/3491102.3517791

[128] Ziliang Zhong, Muhang Zheng, Huafeng Mai, Jianan Zhao, and Xinyi Liu. 2020. Cancer image classification based on DenseNet model. *Journal of Physics: Conference Series* 1651, 1 (nov 2020), 012143. https://doi.org/10.1088/1742-6596/1651/1/012143

[129] John Zimmerman and Jodi Forlizzi. 2014. *Research Through Design in HCI*. Springer New York, New York, NY, 167–189. https://doi.org/10.1007/978-1-4939-0378-8_8

[130] John Zimmerman, Aaron Steinfeld, Anthony Tomasic, and Oscar J. Romero. 2022. Recentering Reframing as an RtD Contribution: The Case of Pivoting from Accessible Web Tables to a Conversational Internet. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 541, 14 pages. https://doi.org/10.1145/3491102.3517789

A ACCESSORY INFORMATION

In this appendix, we provide additional details on the use of AI models in our UI, as well as the severity classification during patient diagnosis, our patient selection, and information about participants. We also discuss the existing system, evaluating performance recognition, thresholds, and strategies for curating patients, the next steps for explanations and tone, and the repositories. As follows, we will provide further information to better understand the details of our work.

A.1 Severity Classification

BIRADS stands for "*Breast Imaging Reporting and Data System*" and is a system used to standardize the way in which radiologists report the findings of mammograms and other imaging exams of the breast [71, 105]. The BIRADS provides a standardized method for reporting the results of breast imaging exams, which can help to ensure that the information is accurate and consistent. This information is useful as an input for AI models that are designed to assist with the diagnosis of breast cancer, as it provides a standardized way of representing the findings of imaging exams [69].

By using the BIRADS system as an input for AI models, it may be possible to improve the accuracy and reliability of the model's predictions, and to help prevent bias in the results. The BIRADS system uses a scale from **0** to **6** for categorizing the findings of breast imaging exams. However, in our study we just considered the scale from **1** to **5**, as the **0** means that the case is inconclusive, where we need to acquire more images, and **6** means we already have biopsy confirmation by previously known lesion.

Here is a brief overview of each category on the BIRADS scale:

(1) **Negative:** The exam did not show any abnormalities and the patient's breast tissue appears normal.
(2) **Benign Finding:** The exam showed a benign (non-cancerous) abnormality in the breast tissue.
(3) **Probably Benign:** The exam showed an abnormality that is likely to be benign, but further testing may be needed to confirm this.
(4) **Suspicious Abnormality:** The exam showed an abnormality that is suspicious for cancer and further testing, such as a biopsy, is needed to determine if it is cancerous.
(5) **Highly Suggestive of Cancer:** The exam showed an abnormality that is highly suggestive of cancer, and a biopsy is recommended to confirm the diagnosis.

It is important to note that the BIRADS system is only a tool for reporting the results of breast imaging exams and does not provide a definitive diagnosis of cancer. A biopsy is usually needed to confirm a cancer diagnosis.

A.2 Patient Selection

In this paper, we used a total of 338 cases and acquired in the HFF clinical institution. From this set of 338 cases, 289 were classified by the head of radiology. Each patient has several images concerning four X-ray MG modalities (two in CC and two MLO views), one or two US images, and roughly 5 volumes in MRI. In the MRI volumes, we take numerous image slices per patient, where the lesion is present.

From the 289 classified cases, we selected a total of 35 patients to be classified by our AI models. Because we aim to test the three trials (*i.e.*, conventional *vs.* non-assertive *vs.* assertive), plus the two groups of medical professional experience (*i.e.*, novice *vs.* expert) we computed at least $2^5 = 32$ the number of patients. Hence, the 35 patients were selected to cover that magnitude of patients. This classification corresponds to assigning a BIRADS value for each modality image of the exam.

A.3 Participants Information

In this study, we collected information about the participants through an initial survey, and this included details about their gender, age, geographic location, and professional experience (Table 3). Additionally, we also ask participants about their professional background, in reading medical imaging data. Regarding the professional background, 11.54% of participants are doing their medical internships, 3.85% were breast medical surgeons, but with knowledge of reading medical images, and 84.61% were medical radiologists, reading and diagnosing patients every day.

Demographic	Group	Frequency	Percentage
Gender	Female	36	69.23%
	Male	16	30.76%
Age	18 - 29	11	21.15%
	30 - 39	12	23.08%
	40 - 49	10	19.23%
	50 - 59	10	19.23%
	>59	9	17.31%
Geographic Location	Hospital 1	15	28.85%
	Hospital 2	12	23.08%
	Hospital 3	2	3.85%
	Hospital 4	9	17.31%
	Hospital 5	1	1.91%
	Hospital 6	1	1.91%
	Hospital 7	6	11.54%
	Hospital 8	1	1.91%
	Hospital 9	1	1.91%
	Hospital 10	1	1.91%
	Hospital 11	3	5.77%
Medical Experience	Interns	6	11.54%
	Juniors	17	32.69%
	Middles	11	21.15%
	Seniors	18	34.62%

Table 3: Characteristics of participants for demographic groups with frequency and percentage. The main characteristics are gender, age, geographic location (clinical institution), and medical experience.

A.4 Existing System

Our new approach allows for a more flexible and dynamic system. Furthermore, the new system addresses some limitations of the traditional system [26], providing a more robust and scalable solution. Ultimately, offering a new and innovative approach for solving the diagnostic task.

The AI models in this study are not fusing the predictions from different imaging modalities (*e.g.*, MG, US, or MRI). Instead, each modality had its own ground-truth score, which refers to the correct or known diagnosis for a given patient. This is because the different modalities may provide different information about a patient's breast tissue, and may, therefore; result in different diagnoses and BIRADS scores for a given patient. Hence, the AI models in this study generated individual final predictions for each modality.

Specifically, the DenseNet model [49] was used to estimate the lesion score for 2D imaging data, such as MG and US images. The 3D ResNet model [5] was used to estimate the lesion score for 3D data, such as MRI volumes. Lesion score refers to the likelihood that a specific area of the breast tissue is cancerous, and is typically used as part of the BIRADS score for reporting the results of exams.

A.5 Evaluating Performance Recognition

We have used the false-positive and false-negative metrics for evaluating the performance of recognition of clinicians, since these metrics are straightforwardly obtained from a classification process. We chose to use these metrics because they are widely recognized as important indicators of performance in medical imaging classification tasks. Particularly, in the context of breast cancer diagnosis, where false-positives and false-negatives can have significant consequences for patient care. Furthermore, we believe that these metrics provide a more balanced and comprehensive evaluation of performance than classification accuracy, which can be misleading in imbalanced datasets. For instance, if a clinician provides a BIRADS of 3 but the real BIRADS is a 5, we consider it as a false-negative result. On the other hand, if the real BIRADS is a 2, but the clinician provides a BIRADS of 4, we consider it as a false-positive. Where the "real" score is the ground-truth provided by the expert

Overall, our goal was to evaluate the performance of our system in terms of its ability to reduce false-positives and false-negatives. We found that our classifiers achieved an average decrease of about 26% for the false-positive rate and about 2% for the false-negative rate, outperforming previous approaches that have been proposed for this task [26]. Moreover, we believe that the false-positive and false-negative metrics we used are appropriate for this purpose. By using these metrics, we were able to demonstrate the potential of our AI-assisted approach for reducing false-positives and false-negatives, and we believe that our findings could help to

A.6 Thresholds & Strategies for Curating Patients

Similar to what was already described (Section A.1), the rationale is the following. The BIRADS score ranges from 0-to-6 scale, with the following meaning: 0 – inconclusive, 1 – no findings, 2 – benign findings, 3 – probably benign, 4 – suspicious findings, 5 – high suspicious malignancy, 6 – previously known lesion. BIRADS of 0 and 6 are ignored in our study, because they are meaningless for prediction purposes.

We have clustered the values above into three classes as follows:

- "No Findings" with BIRADS = 1
- "Benign, probably benign findings" with BIRADS = {2, 3}
- "Probably, highly suspicious malignancy findings" with BIRADS = {4, 5}

Thus, the DenseNet (for 2D MG and US) and the ResNet (for 3D MRI), three classes for the classification. We take the values from 1-to-5, since the other two values do not count for the diagnosis. Notice that both networks are trained with the BIRADS ground truth provided by a radiologist from the HFF clinical institution.

A.7 Next Steps for Explanations and Tone

In this paper, we resort to two main classes of tones: (1) assertive, by having a more authoritative tone, while imposing the AI recommendations; and (2) non-assertive, while being a more suggestive agent. However, and considering the clinical context, this should be expanded not only to test more trials in a near future, but also to a larger extent of the communication tones. Concretely, the explanations should be attached to the concept of the lexicon for each breast modalities [105]. For instance, having the explanation: "scattered areas" (`lex_1`), "fibrogandular density" (`lex_2`) or "scattered fibroglandular tissue" (`lex_3`), and "ring enhancement mass" (`lex_4`). However, we recognize that there may be other communication tones that could be useful in the clinical context, and we plan to explore these in future research.

To study the full effects of the style of tone, we will need the following trials:

(1) Conventional Agent;
(2) Agent with Explanations in Neutral Tone;
(3) Agent with Explanations in Non-Assertive Tone; and
(4) Agent with Explanations in Assertive Tone.

This is an interesting point that presently we are pursuing our research in this direction. By providing more detailed and accurate explanations, we believe that our AI-assisted system can improve the diagnostic performance of medical imaging classification in the clinical domain of breast cancer.

B REPOSITORIES

Our repositories are accessible to the public and can be easily located online. Please follow the `prototype-assertive-reactive` repository (github.com/MIMBCD-UI/prototype-assertive-reactive) for more details about the source code of the prototypes. In terms of results and statistical analysis, all information is available in the `sa-uta11-results` repository (github.com/MIMBCD-UI/sa-uta11-results). The repositories have the linking pointers for the other related repositories, such as the datasets, source code of the AI models, prototypes, documentation, between others. These links were accessed on 25th of January 2023.

C INTELLECTUAL PROPERTY

The work described in this paper is covered by pending patent applications [21], filed by Instituto Superior Técnico. The contents of this paper are intended to be informative to the scientific and technical community. They are not intended to be used to limit the scope of the pending patent application. The patent rights will be enforced to the extent necessary to protect the proprietary interests of the patent holders. For more information, further details are available in the `LICENSE.md` file of the `sa-uta11-results` repository (github.com/MIMBCD-UI/sa-uta11-results). Accessed on 25th of January 2023.