Int. J. Human–Computer Studies 150 (2021) 102607

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

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

# Introduction of human-centric AI assistant to aid radiologists for multimodal breast image classification

Francisco Maria Calisto [*,a], Carlos Santiago [a], Nuno Nunes [b], Jacinto C. Nascimento [a]

[a] Institute for Systems and Robotics, Avenida Rovisco Pais 1, Lisbon, 1049-001, Portugal
[b] Interactive Technologies Institute, Caminho da Penteada, Funchal, 9020-105, Madeira, Portugal

ARTICLE INFO

*Keywords:*
Human-computer interaction
Artificial intelligence
Healthcare
Medical imaging
Breast cancer

ABSTRACT

In this research, we take an HCI perspective on the opportunities provided by *AI techniques* in medical imaging, focusing on workflow efficiency and quality, preventing errors and variability of diagnosis in Breast Cancer. Starting from a holistic understanding of the clinical context, we developed *BreastScreening* to support *Multimodality* and integrate *AI techniques* (using a deep neural network to support automatic and reliable classification) in the medical diagnosis workflow. This was assessed by using a significant number of clinical settings and radiologists. Here we present: i) user study findings of 45 physicians comprising nine clinical institutions; ii) list of design recommendations for visualization to support breast screening *radiomics*; iii) evaluation results of a proof-of-concept *BreastScreening* prototype for two conditions *Current* (without AI assistant) and *AI-Assisted*; and iv) evidence from the impact of a *Multimodality* and *AI-Assisted* strategy in diagnosing and severity classification of lesions. The above strategies will allow us to conclude about the *behaviour* of clinicians when an AI module is present in a diagnostic system. This *behaviour* will have a direct impact in the clinicians workflow that is thoroughly addressed herein. Our results show a high level of acceptance of *AI techniques* from radiologists and point to a significant reduction of cognitive workload and improvement in diagnosis execution.

## 1. Introduction

Artificial Intelligence (AI) has the potential to fundamentally transform many application domains (Ghahramani, 2015). One notable example is clinical radiology (Choy et al., 2018; Hosny et al., 2018), for which a growing number of AI-based systems have been proposed for lesion detection and segmentation, two fundamental steps to accomplish the diagnosis and treatment planning (Graffy et al., 2019; Kooi et al., 2017; Lakhani and Sundaram, 2017; Liang et al., 2019). The AI application within radiology can provide an extraction from large number of medical imaging features using data-characterization. This synergy between AI and medical imaging is currently known as *radiomics* (Lambin et al., 2012) and aims to develop methods that automatically analyze large amounts of data and extract meaningful features to support diagnosis (Ruddle et al., 2016) and clinical decision making (Park et al., 2015).

Many studies have reported robust and relevant findings (Aerts, 2017) in *radiomics*. Specifically, *radiomics* can provide detailed quantifications of medical imaging characteristics of underlying tissues. Without hand-designed features (Ker et al., 2018), the field is including significant hierarchical relationships within the data automatically discovered. This information can be used throughout the clinical care path to improve diagnosis and treatment planning, as well as assess treatment response. However, a growing number of these studies also suffer from deficient experimental or analytic designs (Aerts, 2017). Thus, failing to include a more holistic understanding of the clinical context (Sultanum et al., 2018).

While AI is changing the paradigm both in terms of visualization and interaction in the *radiomics* workflow, it is still not clear neither how it influences radiologists, nor what is the best approach to maximize the benefits of this Human-AI collaboration (Kocielnik et al., 2019). Indeed, existing AI approaches typically assume that there is a single correct answer for any given input, lacking mechanisms to incorporate diverse human perspectives. This assumption is prevalent in various steps of the AI pipeline, including model development and system design. The goal of this work is to study the impact of AI assistance on the diagnostic

* Corresponding author.
*E-mail address:* francisco.calisto@tecnico.ulisboa.pt (F.M. Calisto).

https://doi.org/10.1016/j.ijhcs.2021.102607
Received 13 August 2020; Received in revised form 23 January 2021; Accepted 3 February 2021
Available online 9 February 2021

performance of radiologists across several[1] clinical institutions.

We focus on the breast screening problem, in which radiologists typically require *Multimodality*[2] (MG - MammoGraphy, US - UltraSound, and MRI - Magnetic Resonance Imaging) to detect suspicious regions (lesion) in the breast. This multi-modal setup is very challenging, not only due to the large amount of data being processed, but also for manipulation and visualization of heterogeneous data. To this end, we developed the *BreastScreening* - a multi-modal *AI-Assisted* medical imaging framework that allows the radiologist to view, manipulate and classify images in an effortless way, while also providing diagnostic recommendations.

### 1.1. Motivation

Breast cancer is one of the most diagnosed cancers worldwide (DeSantis et al., 2016; Ferlay et al., 2013; Torre et al., 2015). In fact, breast cancer is the second leading cause of death from cancer in women (Bray et al., 2018). Early detection is one of the most effective ways to increase the survival rate for this disease (Saadatmand et al., 2015; Welch et al., 2016).

Screening mammography aims to identify breast cancer at earlier stages of the disease, when treatment can be more successful (McKinney et al., 2020). However, two main difficulties arise. First, the amount of data to be processed has been increasing significantly and greatly surpasses the throughput capabilities of the radiologists. Second, processing such amount of multi-modal data in timely fashion without compromising the reliability of the diagnosis which is very challenging. These difficulties have been the driving force behind the recent trend in *radiomics* and the integration of AI techniques into medical imaging.

In the context of breast cancer, the requirements for multi-modal data have a significant impact on the clinical workflow. Although MG is the primary imaging modality for breast screening, it may be insufficient to reach a correct diagnosis. For instance, in dense breasts (Fig. 1**B, D**), the lesions can hardly be seen, while in adipose breasts, lesion visualization is clear (Fig. 1**A, C**). Thus, for dense breasts, other modalities constitute a valuable information to complement the diagnosis. Fig. 1**E, F** depicts the US and Dynamic Contrast Enhanced (DCE)-MRI modalities, where the lesions can be easily viewed.

Support for visualization of multi-modal images and *AI techniques* can provide improvements and insights in the breast screening radiology workflow. Our goal is to evaluate the impact of the integration of these techniques in the context of the breast cancer diagnosis using the *BreastScreening* tool. Specifically, we focus on how *Multimodality* and *AI assistance* could add a value in the *radiomics* medical workflow (Calisto, 2017). Our answers to this problem include usage and acceptance by physicians and the improvement of workflow efficiency and quality, as well as reduction and prevention of errors and variability of diagnosis.

### 1.2. Research goals

A vital component of this research was to access a significant number of clinical settings and radiologists. We have established the foundations of our research via a human-centered design process and following the guidelines for Human-AI interaction (Amershi et al., 2019; Cai et al., 2019b; Kocielnik et al., 2019), including: i) findings from a user study in nine health institutions, encompassing *in-situ* observations and interviews (Lim et al., 2019; Sarcevic et al., 2012), and grounded by related work; which informed ii) a list of design recommendations for medical imaging design, including temporal awareness, image processing, *Multimodality*, adoption, usage and trust; leading to iii) findings from an evaluation study of *BreastScreening*, a proof-of-concept prototype we developed to support the clinical translation of *radiomics*, validated by 45 physicians; and finally iv) evidence from the impact of a *Multimodality* and *AI-assisted* strategy in diagnosing and severity classification of lesions.

In particular, focusing on the integration of *Multimodality* and *AI techniques* in *BreastScreening* clinical workflow, we formulate the following three high-level questions:

1. **RQ1.** What is the workflow impact of *AI assistance* for avoiding different types of diagnostics?
2. **RQ2.** What are the design techniques for setting appropriate clinician expectations of *AI assistance* and how to improve diagnostic interpretability?
3. **RQ3.** What is the impact of expectation-setting intervention techniques on satisfaction and acceptance of *AI assistance* in radiology?

Diagnosis can be defined as mapping between visualization of the image and the respective scalar BI-RADS (that should be as accurate as possible), which reflects the severity of the examination diagnosis. The above research questions are placed in a context to encompass the richness of each modality, especially in the *Current* vs. *AI-Assisted* medical imaging scenarios discussed in the evaluation section.

## 2. Background

Two areas motivate this research: a) *radiomics*; and b) Human-AI interaction. In the following sections, we address each one.

### 2.1. Radiomics

In modern healthcare, medical images are crucial to support decision making, both in terms of diagnosis, predictions, and treatment planning (Calisto et al., 2019a). In many situations, the accurate lesion detection and segmentation constitutes a mandatory requirement for image analysis. Over the last decades, several methods have been proposed to automatically perform these two tasks (Litjens et al., 2017).

One way to perform the detection and segmentation is based on the use of deep neural networks, depending on the amount of data used for training. In a supervised (or semi-supervised) setting the training data should be annotated. However this requirement is hard to be accomplished, since the annotations process is usually costly (Calisto, 2020a;

[1] Number of clinicians and institutions that supported our project: 12 clinicians of Hospital Professor Doutor Fernando Fonseca (HFF) ; 10 clinicians of Instituto Português de Oncologia (IPO) de Lisboa (IPO-Lisboa) ; 2 clinicians of Hospital de Santa Maria (HSM) ; 9 clinicians of IPO de Coimbra (IPO-Coimbra) ; 1 clinician of Madeira Medical Center (MMC) ; and 1 clinician of Servios de Assistência Médico-Social do Sindicato dos Bancários do Sul e Ilhas (SAMS) ; 8 clinicians of Hospital do Barreiro (HB) ; 1 clinician of Hospital de Santo António (HSA) ; 1 clinician of João Carlos Costa Diagnostic Imaging (JCC). We collected the patient data from the first institution (*i.e.*, HFF). Then, we applied our User Testing and Analysis (UTA) to the full list of clinical institutions and clinicians.

[2] *Multimodality*: our proposal diagnostic technique for the patient treatment via (1) MammoGraphy (MG); (2) UltraSound (US); (3) Magnetic Resonance Imaging (MRI); (4) text; and (5) annotations. MG modality is a specific type of breast imaging that uses low-dose X-Rays to provide earlier cancer detection. In MG, two views are taken, concretely CranioCaudal (CC) view and MedioLateral Oblique (MLO) view. US modality is an imaging test that sends high-frequency sound waves through your breast and converts them into images on a viewing screen. MRI uses radio waves and strong magnets to make detailed pictures of the inside of the breast. Both text and annotations are inputted by the user. The text represents the *dataset* of co-variables (mimbcd-ui.github.io/dataset-uta7-co-variables) that each clinician can input in our system. The annotations are the groundtruth (mimbcd-ui.github.io/dataset-uta7-annotations) of the lesions that clinicians label on the image, so that the algorithms can learn with clinicians. A sample *dataset* of the used medical images (mimbcd-ui.github.io/dataset-uta7-dicom) is also provided. Furtheremore, we provide a link (mida-project.github.io/prototype-multi-modality-assistant) for the demo of the UTA7 main scenario. For the source code of the prototype, another link (github.com/mida-project/prototype-multi-modality-assistant) is provided with instructions and documentation.

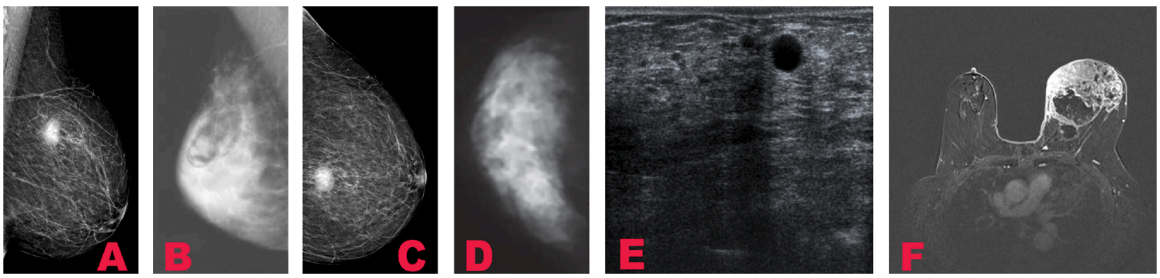


**Fig. 1.** Illustration of the current clinical setup. Adipose breast in mammography where the lesions are easily viewed (A, C). Dense breast where is not possible to view the lesions (B, D). In the latter scenario, the radiologists resort to other image modalities such as US (E) and DCE-MRI (F), to complement the information that is missing in (B,D).

2020b; Calisto et al., 2020). Thus, one has to resort to small datasets. To ensure that the networks' performance is not hampered by the lack of training images, transfer learning (Raghu et al., 2019) is typically used, in which the network is first trained with a large unrelated data set and is then fine-tuned for the target task (Shin et al., 2016).

Recently, Convolutional Neural Networks (CNN) were shown to be effective at diagnosing breast cancer (Carneiro et al., 2017). However, this and many other approaches (Becker et al., 2017; Khan et al., 2019; Wang et al., 2016) are limited to the MG modality. Few works have taken advantage of multi-modality to further enhance the performance of these methods (Murtaza et al., 2019). Specifically in recent years, deep learning (DL) approaches have contributed to the notoriety of *AI-Assisted* healthcare (Topol, 2019). These methods have been outperforming older approaches and setting new State-Of-The-Art results, with some achieving or even surpassing human-level performances (Esteva et al., 2017; Gale et al., 2017).

The success of DL methods is due to the ability of the CNN (Greenspan et al., 2016) to extract meaningful features, obtained using large training datasets. These networks learn to extract and analyze large numbers of image-based features (*e.g.*, through *radiomics*) for quantitative characterization and analysis of tumor phenotype (Kumar et al., 2017). Several deep neural architectures have been proposed, *e.g.*, DenseNet (Huang et al., 2017) for classification (diagnosis) problems, or U-Net (Ronneberger et al., 2015) for segmentation problems.

Despite the success of DL, and even though different studies have shown that *radiomics* (*i.e., AI-Assisted* methods for medical image analysis) can reduce human error and improve outcomes (Cai et al., 2019b; Delvaux et al., 2017; Middleton et al., 2016), their adoption by the medical community has been slow. One of the main reasons is the inability of these systems to provide relevant medical information or to capture the nuances of the human mind (Khairat et al., 2018; Kohli and Jha, 2018; Yang et al., 2016), making them untrustworthy and preventing their clinical acceptance. In particular, DL-based methods have been frequently viewed as black box approaches (Litjens et al., 2017). Therefore, HCIs play an important role in creating user-friendly interactive systems for *AI-Assisted* medical image analysis (Calisto, 2019b; Calisto et al., 2017; 2019b).

### 2.2. Human-AI interaction

Applications of Human-AI collaboration in complex domains are subject to the following two issues: (1) trust, transparency and accountability of the involved AI agent (Amershi et al., 2019); and (2) user's ability to understand and predict agent behavior, *i.e.*, explainability and intelligibility (Cai et al., 2019a; Gunning, 2017; Holzinger et al., 2018; Miller, 2018). Forming accurate mental models of the *AI-Assisted* is useful for: (i) representing the clinician's belief about what the system can do, acquired via interviews and observations, instruction, or inference; (ii) mapping between the observable features of our system and the functionality perceived by the user; and (iii) the prediction for anticipating the AI output in a given scenario.

Whilst eXplainable AI (XAI) (Samek et al., 2019) deals with the implementation of transparency and traceability of statistical 'blackbox' machine learning methods. Particularly, deep learning approaches are in certain domains, a pressing need to go beyond XAI; for example, to reach a level of explainable medicine there is a crucial need for causability. However, causability (Holzinger et al., 2019) is different from causality. In the same way that usability encompasses measurements for the quality of use, causability encompasses measurements for the quality of explanations produced by XAI methods (*e.g.*, the heatmaps of Section 5). Specifically, causability is the property of a human (*i.e.*, human intelligence), whereas explainability is a property of a system (*i.e.*, artificial intelligence).

In the medical domain it is of supreme importance to enable a domain expert to understand (Yang, 2019), 'why' an algorithm came up with a certain result (*e.g.*, this is necessary due to raising legal issues). With certain XAI methods, such as layer-wise relevance propagation, relevant parts of inputs to, and representations in, a neural network which caused a result, can be highlighted (with a heatmap). However, this is only a first - important - but only first step to ensure that end users, *e.g.*, medical professionals (humans), assume responsibility for decision making with AI. The backbone for this approach is interactive ML (Holzinger, 2016), which adds the component of human expertise to AI/ML processes by enabling them to re-enact and retrace the results on demand, *e.g.*, let them check it for plausibility. This requires new human-AI interfaces for XAI and in order to do so, one has to deal with the question of how to evaluate the quality of explanations given by a XAI system, for this we need measurements, *e.g.*, the recently developed System Causability Scale - to measure the quality of explanations (Holzinger et al., 2020).

In this context, expectation theories (Kocielnik et al., 2019; Leung and Chen, 2019) are postulating that user satisfaction and acceptance of a system is directly related to the difference between initial expectations and the actual clinical experience. Specifically, expecting more than the system can deliver will decrease user satisfaction and lead to the rejection of the system. Hence, we created our proposed technique, based on the following contributions: (1) providing users a new control feature on the introduction of AI methods among medical imaging diagnosis; and (2) the impact of the radiologists *behaviour* and the impact in professional practice. We did that to achieve more accurate expectations of the systems capabilities addressing potential gaps.

## 3. Design keys

Patient classification and providing accurately that information to clinicians, as well as resolving the queries coming from the AI outputs

are the main design keys for our solution. In order to cope with the remaining challenges, our human-centered approach revealed to be useful into understanding the medical workflow to properly address the clinician's needs. Specifically, the process that leads to the development of the *BreastScreening-AI* prototype (Calisto et al., 2018a) included a holistic understanding across the clinical context of *radiomics* in breast screening. The simulated AI assistant instantiates the presented design keys by providing human-interpretable rationals for its outputs. We were especially interested in the use of several modalities and *AI assistance* to detect and classify lesions. We conducted quantitative and qualitative studies in nine health institutions to understand the medical practices surrounding *radiomics* in breast cancer, including the classifications of the lesion severity using the BI-RADS (Calisto and Nascimento, 2018) score. Next, we will describe the workflow practices that lead to design implications of our assistant, as well as design goals and design methods.

### 3.1. Radiology room

The main radiology room workflow (Fig. 2) can be defined as a three-stage process: (1) *examination*; (2) *diagnosis*; and (3) *report*. The *examination* stage refers to the time spent on the examination and processing of the patient records (*e.g.*, demographics, clinical records or past medical images). For instance, the radiologist receives this information from both the hospital and radiology information systems. The second stage (*diagnosis*), corresponds to the time that the clinician spends on the interpretation and diagnosis of the patient exams. In this phase, the clinician interacts with a Picture Archiving and Communication System (PACS) retrieving the *Modality Worklist* (Fig. 4) from the Radiology Information System. Finally, in the last stage it refers to the time spent on reporting her/his conclusions after the patient's medical images analysis. In all the above stages, medical images are used with different purposes, as described next.

For the *examination* stage, the clinician relates medical imaging analysis with other exams and medical records, *i.e.*, the clinical situation of the patient. The *examination* of the patient is sometimes supported by other systems that give the clinician a more reliable source of information. Most of the information sought is stored in various formats, including notes from the referring physician (the doctor who sent the patient to the radiology department), past exams and respective reports, and second opinion from other clinicians.

Regarding the *diagnosis*, this is the most crucial from our perspective, since it is the one that most contributes to the treatment choice. The *diagnosis* stage includes image classification on a BI-RADS scale. During this stage, we observed clinicians on the diagnosis process, namely accessing several medical images. We conducted several studies (Section 6) to understand what clinicians need during decision-making and while interacting with an *AI-Assisted* system. In this stage, each medical image takes about 40 s to be analyzed, meaning that the clinician takes about half a minute to interpret an image. However, in most of the cases, each patient requires more than 100 images to be analyzed (*e.g.*, MRI volumes). This means that a complete diagnosis in a real situation takes more than an hour of the clinician's time, making this task cumbersome in the overall time of the *diagnosis* and prone to errors. In our observations, we concluded that most of the clinicians disengage from visualizing all set of medical images, or disregard some pathologies in the images. These handicaps may cause human errors and medical distractions (Bruno et al., 2015).

After a patient classification, the clinician takes into consideration the first and the second stages (*i.e., examination* and *diagnosis*) to improve the patient's clinical records and determine the prognosis. During the classification phase, the clinician examines the patient records and, by analyzing each medical image, records the *report* in a dictating system. The records are transcribed to a report, which are reviewed later by clinicians (signing the report as final).

As previously discussed, using images to display breast lesions vary widely across medical imaging modalities (MG, US and DCE-MRI, being the latter two modalities crucial for dense breasts diagnosis). *Multimodality* is also responsible for increasing the cognitive load of radiologists and increasing detection rates but also False-Positives (Cheung, 2017). It is, therefore, a central issue in *BreastScreening* (Calisto et al., 2018b; Wernli et al., 2019). In the next section, we will describe the relation between workflow stages of medical imaging and the clinical procedures.

### 3.2. Clinicians procedures

In this paper, we provide a modular guideline for study planning, design and implementation within the area of HCI research in medical imaging. Furthermore, our research work allows to acquire the structured description of the clinical workflow within the breast cancer domain and to consider this information for own study design or to perform a comparison of different studies. The investigation model and the corresponding clinical procedures can be used further to create new knowledge bases of HCI assessment in medical imaging.

The main workflow of the radiology room (Wagner et al., 2015) (Fig. 2) usually comprises four different paths that correspond to different observed image acquisition procedures:

1. **Procedure 1** starts with the acquisition of MG, then, if the breast is dense, the US modality is acquired. Finally, if the MG and US are not conclusive, the DCE-MRI is acquired, otherwise the process is concluded;
2. **Procedure 2** starts with the acquisition of MG, then if the clinician detects a high risk of cancer from the image patterns and/or patient records, both US and DCE-MRI are acquired, otherwise the process is concluded;
3. **Procedure 3** MG and US are acquired simultaneously, if the *exam* (Wagner et al., 2015) is still not conclusive, the DCE-MRI is acquired;
4. **Procedure 4** all three modalities (MG, US and DCE-MRI) are acquired simultaneously.

From the interviews and observations (Section 6), we also found that clinicians access medical images in two main scenarios: i) imaging perception process, namely to detect patterns of lesions; and ii) finding relationships between past lesion patterns and possible future diagnosis. Given the time constraints and the amount of information available, clinicians often do not observe all the images with the necessary detail. From our observations and eye-tracking measurements (Section 6), they start by analyzing the patient's clinical history (when available). Clinical history provides the necessary knowledge on how to guide the analysis of the current state. At this point, we took impressions regarding the efficiency of clinicians, and their recommendations based on their experience for improvements of the patient *examination*. In fact, several studies demonstrated (Waite et al., 2017) that radiologist fatigue levels and performance are related to environmental factors such as number of False-Negatives and False-Positives. That said, we start analyzing the potential enhancement that an *AI-Assisted* diagnosis could take in the radiology room (Chatelain et al., 2018; Miglioretti et al., 2007).

### 3.3. Insights and challenges

Our observations and interviews are aligned with previous research on clinician-driven diagnostic *tasks* (Heinrich et al., 2012; Rosset et al., 2004; Sultanum et al., 2018; Weese and Lorenz, 2016; Wolf et al., 2005). From the research insights we identified the following main challenges: i) the heterogeneous visualization mode of a large number of images and file sizes; and ii) the annotation of medical images to support diagnosis and also how the introduction of the *AI techniques* can improve the classification ground truth for. Next, we detail each of the two points above mentioned.

*Radiomics* in general, and *BreastScreening* in particular, require

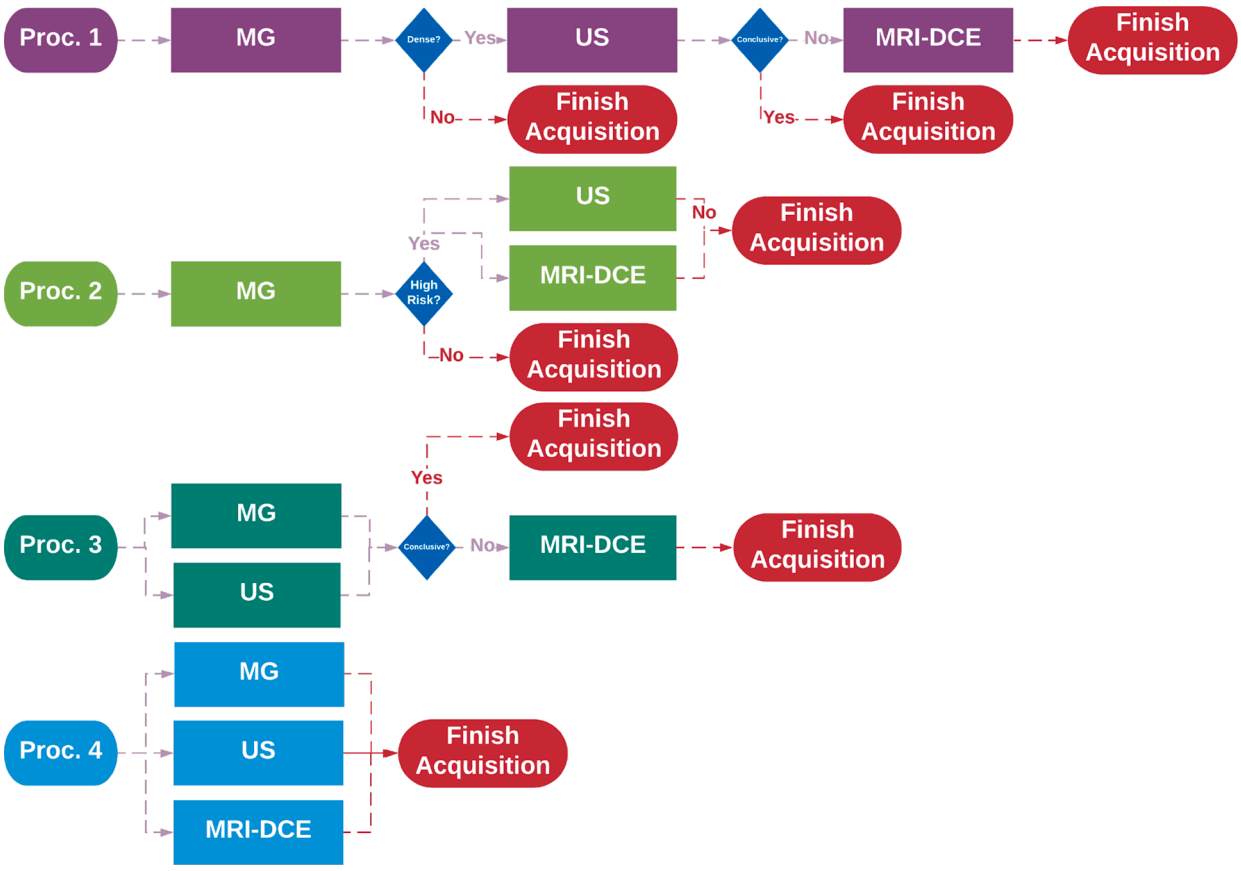


**Fig. 2.** Workflow of the radiology room is commonly adopted in current clinical institutions using several image acquisition strategies. Screening modalities (*e.g.*, MG, US, MRI, etc) constitute important complementary information for a reliable diagnosis. In this work, we intend to demonstrate how the multimodality is used in current clinical setups. The exposition described in **Proc. 1**, is the one followed by Hospital Professor Doutor Fernando Fonseca, a public hospital in Lisbon. Thus, to design a useful interface to a specific hospital institution, it should contain all the modalities involved in the diagnosis. This also supports the use of multimodality in the present study.

managing a significant and heterogeneous number of large image files. This is paramount in MRI volumes. During the MRI acquisition, tens of breast volumes are obtained, comprising different imaging volumes[3] some of them in time intervals (*e.g.*, T1, T2, diffusion, and Dynamic Contrast Enhanced with subtraction (Sorace et al., 2018)).

From the observations and interviews, it was clear that clinicians observe only a fraction of these MRI volumes. Also, the imaging volumes inspected are different depending on the practices of each clinical institution. For instance, we observed that in HFF public hospital only the DCE-MRI at second time instant is considered, while in IPO-Lisboa only the third time instant of imaging volume is used for diagnosis. Consequently, the User Interface (UI) should reflect the specific institution as it will be detailed in Section 6.2.4.

Related to the requirements of *radiomics*, the second challenge involves the generation of ground truth data. This is twofold, namely, (i) for visualization issues, when the radiologist is able to inspect the delineation provided, thus, facilitating an eventual second reading of the exam, also (ii) it constitutes valuable information for training deep neural nets with (semi)supervised learning procedures, as mentioned above (see Section 2.1). This comprises the localization/delineation of anatomical calcification and mass lesions.

In addition, it is easier to detect the lesion patterns by comparing the region of the breast in different modalities. This is particularly relevant in *BreastScreening* since the lesions in dense breasts are almost impossible to detect in MG (in both CC and MLO views) - a recognized problem leading to a large number of non-diagnosed cancers[4] which only manifest later in touch exams or after severe consequences from disease progression (Mohamed et al., 2018).

### 3.4. Design goals

In this section, we investigate how *AI-Assisted* methods could be integrated into the design of a medical imaging diagnostic assistant. Our purpose is to help mitigating breast cancer diagnosis, while meeting overall healthcare design goals. The main design goals are closely related to the research insights and the challenges of the previous section, namely: (1) a collection of a ground truth annotations, namely masses in all imaging modalities and calcification lesions in MG (for both CC and MLO views); (2) classification of the lesion severity using the BI-RADS (Aghaei et al., 2018); (3) categorization of the breast tissues (dense vs non-dense); (4) clinical co-variables, such as personal and family records; and (5) visualizations for clinical summary which is

[3] The MRI comprises different types of volumes, concretely, T1, T2, T2 Fat Sat, Diffusion, DCE-MRI, DCE-MRI with subtraction in five time instants. The sensitivity of MRI makes it an excellent tool in specific clinical situations. Situations such as the screening of patients at high risk, and evaluation of the extent of disease in patients with a new diagnosis. Compared to MG and US, MRI provides higher sensitivity, however its specificity is variable. Moreover, MRI data analysis is time consuming and depends on reader expertise.

[4] The biopsy is the only diagnostic procedure that can definitely determine if a suspicious image area is cancer. For instance, a BI-RADS score of 4 means the patient needs a biopsy. However, there is just 30% of chance of having malign cancer, in other words, a 70% chance of a benign final result.

crucial for a proper diagnosis and to perform patient follow-up.

We fuse these five insights into three corresponding design goals, as follows:

[Medical Imaging Design (MID)] focusing on how to provide the best visualization strategy, given the heterogeneous information coming from the multi-modal sources of information;
[Control Result Design (CRD)] focusing on improving the physician's ability to *accept* or *reject* the *AI-Assisted* results;
[Explanation Design (ED)] focusing on increasing physicians understanding of how the AI techniques operate. By increasing understanding of how AI works, physicians can update their expectations of how well and in which situations the system is likely to work;

Herein, we attempt to holistically integrate these design goals in the context of medical imaging diagnosis supported by *AI-Assisted* methods for the breast cancer domain. Through user studies, we identified the above three (*i.e.*, MID, CRD and ED) design goals. Next, we will describe the design methods to achieve these design goals.

### 3.5. Design methods

In this paper, we actively involved all clinicians in the design of this medical imaging solution. To generate clinician's empathy and involvement, design methods from participatory design were used (Wilde et al., 2017).

Our design methods consist of three aspects:

- *insight*;
- *ideation*; and
- *implementation*.

Interviews and observations are helpful to obtain a synthesized *insight* in the clinical workflow (Section 3.2). As design method according to this aspect, we went through several observations and interviews on clinical institutions. From these methods, we extracted information regarding, not only, workflow (Section 3.2), but also, demographic data (mimbcd-ui.github.io/dataset-uta7-demographics) of clinicians (Section 5.1).

For *ideation*, the process of generating new ideas, is central to design where the goal is to find novel solutions around a set of user needs and requirements. In terms of design methods, we promote several brainstorming techniques. Those techniques are such as workshops (Section 6.2.1), focus groups (Section 6.2.2) and affinity diagrams (Section 6.2.3). Affinity diagramming (Section 6.2.3) has been used in our study to organize the acquired large sets of ideas into data clusters (Section 6.2.4). In this paper, the methods are used to organize our findings and to sort design ideas into *ideation* of a focus group (Section 6.2.2) during several workshops (Section 6.2.1). The techniques will be further (Section 6.2) detailed and discussed.

Finally, the *implementation* is promoted as a way of developing the prototype (Section 6.2.5). We quickly recognized that successful *implementation* would rely on a bare minimum number of requirements. Short iterations enabled the use of many different design methods for prototyping and testing, as we have many different concerns with clinicians.

## 4. BreastScreening

To validate the proposed design goals and research questions, we developed *BreastScreening*, a proof-of-concept fully functional prototype to be evaluated in a realistic clinical scenario. In the following subsections we describe the main features of the *BreastScreening* system.

### 4.1. Implementation

*BreastScreening-AI* was implemented (Fig. 3) using *CornerstoneJS* (Urban et al., 2017) with a *NodeJS* server. To feed the system, we selected image patient sets from the HFF clinical institution and uploaded the images into an *Orthanc* server (Jodogne, 2018). Three imaging modalities (MG, US and MRI) were provided for each patient. The images were pre-processed and anonymized on the *Orthanc* server and then consumed by the *BreastScreening-AI* assistant.

This system is efficiently designed as a set of modules that can be reused in other imaging applications. The *CornerstoneJS* family of libraries provide essential functions, such as i) image rendering; ii) DICOM retrieval; iii) tool support (*i.e.*, developed functionalities); and iv) interpretation. To enable smooth drawing operations for manual labeling of the annotations, *CornerstoneJS* leverages canvas objects (Mullie and Afilalo, 2019), which enables to accelerate the rendering process of dimensional graphics in the browser.

All image data used in this process are stored and retrieved from this *CornerstoneJS* library. Moreover, *CornerstoneJS* is a web-based library with tool support for asynchronous execution, enabling the use of segmentation tools. Lastly, the library permits the interpretation of images rather than the recording of the imaging findings.

The *BreastScreening* core was developed in *JavaScript* with *jQuery* for *HTML* document manipulation, event handling and *dicomParser* for parsing DICOM files. The DICOM files can be loaded by drag-and-drop into the browser window on the *Orthanc* view. Loaded images can be further displayed on both views, but with different visualization configurations. After the loading stage, the images are automatically arranged according to the scan IDs from the DICOM files.

Once the set of medical images is loaded in the visualization viewport (Fig. 5) the user can interact with the data by manipulating the visualization through the mouse and keyboard. For instance, rolling the mouse wheel to navigate across the volume slices (MRI) or even moving the mouse to drag-and-drop the several modalities to the viewport. The goal of our multi-modal strategy is then to provide the visualization and manipulation of several modalities of image to compare lesion patterns among those images. This requires a novel mechanism for visualizing the three modalities of medical images (MG, US and MRI).

### 4.2. User interface

Based on identified user needs (Section 3), we designed and implemented the UI for *BreastScreening*. The UI includes a set of refinement mechanisms to guide clinicians during the diagnostic process. The UI of *BreastScreening* consists of two main components, comprising the list of patients (Fig. 4) and medical imaging views (Fig. 5). Fig. 4 can be framed in the following subcategories: *4.1. List of Patients; 4.2. Header; 4.3. Help; 4.4. About*; and *4.5. Study List Tabs*. Fig. 5 can also be subdivided in the following categories: *5.1. Viewports; 5.2. Toolbars*; and *5.3. Modality Selection*.

The first assistant view is the *4. List of Patients Views*. With this view, clinicians can quickly choose the respective *4.1. List of Patients*. Using the BI-RADS functionality (*6.3.1. Physician Severity*), clinicians can now classify (Fig. 5) the severity of the breast lesion for each patient. This *4.1. List of Patients* contains only the most important and required information, avoiding the excess of data (*e.g.*, gender, access activity on the file, etc), typically shown on these systems. Therefore, improving the medical imaging (*MID*) visualization across the radiology room. The *4.5. Study List Tabs* also allows clinicians to switch between diagnosed patients, *5. Medical Imaging Diagnosis Views*, and *4. List of Patients Views*.

Upon starting the UI in the web browser, the user can select from the *4.1. List of Patients* available in the study list. Hence, it will improve the design around medical imaging (*MID*) at a higher diagnostic accuracy level (Section 6) from a single clinician. From here, users can search all patients using criteria such as **Patient ID, Study Date**, available **Modality** sets, **Study Description**, and number of **Images**. By pressing in a

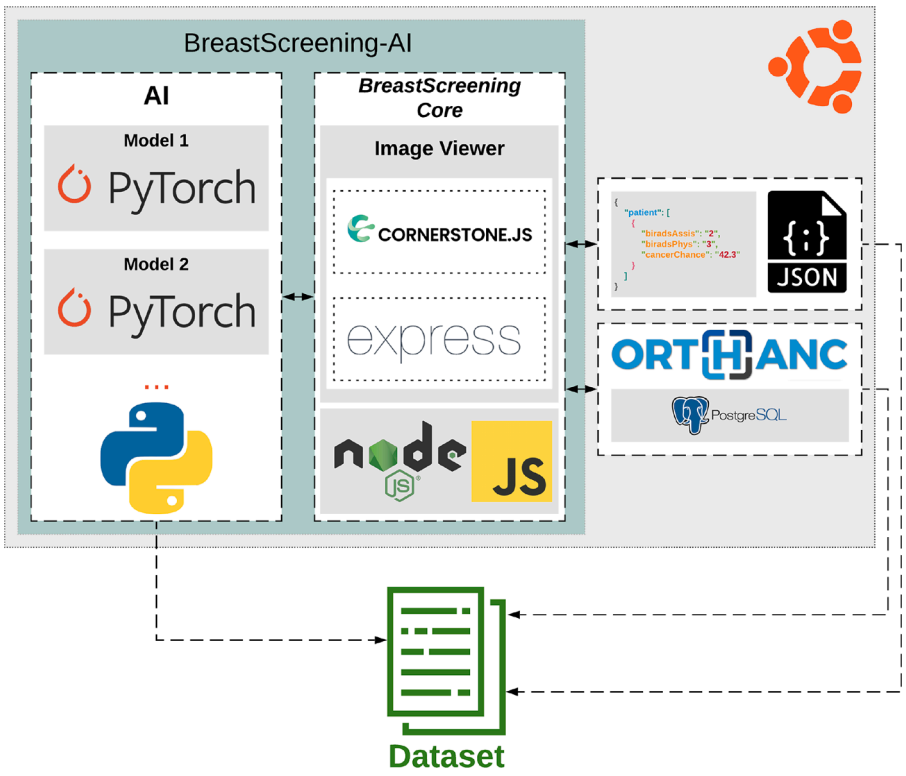


**Fig. 3.** *BreastScreening-AI* Architecture: the main components of the system are AI Application, Image Viewer, Datasets and DICOM Storage. The Image Viewer of *BreastScreening* framework will provide essential interaction tools for radiologists. A study list is fetched from the Orthanc server, and through CornerstoneJS the radiologist can manipulate the image, interacting with the assistant at the same time.

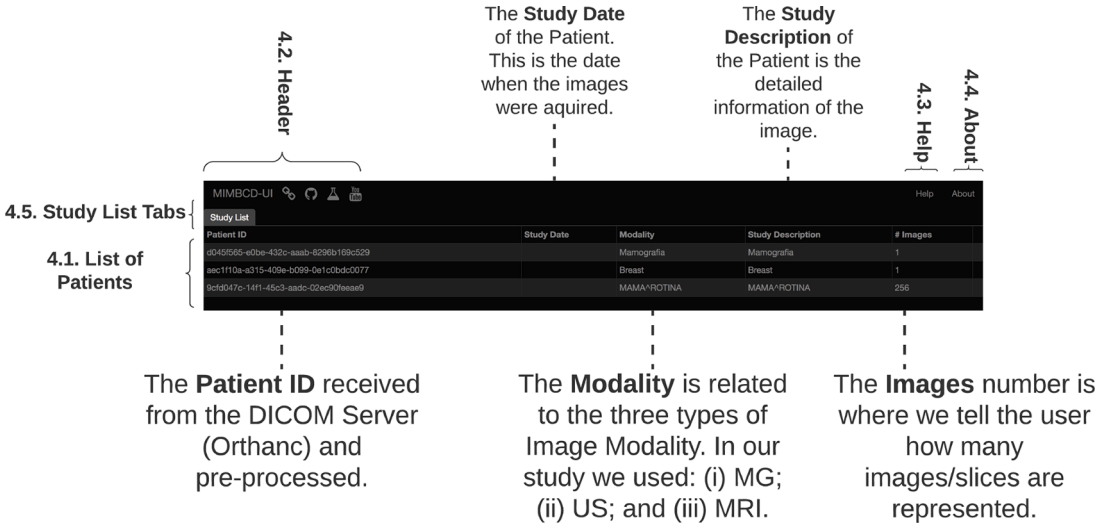


**Fig. 4.** *BreastScreening* provides clinician support to easily access and switch between the patient views. When accessing the assistant, the first screen shows the list of patients. Each row of the table, represents a patient. Each patient has a **Patient ID, Study Date, Modality** (one or more), **Study Description** and number of **Images**.

selected row, clinicians open the image viewers. In the image viewers, it is possible to visualize the image modalities.

To support clinicians, we provide both *4.3. Help* and *4.4. About* components. The components are providing clinicians information about how to classify and information about the system, as well as contact information. For instance, these components were used by Interns, to better understand how they could use the tool for diagnosis.

The *5.2. Toolbars* (Fig. 5) were positioned to match the users's requirements (conclusions drawn during our observations and interviews). This corresponds to the MID goal. The *5.1. Viewports* are displayed at the right, after the *5.2. Toolbars*.

Clinicians can probe for lesion patterns via the *5.1. Viewports* and process the image by using the *5.2. Toolbars* features (*MID*). Each time the *5.2. Toolbars* functionalities are activated, the clinician needs to perform a simple and easy interaction with the medical image to manipulate it as desired. Using the *5.2. Toolbars* on the *5.1. Viewports*, the clinician can locate the lesions and classify its severity (BI-RADS).

#### *4.2.1. Assistant*

The *6.1. Accept* or *6.3. Reject* allows the clinician to accept (or reject)

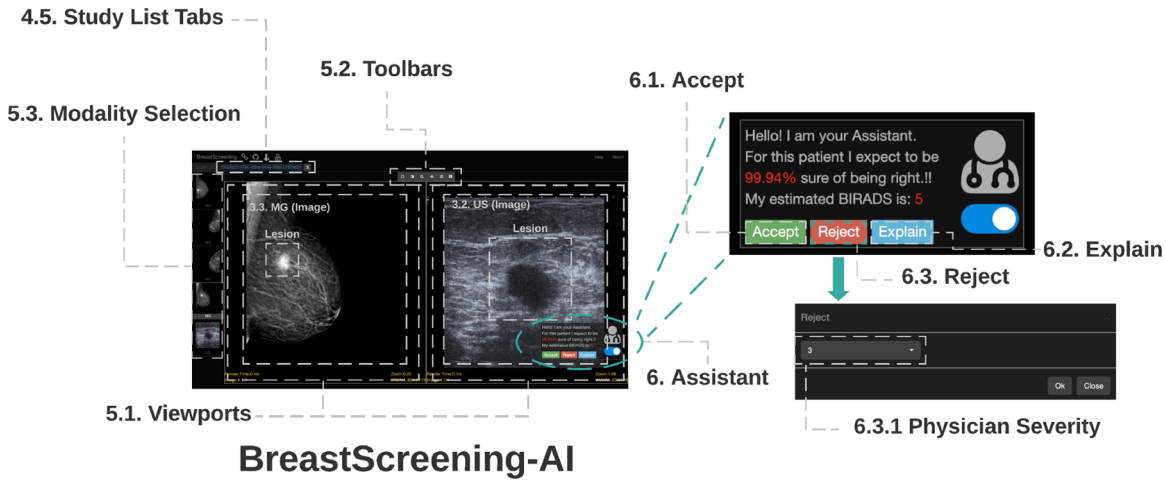


**Fig. 5.** Our *BreastScreening-AI* assistant provides several features regarding the basics of *Radiomics*. From there, we will be able to validate our DenseNet BI-RADS classifier along with clinicians.

the automatic classification of the *6. Assistant*. The *6. Assistant* is based on the use of a DenseNet (Huang et al., 2017), following several steps (Section A.1 of Appendix A) of the training set. However, integrating a deep neural network (DNN) needs special attention.

Typically, the training of a DNN is expensive regarding the time spent, since its classification performance will improve as the training dataset becomes larger. Thus, training the assistant from the scratch is not the best option. For this reason, we pre-train (*i.e.* off-line training before the UI integration) the DNN on ImageNet (Deng et al., 2009) dataset and fine-tuned using our multi-modal breast *dataset*. Specifically, we fine-tune the DNN using supervised learning. Only afterwards, the pre-trained DenseNet is incorporated into the UI.

Our DenseNet takes medical images as an input and outputs the severity probability. The input consists of images (*i.e.* MG, US and MRI slices) with the corresponding label (*i.e.*, the BI-RADS) that is previously classified by an expert. The output consists of having five nodes in the last layer of the DNN. Each node is assigned to a given class that corresponds to each BI-RADS score. After this stage, we have conditions for the integration, since now the DenseNet is tuned to perform the classification in unseen test images. The classification is fast, being tailored for an online diagnosis. When several modalities (*MID* and *CRD*) are correctly used (regarding the *5.3. Modality Selection* on a *Multimodality* view), the clinician can find more accurately the right severity classification, as concluded in Section 6. For the *6.3. Reject* option, the physician will have the opportunity to insert (*CRD*) the proposed BI-RADS on a drag-and-drop menu (*6.3.1. Physician Severity*) of severity options. Both *MID* and *CRD* goals are supporting our **RQ1** question.

#### 4.2.2. Explainability

Finally, the clinician may look for the *6.2. Explain* feature by looking at the generated heatmaps (Fig. 6). The generation of the above maps come from the following information: (i) the area (*Lesion Size*) of the lesion that comes from the delineation process, (ii) the circularity/sharpness (*Lesion Shape*) that can be computed from the annotation in (i), and (iii) the BI-RADS score automatically provided by the DenseNet (*i.e.*, AI assistant). The *ED* design goal is performed as above described helping to answer to research questions **RQ2** and **RQ3**.

## 5. Evaluation

We conducted an evaluation[5] of *BreastScreening* simulating real world conditions with 45 clinicians in nine different clinical institutions. Our goal was to quantitatively and qualitatively assess the proposed design principles that the *BreastScreening* embodies and to understand how these principles would work in practice. The experimental setup aimed at testing two conditions: **Cond. C1** - *Current, i.e.*, a *Multimodality* without any *AI-Assisted* technique; **Cond. C2** - *Multimodality* taking advantage of the *AI-Assisted* (*e.g.*, DenseNet Maicas et al., 2019), supporting clinician's second opinion and autonomous patient diagnostic.

For each condition we defined three classes of patients:

- **Class 1**: patients such that BI-RADS $\leq 1$ (low severity);
- **Class 2**: patients such that $1 <$ BI-RADS $\leq 3$ (medium severity);
- **Class 3**: patients such that BI-RADS $> 3$ (high severity);

The exams were previously annotated by eight radiologists and classified with a BI-RADS severity from an expert doctor. The expert is the head of the radiologist services of HFF clinical institution.

### 5.1. Participants

We asked participants to practice with three predefined patients selected from the three above classes. To accomplish this, we randomly select each patient (*i.e.*, **P1, P2** and **P3**) from each class (*i.e.*, **Class 1, Class 2** and **Class 3**), respectively. Then, we asked participants to diagnose each patient. A natural expectation is that the *AI-Assistant* would minimize the time required and accuracy (*e.g.*, improving False-Positive or False-Negative values).

Our study involved 45 clinicians (Table A.3 of Appendix A), recruited on a volunteer basis from a broad range of clinical scenarios, including nine different health institutions (two public hospitals, two cancer institutes and two private clinics). From the demographic questionnaires: 24.4% of the clinicians have between 31 and 40 years of practical experience (seniors); 31.1% have between 11 and 30 years of experience (middles); 17.8% have between 6 and 10 years of experience (juniors); and 26.7% have limited experience (interns). Interviews were

[5] A link (mida-project.github.io/uta7-statistical-analysis-charts) from the statistical analysis is hereby provided, so that the community can better understand and replicate our results. In this link, we provide the information visualization via graphical charts of the achieved results for usability, BI-RADS severities, and clinicians' time performance, between others.

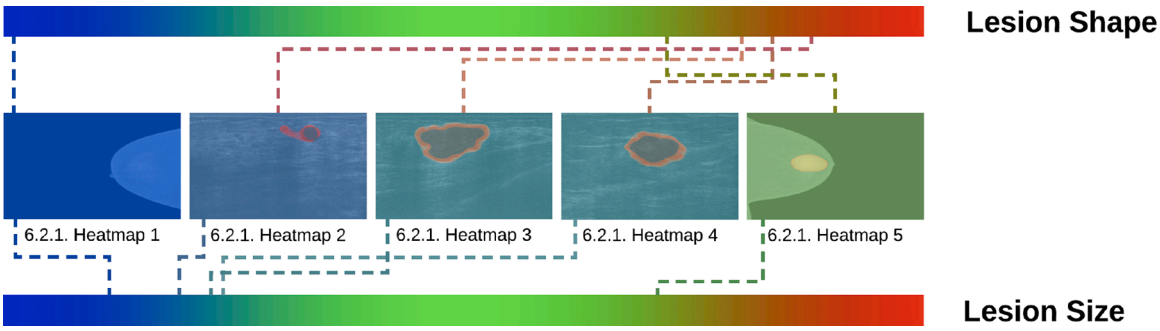


**Fig. 6.** On *6.2. Explain*, the assistant will pop-up several *heatmaps*. The *heatmaps* represent two different scales: (a) *Shape*; and (b) *Size* of the lesion. First, the model computes the BI-RADS using the DenseNet classification. Second, the model computes the lesion circularity (top scale) and size (bottom scale) and associate it to the colors regarding both circularity/size and BI-RADS.

**Table 1**
The Analysis of Variance (ANOVA) factorial analysis table regarding NASA-TLX for *Mental Demand (MD)*, *Physical Demand (PD)* and *Temporal Demand (TD)*, where *F* is the variation between sample means and the variation within samples. To determine whether any of the differences between the means are statistically significant, we used *Sig.* for significance. On the present study, we used a 20-point Likert Scale regarding Workload. The factorial analysis was described assuming $\alpha = 0.05$. Also, each time $p < 0.05$ it is marked with the ★ symbol.

| | MD | | PD | | TD | |
|---|---|---|---|---|---|---|
| Mod. | F | Sig. | F | Sig. | F | Sig. |
| Curre. | 3.392 | 0.027★ | 11.99 | 0.001★ | 10.51 | 0.001★ |
| Assis. | 0.638 | 0.594 | 2.852 | 0.048★ | 0.035 | 0.991 |

conducted in a semi-structured fashion taking about 30 min. Overall, 57 days were spent on the clinical institutions for the observation process and six months for the classification.

### 5.2. Procedure

This section describes the procedure required to reject (or accept) the proposed BI-RADS provided by the *AI-Assistant* (Calisto, 2019a). At this stage, participants will interact with the *BreastScreening* tool using our multi-modal *dataset*. We have a total of 338 cases. In the dataset there exists cases where the patient does not have all the image modalities (recall Fig. 2 where the acquisition may finish before all the modalities are available). Thus, we define the following requirements to conduct our analysis.

The requirements are as follows:

1. All patients must have each of the three available modalities;
2. All patients were annotated and classified by radiologists team of HFF;
3. The patients were grouped in low, medium and high severity according to the BI-RADS;

The above procedure allowed us to obtain a set of 289 cases. Notice that the *dataset* is partitioned according to the three classes above mentioned. The first task was to fill the consent form. The second task was the user characterization form, providing us demographic data about participants. Next, each participant accessed our system via the web browser. We assigned each of the 45 clinicians, three patients (*e.g.*, **P1, P2** or **P3**), for diagnosis. Thus, for each clinician it was assigned one patient with low severity, one with medium severity and another one with high severity.

When analyzing the patients, the main task was to *accept* or *reject* the proposed BI-RADS value provided by the *AI-Assistant*. In case of *rejecting* the proposed value, participants were asked to provide a new BI-RADS value. We also provided an *explain* functionality that could be used to inform participants concerning where and how much sever the lesions are (Fig. 6). This setup will be used to support our results (Section 6).

Participants were given unlimited time to familiarize with the system. Every interaction was shown by the facilitator and upcoming questions were clarified. In the end, for each participant we applied both SUS[6] (Tyllinen et al., 2016) and NASA-TLX[7] (Grier, 2015; Ramkumar et al., 2017) scales on two different questionnaires, respectively. Finally, a *post-task* questionnaire (Calisto, 2019a) was carried out.

### 5.3. Analysis

The study of *BreastScreening* included both quantitative and qualitative analysis. For the qualitative analysis[8] we tracked user interactions across our system, using Hotjar Liikkanen (2017). To record the task activities and the interview, we resorted to screen and sound recordings. Finally, we used Tobii Pro SDK Chatelain et al. (2018), for eye tracking and gaze information. For the quantitative analysis[9] we used the well known SUS[10] (Tyllinen et al., 2016) scale to objectively measure the usability of the *AI-Assistant* setup to answer **RQ3**. To measure the effectiveness performance of the *AI-Assistant* introduction, we used the NASA-TLX[11] (Grier, 2015; Ramkumar et al., 2017), assessing the perceived workload to answer **RQ1**. We also measured the diagnostic to find correlations with the breast severity. Finally, we measured the number of False-Positives and False-Negatives.

Although SUS is more regularly used as single score, in this study we used it as individual questionnaires. Individual SUS scores have typically

[6] For SUS scores, we used a 5 item scale. The scores range from 1 - "**Strong Disagree**" to 5 - "**Strong Agree**". The mean across all individual questionnaires was computed over studies. We provide an available *dataset* (mimbcd-ui.github.io/dataset-uta7-sus) from our SUS data.

[7] For NASA-TLX scores, we used a 20 item scale. The scores range from 1 - "**Very Low**" to 20 - "**Very High**". Again, we provide an available *dataset* (mimbcd-ui.github.io/dataset-uta7-nasa-tlx) from our NASA-TLX data.

[8] By **qualitative analysis** we mean the observational findings from clinicians that identify and answer our design methods and features to use. We divide the **qualitative data** into two groups: (1) **qualitative attitudinal data**; and (2) **qualitative behavioral data**. The first one, can be defined as clinician's thoughts, beliefs and self-reported needs obtained from our user interviews, focus groups and affinity diagrams.

[9] By **quantitative analysis** we mean the use of metrics to measure tasks, which will reflect on the task performance, efficiency and efficacy. Measuring **quantitative data** offer an indirect assessment of the design usability as well.

[10] Regarding **SUS** scores, we provide clinicians an individual questionnaire on a scale that is easily understood. We used **SUS** to measure the *Assistant* usability. The scale provides helpful information about a clinician's takeaways and overall experience during diagnostic.

[11] We used the **NASA-TLX** scale to measure the perceived workload required by the complex, highly demanding tasks of medical imaging diagnosis on the radiology room.

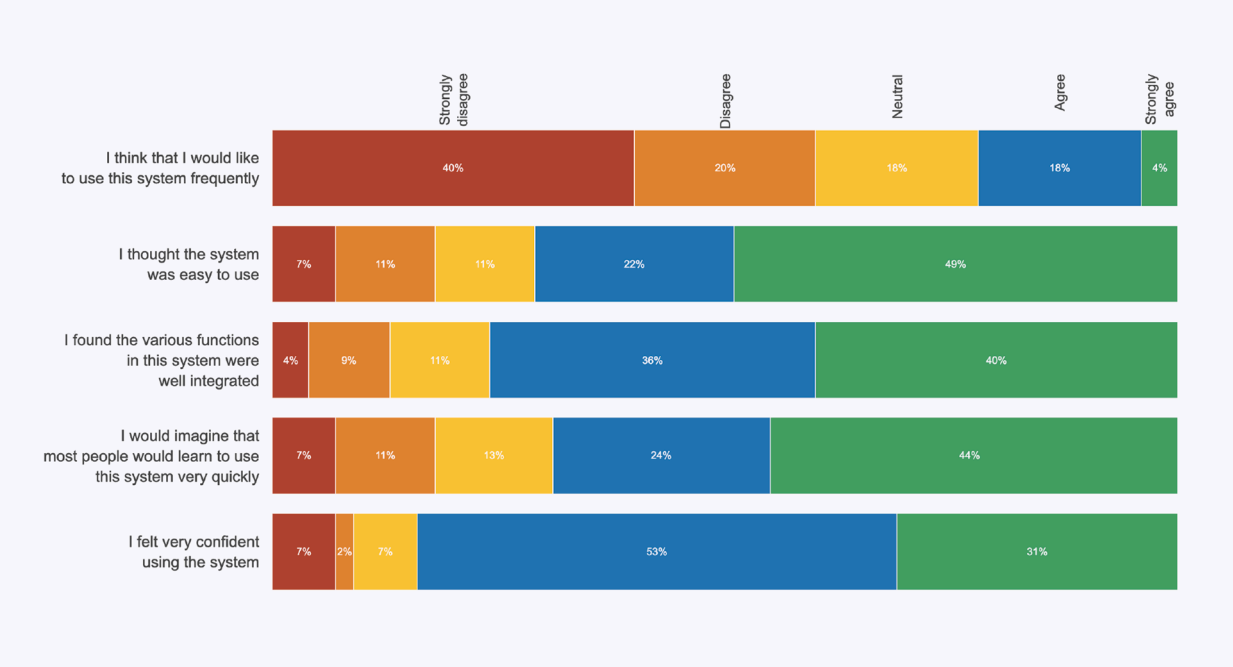


**Fig. 7.** Results of SUS with Positive Questions for the *Current* condition. Each color and respective bar number indicate the mean score for the question, *i.e.*, ranging from 1 = "Strongly disagree" to 5 = "Strongly agree". This figure represents the positive questions. The "Strongly agree" is our optimal value. From the reported results, a majority of the clinicians agree that the system was easy to use and with well integrated functionalities.

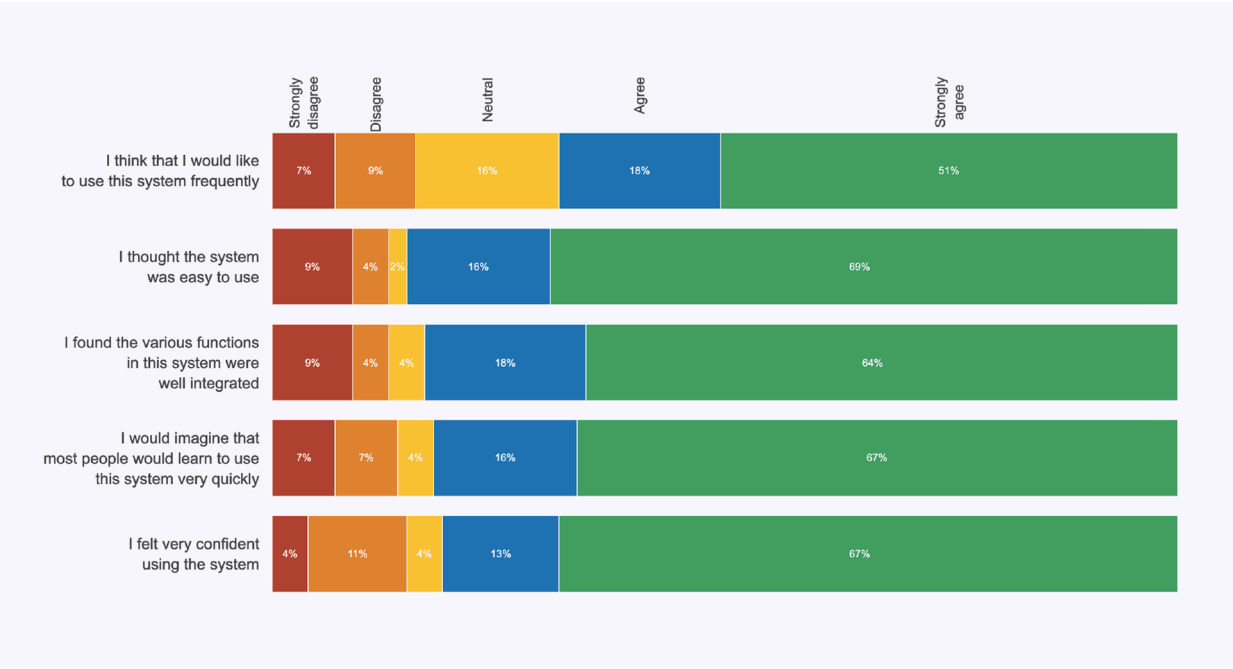


**Fig. 8.** Results of SUS with Positive Questions for the *Assistant* condition. This *Assistant* condition was well accepted by clinicians. Almost all clinicians would like to use this system frequently. Moreover, they consider the system much easier to use and with higher integrated functionalities. Last but not least, clinicians learn more quick and with more confidence this *Assistant* system condition.

negative skew (Lewis, 2018), but the mean sample values are usually normal distributed. Therefore, we took advantage of this statistical behaviour to compute our quantitative analysis (Section 6.1).

On the other hand, we used the basic NASA-TLX scores, which are highly reliable (Ramkumar et al., 2017). NASA-TLX questionnaire consistently exhibits high reliability, user acceptance and low inter-subject variability to measure workload. In our work, NASA-TLX was used to identify clinicians' workload during various stages of the workflow.

Clinicians were asked to complete both SUS and NASA-TLX scores. In the end of the three patients diagnostic, a *post-task* questionnaire was performed. Rating the *Assistant* experience and performance during the diagnostic time period, was our goal with those questionnaires.

The above measurements are part of the quantitative analysis with a

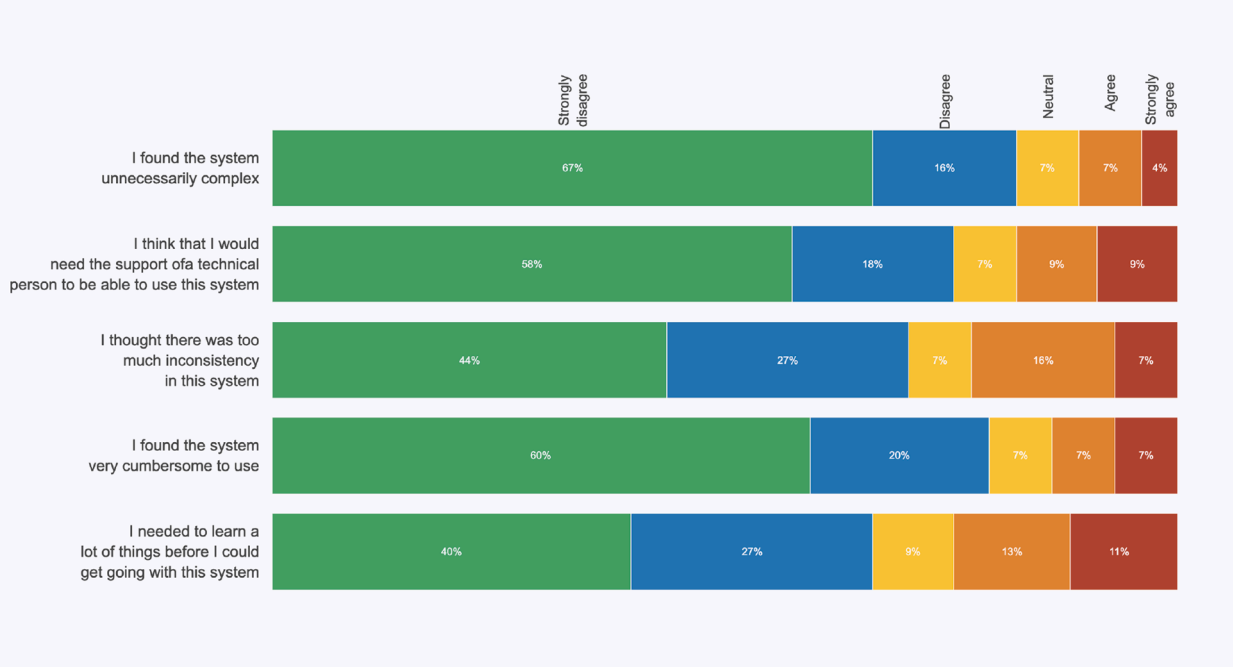


**Fig. 9.** Results of SUS with Negative Questions for the *Current* condition. In this case, we can observe that 23% found the system inconsistent and 24% felt that need to learn before interacting with the system.

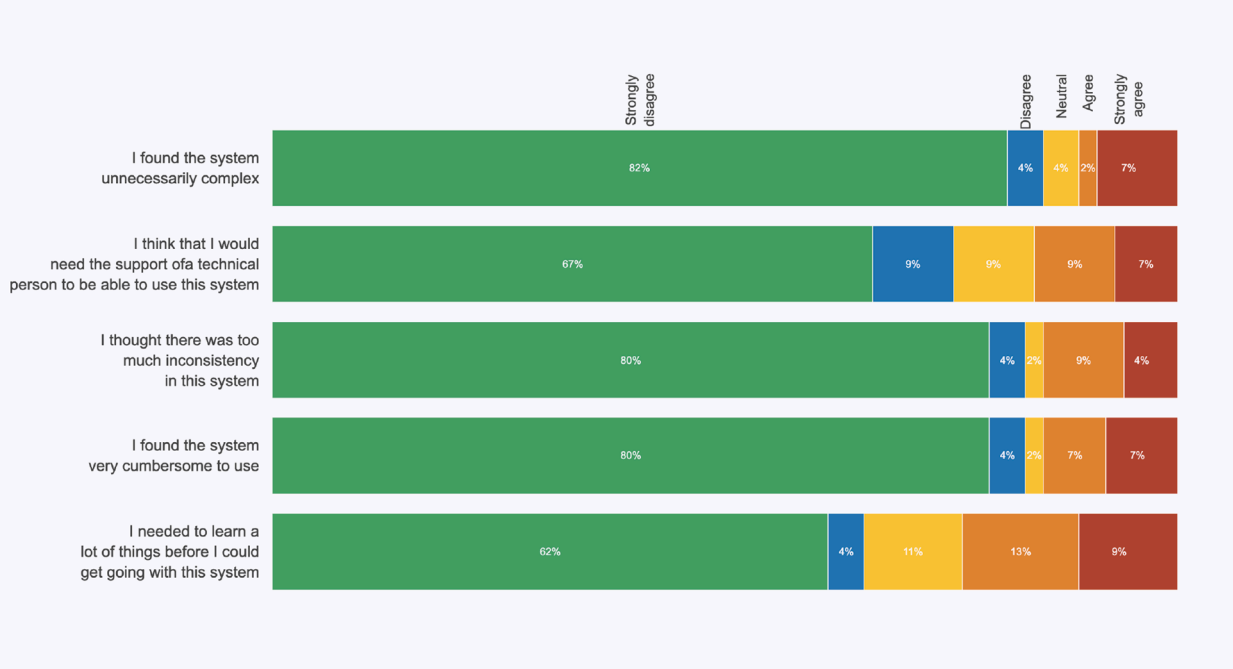


**Fig. 10.** Results of SUS with Negative Questions for the *Assistant* condition. Comparing the *Assistant* with the *Current* condition, we can observe that clinicians found the *Assistant* condition less complex, inconsistent and cumbersome.

comparison between *Current* and *AI-Assistant* setups. With that comparison, we will answer both **RQ1** and **RQ2** providing evidence for the impact and expectations of *AI assistance* on the RR workflow. For the qualitative analysis, we extract opinion-based feedback from the recorded audio. The received feedback was translated to a set of sentences counting the number of clinicians who had the same similar opinion.

## 6. Results

In this section, we will study several concerns by performing a set of quantitative and qualitative analysis of participants' information and behaviour. As follows, we will cover each one.

### 6.1. Quantitative analysis

Our quantitative analysis takes into account differences between medical expert levels (*i.e., Intern, Junior, Middle*, and *Senior*), user

**Table 2**
The Analysis of Variance (ANOVA) factorial analysis table regarding NASA-TLX for *Effort (Eff.), Performance (Per.)* and *Frustration (Fru.)*, where *F* is the variation between sample means and the variation within samples. To determine whether any of the differences between the means are statistically significant, we used *Sig.* for significance. On the present study, we used a 20-point Likert Scale regarding Workload. The factorial analysis was described assuming $\alpha = 0.05$. Also, each time $p < 0.05$ it is marked with the ★ symbol.

| | Eff. | | Per. | | Fru. | |
|---|---|---|---|---|---|---|
| Mod. | F | Sig. | F | Sig. | F | Sig. |
| Curre. | 0.534 | 0.661 | 5.556 | 0.003★ | 2.392 | 0.082 |
| Assis. | 0.664 | 0.578 | 0.319 | 0.811 | 0.408 | 0.748 |

characteristics, and medical imaging interpretation to understand user behaviour during decision making. In this section, we aim at providing a general and straightforward approach to do quantitative analysis and inference understanding data collected during our user studies. The SUS questionnaire was used within the context of assessing *BreastScreening* usability. Hereby, we describe the results obtained from the *SUS Scores* and *SUS Questions* (Figs. 7–10). We provide subjective feedback regarding our interviews from the number of participants who expressed a given comment. By analyzing the numbers, we observe a highly positive adoption of the *Assistant* condition (Figs. 8 and 10) in comparison to the *Current* condition (Figs. 7 and 9), as described next.

Hence, four relations emerged from our analysis:

(a) - differences between *SUS Scores* and *SUS Questions*(Figs. 7–10) across the groups of medical experience (*i.e., Intern, Junior, Middle*, and *Senior*) on the *Assistant* setup;
(b) - workload measurements (Tables 1 and 2) of both *Current* and *Assistant* setups;
(c) - False-Positive and False-Negative ratios (Fig. 12); and
(d) - relation between diagnostic *Time* and lesion severity (*i.e.*, Low, Medium and High values of the BI-RADS) on both conditions (*i.e., Current* and *Assistant*), among different groups of medical experience (Fig. 11);

#### 6.1.1. SUS scores vs SUS positive questions

The generated results of the SUS questionnaire are depicted in Figs. 7 and 8. Specifically, Fig. 7 shows the results obtained with *Current* scenario, and Fig. 8 shows the results with the introduction of the *Assistant*. In short, the *Current* condition obtained 22% of agreement against an obtained 69% for the *Assistant* condition, revealing a higher acceptance for using the *Assistant* (Fig. 8). This means that despite of the clinicians' resistance of change for new tools (Calisto et al., 2017; Gagnon et al., 2014), physicians are now accepting this novel *AI-Assisted* techniques to support their clinical workflow. Indeed, for AI systems to become effective, and consistently accepted and applied within the medical imaging setting, physicians must be able to trust the system and have confidence that the output provided is correct and appropriate for the situation at hand. Another clear fact to explain these values are the results for the ease of use. In those results, the *Assistant* condition reveals an 85% of agreement against the 71% for the *Current* condition. Note that the perceived ease of use is also known as usefulness of the system. Conversely, 82% of clinicians found that the various functions of the *Assistant* were well integrated with the workflow. The confidence level with the *Assistant* was also very high, reaching 80% on this condition.

#### 6.1.2. SUS scores vs SUS negative questions

Similarly to the previous analysis, here we conduct the same study, but now for negative questions (Fig. 10). Regarding the negative questions (Fig. 10), 86% of the physicians disagree that the system is unnecessarily complex. Suggesting that the introduction of an *AI-Assisted* system will not bring more complexity to the diagnostic. This fact is also paired with the NASA-TLX (Tables 1 and 2) results. For the *Current* condition (Fig. 9), 67% of clinicians "Strongly disagree" with the SUS item that the system was unnecessary complex. On the other hand, 4% of clinicians "Strongly agree" that the system was unnecessarily complex. Moreover, 16% of the clinicians "Disagree" with the system complexity SUS item. A total of 83% for disagreement on the *Current* condition. Comparing the *Current* condition with the *Assistant* condition, we observe an improvement of 3% on the total. In fact, the *Assistant* condition improve the total disagreement with the SUS item of system complexity for a 86% as total. More specifically, 82% of the clinicians "Strongly disagree" with that and only 4% "Disagree". For the next SUS item, 60% "Strongly disagree" and 20% "Disagree" that the system is

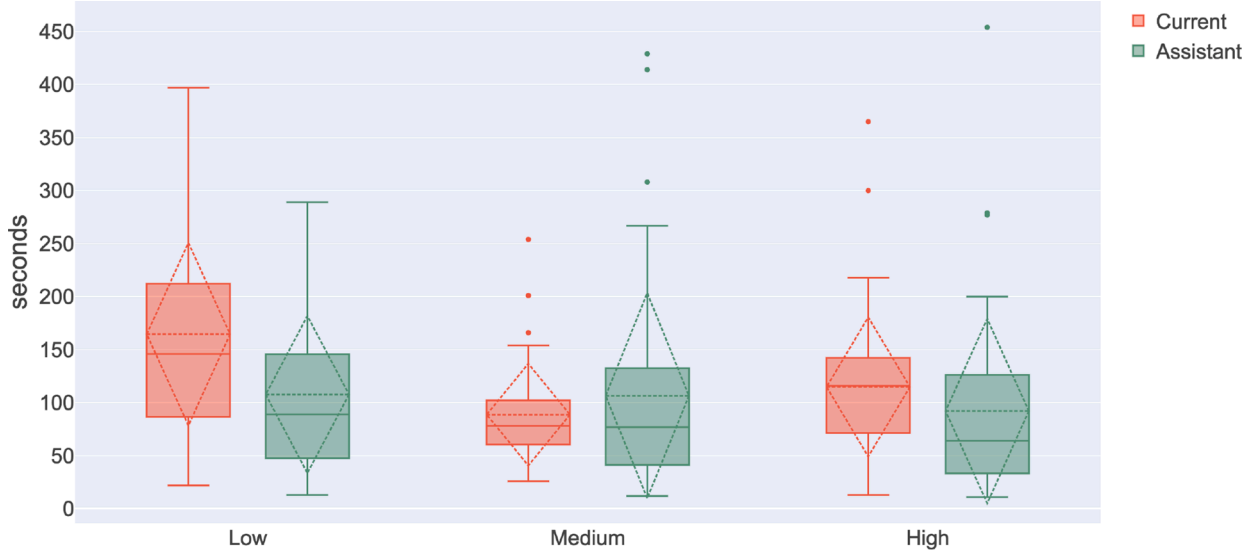


**Fig. 11.** Relation between full diagnostic time length (seconds) and breast severity. We compared both Curre. and Assis. conditions as Low, Medium and High values of BI-RADS. The paper focuses on multimodality, since in current clinical setups the three modalities (MG, US, and MRI) are used for breast diagnosis. Thus, to develop an AI system for exam classification, it is mandatory to take into consideration all the modalities so that it can resemble, and be useful in real scenarios. Indeed, when performing the evaluation in the two conditions, each patient (P1, P2, and P3), has BI-RADS with different severity (Class 1 - Low, Class 2 - Medium, and Class 3 - High) and also, each patient contains all the three modalities.

very cumbersome. Add up to the total of 80% on the *Current* condition. Differently, the *Assistant* condition achieved a total of 84% for this SUS item. Answered by clinicians, 80% "Strongly disagree" and 4% "Disagree". Finally, 40% "Strongly disagree" and 27% "Disagree" with the SUS item that need to learn a lot about the system before start interacting with it. Comparing it with the *Assistant* condition, 62% "Strongly disagree" and 4% "Disagree" with this SUS item. In this item, we can see that the *Current* condition performs better concerning the *Assistant* condition in absolute terms of disagreement. Although, the "Strongly disagree" answer from clinicians was higher for the *Assistant* condition in relative terms.

#### 6.1.3. Workload (Demands)

The results generated from the NASA-TLX (Grier, 2015; Ramkumar et al., 2017) (Mental, Physical and Temporal) Demands are expressed in Table 1. For each NASA-TLX item, the normalized data were first ranked and aligned to the ANOVA[12] measurements. As follows, we present the results for the demands of the NASA-TLX questionnaires. The ANOVA statistical test (Mathews and Marc, 2017; Wobbrock et al., 2011) yields a significant main effect for the Mental Demand ($F_{Curre.} = 3.39$, $p_{Curre.} = 0.027 < 0.05$), Physical Demand ($F_{Curre.} = 11.99$, $p_{Curre.} = 0.001 < 0.05$) and Temporal Demand ($F_{Curre.} = 10.51$, $p_{Curre.} = 0.001 < 0.05$). On the other hand, the *Assistant* condition indicates a significant difference only in Physical Demand ($F_{Assis.} = 2.85$, $p_{Assis.} = 0.048 < 0.05$). A detailed comparison is shown in Table 1. Despite of the higher rates from the NASA-TLX over the several Demands (Table 1), we can point improvements from the *Current* to the *Assistant* setup.

From our study, it can be identified that some functionalities contribute significantly to one (or more) types of workloads (criteria variables) in the NASA-TLX questionnaire. For instance, increasing the number of available image modalities on the viewport is strongly associated to Mental Demand. However, for the *Assistant* condition we could not take conclusions since the fact that their is no significant main effect. The overall time duration of manipulating the images (*i.e.*, zoom, pan, scroll) is strongly associated to the Physical Demand. Comparing both *Current* and *Assistant* conditions, we can observe significant main effect and improvements on the *Assistant* condition. The time duration of decision-making is strongly associated with Temporal Demand. Nonetheless, only the *Current* condition follows a significant main effect making it difficult to do a strong comparison with the *Assistant* condition.

#### 6.1.4. Workload (Non-demands)

The NASA-TLX on the Non-Demands scales only yields significant difference among groups for Performance ($F_{Curre.} = 5.56$, $p_{Curre.} = 0.003 < 0.05$). A more detailed comparison is shown on Table 2. Again, despite of a higher rates one can point improvements from *Current* vs *Assistant*. This is a result of the increasing number of visualization modalities, at the same time, from one (*Current*) to three (*Assistant*) assisted by an AI model. In fact, the improvement scores (**F**) of our *Assistant* are positive. Note that, as far as the scores are less than three times the results of the *Current* condition, one can conclude that we are getting better results.

Effort and Frustration do not provide any significant main effect on both *Current* and *Assistant* conditions. Therefore, we could not consider any findings regarding these issues. Nor even the Performance results, since the fact that the *Assistant* does not represent any significant main effect. Notwithstanding, we will pair (Section 7) the above (Demands) and these (Non-Demands) NASA-TLX results with other metrics to discuss the results with more evidence. A decrease in drawing time will decrease the workload of clinicians, which was confirmed by the lower levels of frustration found with NASA-TLX using the *Assitant* condition. In our study, the frustration measure on the NASA-TLX questionnaire is including aspects of the HCI process while performing the task and are not just limited to the end result.

#### 6.1.5. Diagnostic time vs Breast severity

The results[13] expressing the full diagnostic time length and breast severity among the 289 Patients (*i.e.*, **P1** - Low, **P2** - Medium and **P3** - High severities) are shown in Fig. 11. For the **P1** - Low severity, the *Current* ($M_{Curre.} = 146$, $SD_{Curre.} = 86.17$) condition was longer than the *Assistant* ($M_{Assis.} = 89$, $SD_{Assis.} = 74.13$) condition. Also, for the **P2** - Medium severity, the *Current* ($M_{Curre.} = 78$, $SD_{Curre.} = 48.05$) condition was, again, longer than the *Assistant* ($M_{Assis.} = 77$, $SD_{Assis.} = 96.80$) condition.

Finally, for the **P3** - High severity, the *Current* ($M_{Curre.} = 116$, $SD_{Curre.} = 65.70$) condition was longer than the *Assistant* ($M_{Assis.} = 64$, $SD_{Assis.} = 86.94$) condition. The ANOVA statistical test shows a significant effect over the total *Time* for the *Current* ($F_{Curre.} = 3.25$, $p_{Curre.} = 0.03 < 0.05$) condition regarding the clinical experience groups on a **P1** - Low severity case.

Planned post-hoc testing Zhang et al. (2016), using the Tukey's HSD Post-Hoc Comparison ($p_{Curre.} < 0.05$), revealed that for the groups of *Juniors* significantly increased the time performance and severity accuracy compared to *Interns*. Therefore, our assistant, not only improves time performance (Fig. 11) and diagnostic accuracy (Fig. 12) among the others physicians' categories of medical experience, but also provides important support for *Interns*.

These results support **RQ1**, suggesting that our *BreastScreening* tool could impact positively the clinical workflow, mitigating the different types of errors on clinical perception by improving time performance and diagnostic accuracy.

#### 6.1.6. False-negatives vs False-positives

We also measured the rates of False-Negatives and False-Positives (Fig. 12) between *Current* and *Assistant* conditions. The False-Negative rates decrease from 33% on the *Current* condition to 14% on the *Assistant* condition. From our *dataset*,[14] the results show a significant potential reduction of False-Negatives, i.e., cases where the diagnosis leads to a low severity (BI-RADS) against expert ground truth.

Notwithstanding, the False-Positive rates decrease from 39% on the *Current* condition to the 15% on the *Assistant* condition for an overall (*i.e.*, Total) condition. In essence, we will have a 24% decrease of situations where radiologists are providing a BI-RADS higher than the real one.

### 6.2. Qualitative analysis

We complemented our quantitative analysis with insights and results from interviewing participants. In this section, we describe the study of a preliminary design for the development of our *Assistant*, informed by an iterative process to identify clinician's needs and recommendations. First of all, from a set of workshops (Section 6.2.1) we introduce participants to aspects or open-ended questions that can drive this initial

[12] *N*: the number of users (Clinicians); $F_{var}$: the F-test used for comparing the factors of the total deviation per each variable (*var*) categorized by clinical experience; $M_{var}$: Mean value of the variable (*var*); $SD_{var}$: the Standard Deviation (SD) per each variable (*var*). Notice that from the statistical significance analysis described in Table 1 and setting a significance threshold to 0.05, two scenarios are possible to occur. First, if we obtain a *p*-value $> 0.05$, this means that the approaches are not statistically different, better saying, we can not state anything about the data. On the contrary, if the *p*-value $< 0.05$ the approaches are statistically different, since now we can reject the null hypothesis that states there is not a statistically significant difference between results of the proposed method and the other methods compared.

[13] We provide an available *dataset* (mimbcd-ui.github.io/dataset-uta7-time) from our *time* data.

[14] We provide an available *dataset* (mimbcd-ui.github.io/dataset-uta7-rates) from our severity *rates* (BI-RADS) data.

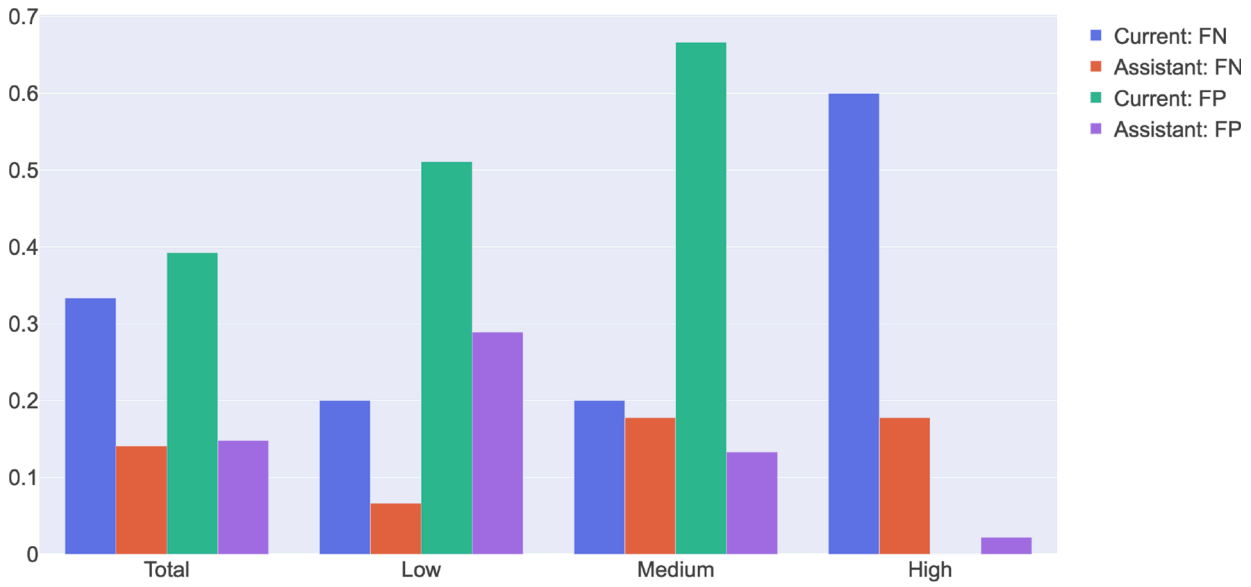


**Fig. 12.** *Current* **vs** *Assistant* rates for False-Negatives and False-Positives. A False-Positive is considered when the BI-RADS$_{provided}$ > BI-RADS$_{real}$. A False-Negative is considered when the BI-RADS$_{provided}$ < BI-RADS$_{real}$. Each patient (P1 - Low, P2 - Medium, and P3 - High) comprises all three (MG, US, and MRI) modalities.

stage of qualitative data analysis. In this workshops, participants are divided into small groups. Second, by joining our research team with participants, it is established the focus group (Section 6.2.2). And third, to cluster and categorize (Section 6.2.3) the set ideas, features and priorities, we introduce in this focus group a lightweight approach called affinity diagrams (Section 6.2.4). The following sections will detail and describe the qualitative analysis.

### 6.2.1. *Workshops with clinicians*

The first step of our methodology was to record *workflow* practices and routines from clinicians (Høiseth et al., 2013a; 2013b). To this end, several invitations were sent among the various medical institutions and at least one workshop was formed per each institution as described next. We grouped all participants volunteered to our study, at least one day per each institution. For HFF (public hospital), we did four workshop meetings, while it was the institution with more clinicians and, therefore, harder to schedule. For IPO-Lisboa, a public cancer institution, we did two workshop meetings, as well as for the HB public hospital. For the other institutions, we did just one workshop meeting. A total of 45 professionals from the sector of healthcare (*i.e.*, Radiologists, Oncologists, and Surgeons), as well as six members of the *BreastScreening* project (*i.e.*, HCI and AI Researchers) participated in these series of workshops.

Based on a preliminary content analysis of the semi-structured interviews with clinicians, we conducted the workshop as part of the *BreastScreening* project development. Participants worked in groups and brainstormed around their clinical practices and routines. Most of the practices are recorded and written. Moreover, notes were digitally transcribed.[15]

The duration of the workshop per session was roughly two hours and ended with joint sessions wherein each group of clinicians highlighted important aspects of the clinical *workflow* for their institutions. At the end of the workshop sessions, we collected four different procedures (Fig. 2) of acquiring medical images. Participants engaged into our planed design activities (Calisto, 2019a), in which they provide us inputs regarding the current *Assistant*. Using the provided input from the workshops, we designed a prototype that was evaluated during several sessions at our nine clinical institutions, such as public and private hospitals, public cancer centers, as well as private clinics.

Another important aspect is that the MG image modality is always present on a first stage of medical image acquisition, mainly because of its low cost. The US is the second most preferred modality to cross information between MG image modality views (*i.e.*, CC or MLO). Finally, because of the high costs (*e.g*, time of acquiring the images) associated with the MRI, clinicians said (in nine institutions only one follows the **Proc. 4**, see Fig. 2) it was typically recommended only for highly risk patients.

### 6.2.2. *Focus group*

Building on the qualitative data from the workshops with clinicians, a focus group consisting of six Researchers (MSc, PhD and Post-Doc students, as well as Assistant and Full Professors) and another six Radiologists (2 seniors; 1 middle; 2 juniors; and 1 intern) from HFF public hospital, organized the *workflow* practices and main feature ideas to greater detail by using affinity diagrams (Harboe et al., 2012; Høiseth et al., 2013b). The affinity diagrams enable us to identify several functionalities, such as the need to *accept* or *reject* (Fig. 5) our *Assistant* result. Also important, it was from the affinity diagrams that we achieve a high priority feature of developing a technique so that in case of *reject* our *Assistant* result, the clinician can provide new information to the DenseNet.

This technique is novel and, while applying these HCI practices, it provides a twofold of contributions:

1. We created a new way of control on the introduction of AI methods among medical imaging diagnosis; and
2. The inclusion of a DNN in the UI, particularly, the introduction of a pre-trained DenseNet capable to provide a fast and reliable classification.

### 6.2.3. *Affinity diagrams*

This was an interactive process that consisted of adding or removing items until a final pattern configuration is reached. As these ideas and functionalities from the clinicians could be relevant for several purposes, it was considered useful to short on a more general level to start with.

[15] Affinity diagramming is a powerful method for performing qualitative data (mimbcd-ui.github.io/dataset-uta7-ad) organization and analysis. The method was used to help understand the role of technology into the RR workflow. More precisely, affinity diagrams were used to organize the provided information from clinicians to group it with related ideas or topics.

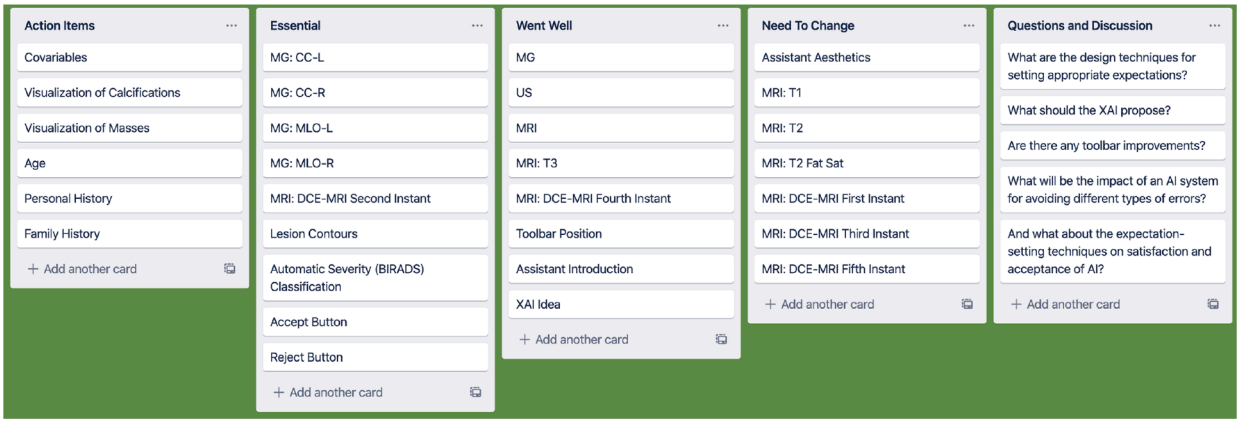


**Fig. 13.** Resulting affinity diagrams passed to a digital software tool. The overall ideas and features categorization. Each idea and feature has a category (*e.g.*, "Action Items", "Essential", "Went Well", "Need To Change" or "Questions and Discussion"), so that we can manage the final requirements and development priorities of our *Assistant*.

Ideas, functionalities and priorities were translated (Fig. 13) to a digital tool,[16] while the affinity diagrams (Subramonyam et al., 2019) are used for: (i) categorization of the focus group items; and (ii) data clustering of the chaotic information. This solves our problem (Section 6.2.4) for chaotic data (McKay et al., 2020; Yang et al., 2016). From workshops (Section 6.2.1), participants (*i.e.*, researchers and some of the clinicians) of the focus group (Section 6.2.2) were asked to review and re-position ideas and functionalities, within each category, in order to organize them.

Every time an idea or functionality was triggered, we put it on the "Action Items" category. After that, participants discuss where it should be, answering the workshop needs for item categorization. For instance, several clinicians (Section 4) listed[17] their preferred components as the position (32/45) and simplicity (28/45) of the *Toolbar*: "The *Toolbar* position is in a better place on the top in contrary what we usually [Current] see" (C30). From here, we created an item titled as "Toolbar Position" and since it was accepted by a major number of clinicians (and rejected or omitted by another minor number), we stuck the item into the "Went Well" category.

Findings were documented as notes and arranged into a hierarchical organization of common themes based on the provided data of clinicians' ideas and opinions, defined in categories. This data was also organized into consolidated needs that characterized each institution work practices, structure, and requirements. After completing the workshop and focus group sessions, researchers met and analyzed the data. At the conclusion of this process and completion of all data interpretation, consolidated affinity diagrams were created to represent the structure of clinician's needs.

Through the affinity diagrams, we found that specific items, within each own categories, correspond to three needs of clinicians:

(a) - new strategies among the medical imaging visualizations;
(b) - to control (*e.g., accept* or *reject*) the final *Assistant* result; and
(c) - to understand the *Assistant* result.

The approach of the three needs was defined by the focus group as mapping all *workflow* processes and activities that clinicians should perform in order to finish all pipelines of the diagnostic. For example, it was here that the importance for the introduction of such AI systems was emphasized. One of the clinicians even argue that: "If we have an intelligent assistant like this in our workflow, it will be more simple and easy to do our job" (C45). The settings for the clinicians' need include all the preconditions that helped to enroll with the diagnosis more promptly finished.

Throughout this process a set of design guidelines have been derived. In order to investigate how the clinical *workflow* proceeds, the affinity diagrams were used to connect ideas and features to a set of guidelines that we will describe next. For this purpose, we created three central design components that can be applied for medical imaging systems with AI behind: (i) explaining the important lesion regions; (ii) higher interpretability of the AI results; and (iii) providing control for the final result.

From Fig. 13 and from the feedback obtained when building the affinity diagrams, the following design guidelines were considered:

> [Relevance to Diagnostic] our AI system (*Assistant*) should provide relevant clinical information so that clinicians can explore various aspects of the diagnosis. For instance, information that allows clinicians to understand where are the relevant lesions (Fig. 6) and what are the respective levels of severity regarding both shape and size of the lesion.
> [Clinician-Centered Activities] since our data must be shared with a team of clinicians, our AI system should provide collaboration. Lesion annotations in context for collaboration should, for instance, be visualized synchronously by two (or more) clinicians. If a clinician starts annotating a lesion, the same annotation should be visualized remotely by another.
> [Provide Explanations] the system must provide answers regarding the final AI result. In our work, we created an "Explain" (Fig. 5) so that the clinician can open the heatmaps (Fig. 6) on the image. The heatmaps will show the variability (color) of the important regions. Which is information that will explain the final BI-RADS.
> [Feeling in Control] the *Assistant* must provide control for the final decision. Clinicians must feel that, in case of a wrong AI diagnostic, the final result must be changed (*reject*) by them. So that we can guarantee the patient safety and right treatment of the lesion.

The data was acquired based on the *affinity* of the collected ideas and functionalities. Next, we describe how we clustered the acquired data.

### 6.2.4. Data clustering

Data clustering is an important step for the developed UI, as we aim

[16] For that, we used the collaboration software Trello (trello.com), which allowed us to digitally organize and manage the group ideas. While inserting the ideas in the Trello board to be used within the affinity diagramming process, this workflow makes it possible for a facilitator to either capture the affinity diagrams as they are created.

[17] We transcribe the workshop answers and feedback so that we can join similar opinions in different items. A "(32/45)" means that 32 clinicians for a total of 45 clinicians appointed a similar sentence of the clinician number 30, *i.e.*, "(C30)" on that example.

to extract themes from clinicians across diverse institutions and clinical practices. The process of gathering information and affinity diagramming is inspired from chaotic data itself. It is not based on fixed quantitative data, and it does not verify an hypothesis but rather inspires an hypothesis itself. Indeed, the MRI data is inherently chaotic, since each exam contains tens of volumes.

Specifically, we have:

1. T1;
2. T2;
3. T2 Fat Sat;
4. T2 TIRM - Turbo Inversion Recovery Magnitude;
5. Diffusion - with and without SUB;
6. DCE MIP - Maximum Intensity Projections;
7. DCE WO;
8. DCE PEI;

As notes are placed, and in moving them around later, they are clustered based in their *affinity, i.e.*, their similarity or relevance to a shared topic. This leads to the creation of data groups, which are labeled and recursively clustered. The process is repeated until the highest level has only a few groups and the initially unstructured items have been organized bottom-up (Harrington, 2016; Subramonyam et al., 2019; Yang et al., 2018). Clusters and then given titles are grouped into more abstract groups, giving rise to general and overarching ideas. Ideas are then clustered again to identify common issues and potential solutions, ultimately helping to frame the user needs and design problems.

While doing the affinity diagrams, the focus group responded (Fig. 13) with several user needs and requirements. Such user needs and requirements are crucial to define which are the most important (*i.e.*, "Essential" column of Fig. 13) modalities and what are the procedures (Fig. 2) on the workflow. More specifically, as we have several MRI sequence[18] options, it is really hard for clinicians to proceed to the visualization of all MRI volumes. What we call chaotic data problem. Thus, from a set of MRI volumes, we need to figure out, what is the ones that best matches the radiologist needs.

We should highlight that, during the focus group, we could not reach a consensus within the radiologists from HFF public hospital and IPO-Lisboa public cancer center for the MRI volumes. On one hand, HFF are using DCE-MRI in second instant. On the other hand, IPO-Lisboa are using T3. Since we were using medical imaging data from HFF public hospital, and we had consensus from the clinicians of this hospital, we choose their main sequences (*i.e.*, DCE-MRI in second instant) as standard. From data clustering, we could understand the need for each modality (*e.g.*, MG, US and DCE-MRI at the second time instant) and what is the meaning to the clinical workflow (Fig. 2). For instance, thanks to the workshops and the focus groups, we could take important data regarding how these modalities constitute complementary information for a reliable diagnosis per institution.

#### 6.2.5. Prototype

A concept prototype of a medical imaging diagnosis system has been developed to provide clinicians with decision making recommendations, thanks to the integration of our AI methods as part of the *BreastScreening-AI* (Fig. 5) research work. The *AI-Assisted* prototype consists of two parts: (1) the diagnosis via medical images, in which our prototype has several functionalities to support clinicians with tools for a proper final result; and (2) the automatic severity classification of lesions, in which we have a DenseNet providing the BI-RADS values. Our prototype is based on insights from a human-centered approach, which included investigations such as observations, interviews, dialogues, and workshops, as well as a co-design process with clinicians.

#### 6.2.6. Final visual configuration achievements

Thanks to the process of putting something visual with real cases in front of clinicians, the workshops and the affinity diagrams, we could explore in a higher manner the good insights of them. It was at this stage of the study, that several clinicians provided important modifications and feedback for the future final system. Such modifications were, for instance, bringing the *Assistant* (Fig. 5) from the top and middle to the button right of the screen inside the *5.1. Viewports*. Again, the comment was provided during the interaction with the prototype at this phase. Otherwise we did not test the suggestion. The suggestion made us improve at about a 40% factor of the time (Section 6) over interaction with our *Assistant*.

The explanation was simple, clinicians are use to work inside the *5.1. Viewports* (Fig. 5). Meaning that it is less distance, time and effort to achieve the 6. *Assistant* avatar. Furthermore, the visibility was not compromised because of two reasons: (a) first of all, most of the cases (typically the majority of MG and MRI modalities) has no relevant image information on the button right of the screen inside the *5.1. Viewports*; and (b) we developed a hidden button, so that if a clinician really need to look at that area, the 6. *Assistant* avatar will disappear, showing the full image with nothing overlaid.

#### 6.2.7. Qualitative feedback from clinicians

Our qualitative data present an initial attempt to exploit knowledge from clinicians into useful guidelines. Now, we want to address a presented result of applying these guidelines to our work and which was the clinician's final opinions and feedback. The following sentences are addressing the most important clinician's final opinions and feedback for the final solution. At the end several clinicians (28/45) answered that the assistant will be an asset of an immense importance for the current RR situation: "The system [*BreastScreening* assistant] will be a great asset for us" (C6). Several clinicians reported that the system is intuitive (33/45) and easy to learn (28/45): "When I start exploring the new system [*BreastScreening* assistant] I found it very fast to learn and intuitive to use" (C15); and "The interface [*BreastScreening* assistant] is easy to learn and I do not need any help" (C10). Another positive answer was the one related to the frequency of use (41/45) for this new assistant regarding the current system used by the clinicians on the daily practice: "I would like to frequently use this on my daily practice" (C1).

The above statements confirm the efficiency and acceptability of the proposed *AI-assisted* prototype. This suggests that it is an improvement on the current setups, and fits the Reading Room (RR) workflow.[19] From the positive feedback, we expect that this tool will be helpful to improve the diagnostic results over time, providing higher health care to the patients. Several reports with details of clinicians comments, system architecture, and description of *dataset* (results of the usability, workload scales and other metrics) will be made available.

## 7. Discussion

The optimal use of the *Assistant* within clinical workflows (**RQ1**) remains to be determined. The specificity advantage exhibited by the *Assistant* suggests that it could help to improve diagnostic accuracy and, therefore, unnecessary biopsies. Our findings suggest that there is an

[18] As most sensitive method for detection of breast cancer (radiopaedia.org/articles/breast-mri), the breast MRI aims to obtain a reliable evaluation of any lesion within the breast. It is always used as an adjunct to the standard diagnostic procedures of the breast, *i.e.*, clinical examination, MG and US.

[19] The **Reading Room (RR)** is the space in which we have non-invasive imaging scans to diagnose a patient. The tests and equipment involves low dose of radiation to create a highly detailed image of the breast area. Each room accommodates the controls and appropriate accessories. There are several goals for these reading rooms: (1) read a large number of exams; (2) find as many of the screening cancers as possible; (3) provide a comfortable work environment for the radiologists; and (4) instill confidence in the quality of the work.

evidence that the integration of the AI in the UI, can help for an earlier cancer detection. Moreover, as presented on the results section (Section 6), the introduction of AI was well received by clinicians, while our *Assistant* is above their expectations (**RQ2**). Finally, supported by our results (Section 6), we show that clinicians' satisfaction and acceptance (**RQ3**) of AI assistance successfully impacted the intended aspects of expectations. The next sections will describe and discuss this paper contributions in terms of clinical expectations and AI assistance.

### 7.1. Contribution for clinical expectations

Our work provides insights into feasible *AI-Assisted* mechanisms on medical imaging diagnosis. In this study, clinicians agree that an *Assistant* could improve their workflow. Actually, by accessing our usability measures (*i.e.*, SUS) we observe that 31 clinicians in 45 are open to adopt this new paradigm. In terms of workload, the *Assistant* results are much better than the *Current* results. In fact, only Effort was outperformed. The diagnostic time performance was also improved for low, medium and high severities. Although, the improvements for the medium severities are merely small improvements. The rates of False-Negatives and False-Positives were also reduced. Which means that we are decreasing the number of medical-errors.

If we address the provided feedback, we can see that 41 clinicians would like such as a system to enter their daily practice. Of those, 28 mentioned that the system will be part of an important asset, improving their job. One important aspect of our approach was giving clinicians the opportunity to revise the AI recommendation. At the same time, the AI recommendations are also providing the opportunity to *accept* or *reject* the suggested diagnosis.

In this study, we show that AI can be integrated on the clinical workflow (**RQ1**), impacting the same diagnostic types. Importantly, the proposed *Assistant* was integrated into a clinical radiology RR scenario. The *Assistant* has significantly benefit clinicians on diagnostic time (Section 6). Actually, 86% of the clinicians are finding the *Assistant* is not complex. Further, 84% of the clinicians found that the system was not cumbersome. Making the *Assistant* quick to interact with and trivial. This will impact the workflow on a positive manner, in consideration of clinicians perceiving that the *Assistant* will bring less complexity to their workflow. Indeed, we also verified that for both low and high severities, the *Assistant* improve the diagnostic time improving the final decision in more than 50 s per patient. In addition, we are also improving the numbers of False-Positives and False-Negatives. Therefore, the *Assistant* will impact the clinical workflow in terms of time and accuracy.

Our findings show that supporting explainability[20] and intelligibility[21] (Abdul et al., 2018), improved the acceptance and confidence of clinicians. In fact, results are showing that about 37 clinicians are accepting our approach and are feeling confident using it. A key factor to our results is pairing the *accept* or *reject* features with several visual explainability techniques (Fig. 6) that empower the clinicians' choice and sense of control. Hence, the outcome is achieved for clinical expectations, answering the **RQ2** question on how to improve (via interpretation of *heatmaps*) diagnostic interpretability.

We believe our *BreastScreening* framework can offer a substantial contribution to the breast cancer domain. We show (Figs. 7–10) that user satisfaction and acceptance (**RQ3**) can be improved, not only through even more accurate models, but also higher expectation adjustment techniques. Such techniques, could also contribute to improve our UI explainability methods. Not just using simple *heatmaps*, but also using other important image feature information. This addresses an important gap in existing research on preparing clinicians for the introduction of *AI-Assistive* techniques of their workflow (Alkhatib and Bernstein, 2019; Challen et al., 2019; Shah et al., 2019; Szolovits, 2019).

With our results, we could verify the proposed research questions providing evidence of the design and integration of *AI-Assisted* methods on a medical imaging domain. For the breast cancer diagnosis, the introduction of an *Assistant* could impact the clinical workflow (**RQ1**) on a positive way. By setting visual explainability techniques, we show how to support clinician expectations (**RQ2**) of AI assistance and how to improve diagnostic interpretability. Finally, we successfully measure the impact of expectation-setting intervention techniques (**RQ3**) on satisfaction and acceptance of AI assistance in radiology.

### 7.2. Contribution for AI assistance

While Human-AI interaction tools have traditionally been used to improve algorithms, we found that *AI-Assisted* mechanisms empowered clinicians in the medical imaging diagnosis. More precisely, a medical *Assistant* can make itself more understandable to clinicians by providing some kind of explanation (*heatmaps*).

In this work, we assume a setting in which either the Human or AI (or both) has ground-truth knowledge of how to diagnose the patient and classify the breast severity (BI-RADS). With the *accept* and *reject* features, both Human and AI are learning together about the diagnostic task. Such Human-AI interaction will improve clinicians' transparency as a result of this bidirectional process. Nevertheless, our findings suggest new ways of improving AI transparency (Cai et al., 2019a).

Quantitative results, such as usability (*i.e.*, SUS) or workload (*i.e.*, NASA-TLX), point to improvements in terms of satisfaction (Bonham, 2019) and acceptance (Gambino et al., 2019; Sonntag et al., 2012), as well as a positive *workflow* impact (De Backere et al., 2015). The magnitude of the SUS usability measure employed by clinicians to describe their experience with our *Assistant* showed "good" usability (Yu et al., 2020) (*i.e.*, SUS $>$ 68) and improvements on the workload values. We also achieve improvements in performance and accuracy. From our results, we can point that the time performance was almost 2x faster with the *Assistant* setup and more than 2x accurate. Apart from being passive recipients of the machine outputs, clinicians could play an active role, improving data for the learning process on a Human-AI collaboration. Indeed, this interactive collaboration could help clinicians from mental models and increasing assistant transparency (Amershi et al., 2014; Cai et al., 2019b; Eslami et al., 2016).

Qualitative results, such as workshops, focus groups, affinity diagrams and feedback from the interviews, are making a strong contribution in terms of how decisions are reached across many clinician roles and contexts. We received positive feedback regarding our *BreastScreening* assistant. Specifically, the decision benefits from using an integrated AI for automatic breast classification on the UI. Two main outcomes resulted from this qualitative study. First of all, the adoption of *heatmaps*. Second, what are the main requirements for clinicians, such as imaging modalities, the need for control (*i.e.*, the *accept* and *reject* features), and what volumes to choose. With these two implemented outcomes, we could cover some of the hazards on our *BreastScreening* assistant prototype.

Since clinicians prefer to see the lesion with context, we need to provide color information regarding both the shape and size of the lesions (Fig. 6). Therefore, we chose this technique to support the lesion visualization use case. Also, the *AI-Assistant*, for a given patient, provides a BI-RADS classification for each modality. However, from the interviews, we realized it was paramount to provide a "global" classification of the exam, instead of modality based. This means that our *AI-Assistant* must provide the worst-case classification as the global score. Furthermore, the image-modality assigned to the highest BI-RADS should be displayed in the UI, providing the "explainability" (*6.2 Explain* button of Fig. 5) of such a global score.

[20] **Explainability:** is the use of models that are able to summarize the reasons for Neural Network behaviour, gaining the user's trust, or producing insights about the decisions causes.

[21] **Intelligibility:** is defined by the use of inherently interpretable models or by developing methods for explaining otherwise overwhelmingly complex decisions.

In short, *BreastScreening* uses large amounts of data, due to its multi-view and multi-modality nature. The very first step on both quantitative and qualitative studies was to extract from our study information a set of analysis - a design-thinking method, which was crucial to perform data collection and cluster information. Specifically, data clustering resulted in the following clusters: (i) MG (both CC and MLO views); (ii) US; and (iii) DCE-MRI volume. Note that in (i) a large number of views are available, *e.g.*, ML, LM, LMO, late ML, among others. Concerning (iii), radiologists acquire a large set of MRI volumes (T1, T2, T2 Fat-Sat, Diffusion, Dynamic Contrast-Enhanced (DCE), etc). This initial study resulted in a consensus that allowed the selection of CC and MLO views, in the case of MG, and DCE-MRI in MRI. This information was suppressed in the paper since our main focus was on the evaluation of the (quantitative and qualitative) impact of an AI integration, and where we assumed that the data collection was readily available.

## 8. Conclusion

Clinical translation of *radiomics* is a promising but challenging research topic. The integration of AI techniques into the clinical workflow requires a holistic approach which can benefit from an HCI perspective such as the one provided here. In our work, we need to consider several factors when making decisions about how to add AI to medical practice. For instance, bias is a familiar concept for clinicians, who are already trained to practice evidence-based medicine. On the other hand, AI can suffer from selection bias (*i.e.*, of training *datasets*), automation bias, and data shift. We must carefully balance trust and risk in AI implementation, by acknowledging remarks and limitations, as well as potential consequences of AI-driven diagnosis. Our *Assistant* design concerns the interaction of clinicians and AI in a real-world setting. Further work is necessary to transition AI from highly controlled experimental environments to real life practice.

### 8.1. Main contributions

In this paper, we implemented and studied the introduction of an assistant in the clinical workflow as a tool to classify and explain medical imaging diagnosis in the breast cancer domain. Indeed, the results are showing that AI can positively impact the clinical workflow at a cost of providing clinician control over the proposed BI-RADS classification. Specifically, medical assistants are contributing to an increase of clinicians' perception of trust. The proposed investigations focus on two different contexts. As follows, each of the two different contexts will be detailed.

First of all, we applied these developments as a second reader assistant mimiking the expert domain of breast cancer interpretation and classification for a second opinion. *BreastScreening-AI* was developed on the top of *BreastScreening* core to study the impact of AI-assistance on the medical imaging workflow as an autonomous and imidiatly available second reader. Insights from an experiment with 45 clinicians showed that a medical assistant can make itself more understandable by providing some type of explanation (*i.e., heatmaps*) without comprising its reliability. Moreover, satisfaction and acceptance were significantly improved by the assistant with this strategy.

Second, a study was promoted to understand how AI assistants are (positively) affecting the medical workflow. Clinicians' accuracy was compared with and without AI, while studying the impact of several design techniques in terms of clinicians' expectations and satisfaction. Consistent with expectations, the results are showing clinicians' accuracy and acceptance of an AI system optimized for high specificity can be significant higher than a system optimized for higher sensitivity values. Furthermore, by measuring FP/FN and relating to time performance, results are suggesting that the proposed adjustment techniques successfully impact the intended aspects of expectations. Finally, the results are showing that these techniques are successful in reducing variability among the clinical groups of experts on an AI system. Thus, converging to a more reliable final diagnostic by reducing the independence levels of less experient clinicians (*i.e.*, Interns and Juniors) in disagreement (Schaekermann et al., 2018) with higher expert clinicians' (*i.e.*, Middles and Seniors) results.

To conclude, it was demonstrated how an AI system focus on the communication of model performance (*e.g.*, specificity and sensitivity) can lead to much higher perceptions of accuracy. Due to a model performance communication, intelligent agents are increasing the acceptance of medical professionals by showing the visual representations for classification, performance and segmentation of lesions. With these visual representations, model ambiguity (Schaekermann, Mike, 2020) can be adjudicated (Schaekermann et al., 2019b; 2019c) by clinicians to control the results of the final diagnostic. Therefore, clinicians are increasing their AI confidence with higher values of *trust*.

### 8.2. Findings

Medical AI-assistants are representing a type of system that is passive (*i.e.*, assistive as a second reader) and casual. Clinicians can *accept, reject* and ask the system to *explain*, or even *ignore* the assistant result. While this is only one class of CDSS, we believe that this class represents many current efforts of integrating HAII into other (Savage, 2019; Shah et al., 2019; Topol, 2019) clinical domains (*e.g.*, systems for clinical drug development, epidemiology, dementia treatment, etc).

Findings should be generalized to other clinical systems and tasks according to the following claims:

1. Evaluations are task agnostic as they are informed by the high-level workflow actions and autonomous diagnostic mechanisms in which clinicians make decisions, while a workflow understanding is crucial to recognize clinicians' needs;
2. Intelligent agents should provide concrete empirical evidence and insights by detailing the benefits of avoiding the different types of AI errors and diagnostic mistakes to address model ambiguity in various steps of the AI pipeline;
3. An AI system based on *Precision* and *Recall* optimization will improve higher values of FPs and FNs as model uncertainty and explanations are provided to clinicians so that they can be aware of model misunderstandings;

The importance of avoiding different types of errors depends on the information available from a specific clinical domain. Indeed, this is a complex issue. While it could be argued that for the breast cancer diagnosis FNs are always better to avoid, such results can not always suggest that they are independently to the clinical domain. In this thesis, avoiding FNs might be better. Actually, avoiding FNs can be better as a result of escaping from saying there is no cancer, when surely there is. However, in other clinical domains an FP might be more important to avoid. For instance, in drug development, the domain wants to prefer avoiding the FP rates. As a consequence, the applications of AI systems are optimizing the number of chemicals and drugs identified incorrectly by the system (Raja et al., 2017).

We believe that the achieved main findings are generalized to a clinical class of passive systems (*i.e.*, clinicians are making the final decisions) in which the end ratios of workload regarding both FPs and FNs are lower. In critical systems, it is more important to analyze the severity of consequences. However, the severity of consequences shall be applied to different clinical errors rather than medical imaging workload improvements.

### 8.3. Design implications

The main findings have implications for different stages of the design of an intelligent agent for CDSS solutions. From data collection across model training to the design of UI for AI systems, the implications of design are influencing the final proposed solution of this work. In this

section, we describe several design implications and recommendations applied until now, as well as the ones that will be applied in the future.

#### 8.3.1. Data collection

Developing an AI system for providing explanations across patient cases would require that structured information is given in the training data. On the one hand, several approaches are addressing the problem of collecting unstructured medical data (Schaekermann et al., 2019a; Schaekermann, Mike, 2020) with open-ended arguments for classification cases. On the other hand, recent works are demonstrating that imposing structure in the data collection process can facilitate a deeper understanding of clinician disagreement with AI (Schaekermann et al., 2019a) and accelerate consensus formation (Schaekermann et al., 2020a). Thus, a data collection procedure is recommended for AI-based systems by being equipped with structured procedures to benefit from these findings and facilitate the development of AI-assistance for medical imaging.

#### 8.3.2. Model training

This study suggests that medical workflows and trust can be positively affected by the introduction of intelligent agents which are endowing AI-based CDSS with the ability to not only make BI-RADS classification suggestions, but also to identify potential lesions. Implementation of such AI systems would require that supervised ML models are equipped with additional prediction targets (*e.g.*, automatic segmentation and classification of patient co-variables) beyond severity classification (BI-RADS) alone. These additional targets could include automatic classification of breast density, while warning clinicians the model accuracy for being in the presence of these cases. These another additional target is showing information concerning the weights of each co-variable for the achieved classification. Additional targets could be integrated either into one joint training process or by developing several separate models, one for each target. Mehta et al. (2018) describes how to bring a multi-target approach into the classification and segmentation of medical images in one joint training process. These design implications were brought from the focus groups, but were not yet addressed under this thesis proposal. However, they will be addressed as future work.

#### 8.3.3. User interface considerations

In this paper, several ways of displaying and explaining breast cancer diagnosis were evaluated by visually showing explainable (XAI). These explainable (XAI) techniques are coming from the trained models and are informing clinicians in several ways. While the achieved results may suggest that these representations should be effective, it is recommended further future work.

Further directions could explore more complex techniques for proper information visualization and design considerations. For instance, it is important to understand if heatmaps are effective and final ways for the representation of lesion contours with respect (colors) to the severity classification, or if there are other techniques (*e.g.*, bounding-boxes) to better inform clinicians. Moreover, it could be important to study whereas the design for information visualization of model performance would support a better patient diagnostic.

UI design considerations are facilitating and constraining the communication between human users and AI models (HAII). The design should be carefully decided whether and how the information is exposed to the end-users, so that the end-users can strike the right balance for a trusted, efficient and effective interaction. Clinicians' trust can be promoted by highlighting and explaining instances (*i.e.*, co-variables) of important patient information as predicted by an AI model. However, factors like domain-specific tolerance may affect whether exposing wrong patient information contributes to the establishment or erosion of clinicians' trust.

A predicted likelihood of showing AI accuracy for a given case can be a useful criterion for clinicians. Actually, by choosing which AI-suggestion to review first, the UI design is saving clinicians' time and cognitive resources (*e.g.*, not showing heatmaps immediately) while diagnosing each patient. In fact, the given model explanations should be human-interpretable, case-specific, and accurate.

An inaccurate or irrelevant explanation may significantly harm a clinician's ability to diagnose a patient case correctly. As a consideration, if accurate explanations cannot be reliably produced, they should not be exposed to the end-user. Hence, low accuracy classification must be actively warned to the clinician about the inaccurate final result.

Impact of explanation is a factor that should determine whether clinicians are exposed to explainable (XAI) information. For an efficient interaction, it is recommended that clinicians should not be exposed to granular explainable information for patient cases where accuracy interpretation cannot yield straight decision-making outcomes. As an example, if the model accuracy is low for a specific co-variable classification, the assistant should omit that co-variable, since it was not properly classified. Thus, it will be of chief importance to study at what level of accuracy are clinicians accepting each co-variable to be less accurately classified.

#### 8.3.4. Medical imaging perspectives

Due to its multi-view and multi-modal nature, the assistant uses large amounts of data to inform clinicians. The very first step in both quantitative and qualitative studies was to extract from our study information a set of analysis – a human-centered design methodology – which was crucial to answer our design choices. Specifically, data clustering resulted in the following clusters: (i) MG images (both CC and MLO views); (ii) US images; and (iii) MRI volumes. Note that in (i) a large number of views are available, *e.g.*, ML, LM, LMO, late ML, among others. Concerning (iii), radiologists acquire a large set of MRI (*e.g.*, T1, T2, Diffusion, DCE) volumes (Seifabadi et al., 2019). This initial design choices resulted in a consensus that allowed the selection of CC and MLO views, in the case of MG images, and MRI volumes.

### 8.4. Limitations

Unfortunately, due to time constrains of clinicians, it was not possible to investigate the impact of other AI-assistance techniques. Furthermore, the assistant represents a type of system that is passive and casual (Kocielnik et al., 2019), where the impact of AI imperfections in the clinical workflow is arguably critical. Breast cancer screening may only be effective if clinicians accept the new diagnostic paradigm with the introduction of useful, yet imperfect, AI techniques. Despite of this limitation, the achieved results are indicating that the proposed assistant may successfully introduce behavioral changes into the clinical workflow.

Another important point to follow are the limitations of DL systems. These methods are typically seen as black-boxes (Litjens et al., 2017) and, therefore, difficult to "explain" without more varied data (Schaekermann et al., 2020b). In this work, it was only addressed the BI-RADS classification problem and *heatmap* segmentation features. However, clinical co-variables are also important to better inform and explain to

clinicians the DL results. Additionally, it is important to understand the best communication strategies to inform clinicians about the importance of each co-variable in the final DL result. Thus, such concerns are addressed in the future work.

### 8.5. Future work

As a future work, the paper proposes the application of a model based on the UTAUT (Venkatesh et al., 2016) to study the adoption of AI systems in medical imaging diagnosis. The idea is to test the model via confirmatory factor analysis and structural equation modeling while using clinicians' responses to a formulated UTAUT questionnaire. Findings will provide valuable contributions to HCI and AI researchers concerning the design and implementation of intelligent agents. Future directions will analyse the critical behaviour and implement persuasive mechanisms to reduce the rates of FPs and FNs. The goal will be to understand how the level of assertiveness (Pacheco and Martinho, 2019; Paradeda et al., 2019) will impact the clinicians' decision-making process during breast cancer diagnosis. A proof-of-concept prototype will be developed with two major scenarios of assistant behaviour: (1) Assertive; and (2) Non-Assertive; paired with two assistant behaviours: (i) Proactive; and (ii) Reactive.

Future developments must follow the literature implementation of explainable (XAI) and interpretability methods suggested by research works (Tjoa and Guan, 2020). The various methods will show to clinicians different dimensions in interpretability research, from providing "obviously" explainable and interpretable information on complex patterns of lesions and patient co-variables. However, it will be also important to evaluate the quality of explanations and levels of interpretability given by the introduction of these new XAI methods. Therefore, the System Causability Scale (SCS) (Holzinger et al., 2020) will support future studies of this thesis. Image data availability is an important hurdle for the implementation of AI in the clinical setting. Thus, a dataset will be published including not only clinical data, but also user results. Clinical data will be published containing important information to the ML algorithms, such as medical images, classifications and segmentations. Finally, user results will be also published containing usability (SUS) and workload (NASA-TLX) measures, as well as trust and causability (SCS), between other future metrics to be applied.

## CRediT authorship contribution statement

**Francisco Maria Calisto:** Conceptualization, Methodology, Software, Data curation, Writing - original draft, Formal analysis, Investigation, Visualization. **Carlos Santiago:** Validation, Formal analysis, Data curation, Writing - review & editing. **Nuno Nunes:** Conceptualization, Methodology, Resources, Writing - review & editing, Supervision, Funding acquisition. **Jacinto C. Nascimento:** Conceptualization, Methodology, Writing - review & editing, Supervision, Project administration.

## Declaration of Competing Interest

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

## Appendix A. Appendix

### A1. DenseNet steps

The DenseNet steps are as follows:

1. The DenseNet is pre-trained with the ImageNet model that contains roughly 1.2 millions images. This gives generalization capabilities to new datasets.
2. Next, we remove the last layer of the DenseNet (that contains a large number of classes, about 1000 classes or end-nodes) and replace it by a fully connected layer containing now three nodes. Each node of the network corresponds to one of the following classes: (i) low severity (Class 1: BIRADS $<= 1$), (ii) medium severity (Class 2: $1 <$ BIRADS $<= 3$) and (iii) high severity (Class 3: BIRADS $> 3$). Thus, we have a DenseNet with three nodes in the last layer.
3. Now, the pre-trained DenseNet goes through a process of fine-tuning in our breast dataset. The dataset comprises 1125 training images (*i.e.*, MG in both CC and MLO views, US, and MRI). Notice that here, we have 338 cases or patients (Section 5.2). The larger number of training images, when compared to the number of patients, stems from the fact that one patient may have more than one image.
4. We split the above dataset in the following way: 80% for training and the remaining 20% for testing. In this partition, we guarantee that each set (training and test) are properly balanced, that is, each set contains samples that belong to all classes (recall that there are three classes as mentioned in Section 5.2, see also point 1) and all modalities (*i.e.*, MG, US, and MRI). Therefore, the experiments take into consideration the multimodality.
5. The pre-training follows a supervised learning strategy. Specifically, during the training we provide the image, as well as the corresponding label, that is, the classification in one of the three classes for the image. This classification (*i.e.*, ground-truth) is provided by a set of eight radiologists (Section 5). Say that this training stage gives to the DenseNet the ability to establish associations between the morphology of the lesion (image) and its severity (classification label).
6. Finally, we use the held-out test set to measure the performance of the DenseNet. An accuracy of 98.2% was obtained in this test set.
7. As a final note, we should mention that the three patients (as mentioned in Section 5.1) are obtained from the test set as mentioned above (see point 4). From the above the number of test samples is high, however, only three samples are taken for the experiments purpose. Thus, the training and testing phases take into account a large number of samples.

**Table A1**
This table represents the demographic data of our participants. The **ID** is the participants identifier for each clinician, *e.g.*, Clinician 1 (C1), Clinicia 2 (C2),..., Clinician 45 (C45). The **Group** represents the professional medical experience of clinicians (*i.e.*, Intern, Junior, Middle or Senior) and the **Speciality** represents the medical stage of speciality. The **Medical Experience** represents the number of years working as a clinician or as an internship of medical studies. The **Education** is the clinicians' level of education or background. The **Work Sector** represents what type of place and the **Institution** is the respective working host institution, as well as where the user tests where led for this study. (mimbcd-ui.github.io/dataset-uta7-demographics).

| ID | Group | Speciality | Medical Experience | Education | Work Sector | Institution |
|---|---|---|---|---|---|---|
| 1 | Senior | Head of Radiology Director | more than 10 years of speciality | Medicinae Doctor (M.D.) | Public | Hospital Fernando Fonseca |
| 2 | Intern | Medical General Internship | doing medical internship | Bologna Master Degree | Public, Private | Hospital Fernando Fonseca |
| 3 | Intern | Radiology Internship | did a medical internship as Gynecologist and Oncology | Post-Bologna Degree and Bologna Master Degree | Public, Social | Hospital Fernando Fonseca |
| 4 | Intern | Medical General Internship | doing medical internship | Medicinae Doctor (M.D.) | Public | Hospital Fernando Fonseca |
| 5 | Intern | Medical General Internship | doing medical internship | Bologna Master Degree | Public | Hospital Fernando Fonseca |
| 6 | Intern | Medical General Internship | doing medical internship | Bologna Master Degree | Public | Hospital Fernando Fonseca |
| 7 | Junior | Radiologist | a specialist with less than 5 years of speciality | Medicinae Doctor (M.D.) | Public | Hospital Fernando Fonseca |
| 8 | Senior | Radiologist | more than 10 years of speciality | Bologna Doctoral Degree (PhD) | Public, Private | Hospital Fernando Fonseca |
| 9 | Junior | Radiologist | a specialist with less than 5 years of speciality | Medicinae Doctor (M.D.) | Public | Hospital Fernando Fonseca |
| 10 | Senior | Surgeon | more than 10 years of speciality | Medicinae Doctor (M.D.) | Public | Hospital de Santa Maria |
| 11 | Middle | Radiologist, Senology, Mastology | between 5 to 10 years of speciality | Pre-Bologna Degree and Specialist | Private | SAMS Hospital |
| 12 | Middle | Immunotherapist | between 5 to 10 years of speciality | Medicinae Doctor (M.D.) | Private | Madeira Medical Center |
| 13 | Middle | Head of Radiology Director | between 5 to 10 years of speciality | Medicinae Doctor (M.D.) | Public | IPO Coimbra |
| 14 | Middle | Radiology Coordinator | between 5 to 10 years of speciality | Medicinae Doctor (M.D.) | Public | IPO Lisboa |
| 15 | Intern | Radiology Internship | doing medical internship | Bologna Master Degree | Public | Hospital do Barreiro |
| 16 | Intern | Radiology Internship | doing medical internship | Medicinae Doctor (M.D.) | Public | IPO Lisboa |
| 17 | Middle | Radiologist | between 5 to 10 years of speciality | Medicinae Doctor (M.D.) | Public | IPO Lisboa |
| 18 | Intern | Medical General Internship | doing medical internship | Medicinae Doctor (M.D.) | Public | IPO Lisboa |
| 19 | Middle | Radiologist | between 5 to 10 years of speciality | Medicinae Doctor (M.D.) | Public | IPO Lisboa |
| 20 | Senior | Radiologist | more than 10 years of speciality | Medicinae Doctor (M.D.) | Public | IPO Lisboa |
| 21 | Senior | Radiologist | more than 10 years of speciality | Medicinae Doctor (M.D.) | Public | IPO Coimbra |
| 22 | Middle | Radiologist | between 5 to 10 years of speciality | Medicinae Doctor (M.D.) | Public | IPO Coimbra |
| 23 | Middle | Radiologist | between 5 to 10 years of speciality | Medicinae Doctor (M.D.) | Public | IPO Coimbra |
| 24 | Middle | Radiologist | between 5 to 10 years of speciality | Medicinae Doctor (M.D.) | Public | IPO Coimbra |
| 25 | Middle | Radiologist | between 5 to 10 years of speciality | Medicinae Doctor (M.D.) | Public | IPO Coimbra |
| 26 | Middle | Radiologist | between 5 to 10 years of speciality | Medicinae Doctor (M.D.) | Public | IPO Coimbra |
| 27 | Senior | Radiologist | more than 10 years of speciality | Medicinae Doctor (M.D.) | Public | IPO Coimbra |
| 28 | Junior | Radiologist | a specialist with less than 5 years of speciality | Medicinae Doctor (M.D.) | Public | Hospital Fernando Fonseca |
| 29 | Senior | Radiologist | more than 10 years of speciality | Medicinae Doctor (M.D.) | Public | Hospital Fernando Fonseca |
| 30 | Middle | Radiologist | between 5 to 10 years of speciality | Medicinae Doctor (M.D.) | Public | Hospital Fernando Fonseca |
| 31 | Intern | Radiology Internship | doing medical internship | Medicinae Doctor (M.D.) | Public | IPO Lisboa |
| 32 | Junior | Radiologist | a specialist with less than 5 years of speciality | Medicinae Doctor (M.D.) | Public | Hospital Fernando Fonseca |
| 33 | Intern | Radiology Internship | doing medical internship | Bologna Master Degree | Public | Hospital do Barreiro |
| 34 | Middle | Radiologist | between 5 to 10 years of speciality | Medicinae Doctor (M.D.) | Public | Hospital do Barreiro |
| 35 | Junior | Radiologist | a specialist with less than 5 years of speciality | Medicinae Doctor (M.D.) | Public | Hospital do Barreiro |
| 36 | Junior | Radiologist | a specialist with less than 5 years of speciality | Medicinae Doctor (M.D.) | Public | Hospital do Barreiro |
| 37 | Intern | Radiology Internship | doing medical internship | Bologna Master Degree | Public, Private | Hospital do Barreiro |
| 38 | Intern | Radiology Internship | doing medical internship | Medicinae Doctor (M.D.) | Public | Hospital do Barreiro |
| 39 | Senior | Radiologist | more than 10 years of speciality | Medicinae Doctor (M.D.) | Public | Hospital do Barreiro |
| 40 | Senior | Radiologist | more than 10 years of speciality | Medicinae Doctor (M.D.) | Public | Hospital do Barreiro |
| 41 | Senior | Radiologist | more than 10 years of speciality | Medicinae Doctor (M.D.) | Public | IPO Lisboa |
| 42 | Junior | Oncologist | a specialist with less than 5 years of speciality | Medicinae Doctor (M.D.) | Public | Hospital de Santa Maria |
| 43 | Middle | Radiologist | between 5 to 10 years of speciality | Medicinae Doctor (M.D.) | Public | IPO Lisboa |
| 44 | Junior | Radiologist | a specialist with less than 5 years of speciality | Bologna Master Degree | Public | IPO Lisboa |
| 45 | Senior | Radiologist | more than 10 years of speciality | Medicinae Doctor (M.D.) | Public | IPO Coimbra |

## References


Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M., 2018. Trends and trajectories for explainable, accountable and intelligible systems: an HCI research agenda. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 582:1–582:18. https://doi.org/10.1145/3173574.3174156.

Aerts, H.J., 2017. Data science in radiology: a path forward. Clin. Cancer Res. 24 (3), 2804.

Aghaei, F., Mirniaharikandehei, S., Hollingsworth, A.B., Stoug, R.G., Pearce, M., Liu, H., Zheng, B., 2018. Association between background parenchymal enhancement of breast MRI and BIRADS rating change in the subsequent screening. Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications. SPIE, Houston, Texas, United States, pp. 1–8.

Alkhatib, A., Bernstein, M., 2019. Street-level algorithms: a theory at the gaps between policy and decisions. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 530:1–530:13. https://doi.org/10.1145/3290605.3300760.

Amershi, S., Cakmak, M., Knox, W.B., Kulesza, T., 2014. Power to the people: the role of humans in interactive machine learning. AI Mag. 35 (4), 105–120.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P.N., Inkpen, K., Teevan, J., Kikin-Gil, R., Horvitz, E., 2019. Guidelines for human-ai interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 3:1–3:13. https://doi.org/10.1145/3290605.3300233.

Becker, A.S., Marcon, M., Ghafoor, S., Wurnig, M.C., Frauenfelder, T., Boss, A., 2017. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. Invest. Radiol. 52 (7), 434–440.

Bonham, M., 2019. Augmented reality simulation toward improving therapeutic healthcare communication techniques. Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion. ACM, New York, NY, USA, pp. 161–162. https://doi.org/10.1145/3308557.3308726.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A., 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA A Cancer J. Clin. 68 (6), 394–424. https://doi.org/10.3322/caac.21492.

Bruno, M.A., Walker, E.A., Abujudeh, H.H., 2015. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radiographics 35 (6), 1668–1676.

Cai, C.J., Jongejan, J., Holbrook, J., 2019. The effects of example-based explanations in a machine learning interface. Proceedings of the 24th International Conference on Intelligent User Interfaces. ACM, New York, NY, USA, pp. 258–262. https://doi.org/10.1145/3301275.3302289.

Cai, C.J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., Wattenberg, M., Viegas, F., Corrado, G.S., Stumpe, M.C., Terry, M., 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 4:1–4:14. https://doi.org/10.1145/3290605.3300234.

Calisto, F.M., 2017. Medical imaging multimodality breast cancer diagnosis user interface. Master's thesis, Avenida Rovisco Pais 1, 1049-001 Lisboa - Portugal (EU). A Medical Imaging Tool for a Multimodality use of Breast Cancer Diagnosis on a User Interface.

Calisto, F.M., 2019. Assistant Introduction: User Testing Guide For A Comparison Between Multi-Modality and AI-Assisted Systems. Technical Report. Instituto Superior Técnico. https://doi.org/10.13140/RG.2.2.16566.14403/1.

Calisto, F. M., 2019b. It-medex closing workshop: towards touch-based medical image diagnosis annotation. 10.13140/RG.2.2.30479.43682.

Calisto, F.M., 2020. Breast Cancer Medical Imaging Multimodality Lesion Contours Annotating Method. Technical Report. Instituto Superior Técnico, Avenida Rovisco Pais 1, 1049-001 Lisboa - Portugal (EU). https://doi.org/10.13140/RG.2.2.14792.55049. Method and process using a system to annotate and visualize masses and microcalcifications of breast cancer lesions in a multimodality strategy.

Calisto, F.M., 2020. Medical imaging multimodality annotating framework. PhD Open Days 2020. Instituto Superior Técnico, pp. 1–2. https://doi.org/10.13140/RG.2.2.16086.88649.

Calisto, F.M., Ferreira, A., Nascimento, J.C., Gonçalves, D., 2017. Towards touch-based medical image diagnosis annotation. Int'l Conf. Interactive Surfaces and Spaces (ISS). ACM, New York, NY, USA, pp. 390–395. https://doi.org/10.1145/3132272.3134111.

Calisto, F.M., Lencastre, H., Nunes, N.J., Nascimento, J.C., 2019. BreastScreening: towards breast cancer clinical decision support systems. National Science Summit 2019. Fundação para a Ciência e Tecnologia, pp. 1–2. https://doi.org/10.13140/RG.2.2.25718.65606.

Calisto, F.M., Lencastre, H., Nunes, N.J., Nascimento, J.C., 2019. Medical imaging diagnosis assistant: AI-assisted radiomics framework user validation. Keep In Touch 2019. Instituto Superior Técnico, pp. 1–2. https://doi.org/10.13140/RG.2.2.33421.59360.

Calisto, F.M., Miraldo, P., Nunes, N.J., Nascimento, J.C., 2018. BreastScreening: a multimodality diagnostic assistant. LARSyS 2018 Annual Meeting. Interactive Technologies Institute, pp. 1–2. https://doi.org/10.13140/RG.2.2.29816.70409.

Calisto, F. M., Nascimento, J. C., 2018. Medical imaging multimodality breast cancer diagnosis user interface: Breast imaging reporting and data system (BI-RADS) survey template file. 10.13140/RG.2.2.36306.86725.

Calisto, F.M., Nunes, N., Nascimento, J.C., 2020. BreastScreening: on the use of multi-modality in medical imaging diagnosis. Proceedings of the International Conference on Advanced Visual Interfaces. Association for Computing Machinery, New York, NY, USA, pp. 1–5. https://doi.org/10.1145/3399715.3399744.

Calisto, F.M., Nunes, N.J., Nascimento, J.C., Miraldo, P., 2018. BreastScreening: a multimodality diagnostic assistant. National Science Summit 2018. Fundação para a Ciência e Tecnologia, pp. 1–2. https://doi.org/10.13140/RG.2.2.25412.68486.

Carneiro, G., Nascimento, J., Bradley, A.P., 2017. Automated analysis of unregistered multi-view mammograms with deep learning. Trans. Med. Imaging 36 (11), 2355–2365.

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., Tsaneva-Atanasova, K., 2019. Artificial intelligence, bias and clinical safety. BMJ Qual. Saf. 28 (3), 231–237.

Chatelain, P., Sharma, H., Drukker, L., Papageorghiou, A.T., Noble, J.A., 2018. Evaluation of gaze tracking calibration for longitudinal biomedical imaging studies. IEEE Trans. Cybern. (99), 1–11.

Cheung, Y.-C., 2017. Integral multimodality imaging in breast cancer diagnosis. Ultrasound Med. Biol. 43, S17.

Choy, G., Khalilzadeh, O., Michalski, M., Do, S., Samir, A.E., Pianykh, O.S., Geis, J.R., Pandharipande, P.V., Brink, J.A., Dreyer, K.J., 2018. Current applications and future impact of machine learning in radiology. Radiology 288 (2), 318–328.

De Backere, F., Verstichel, S., Van den Bergh, J., Elprama, S.A., Ongenae, F., De Turck, F., Jacobs, A., Coninx, K., 2015. Discovery of the potential role of sensors in a personal emergency response system: what can we learn from a single workshop? Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare. Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, Brussels, Belgium, pp. 330–333.

Delvaux, N., Van Thienen, K., Heselmans, A., de Velde, S.V., Ramaekers, D., Aertgeerts, B., 2017. The effects of computerized clinical decision support systems on laboratory test ordering: a systematic review. Arch. Pathol. Lab. Med. 141 (4), 585–595.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: a large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp. 248–255.

DeSantis, C.E., Fedewa, S.A., Goding Sauer, A., Kramer, J.L., Smith, R.A., Jemal, A., 2016. Breast cancer statistics, 2015: convergence of incidence rates between black and white women. CA Cancer J. Clin. 66 (1), 31–42.

Eslami, M., Karahalios, K., Sandvig, C., Vaccaro, K., Rickman, A., Hamilton, K., Kirlik, A., 2016. First i "like" it, then i hide it: folk theories of social feeds. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 2371–2382. https://doi.org/10.1145/2858036.2858494.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542 (7639), 115.

Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D., Forman, D., Bray, F., 2013. Cancer Incidence and Mortality Worldwide: IARC Cancer Base No. 11. International Agency for Research on Cancer, France.

Gagnon, M.-P., Ghandour, E.K., Talla, P.K., Simonyan, D., Godin, G., Labrecque, M., Ouimet, M., Rousseau, M., 2014. Electronic health record acceptance by physicians: testing an integrated theoretical model. J. Biomed. Inf. 48, 17–27.

Gale, W., Oakden-Rayner, L., Carneiro, G., Bradley, A. P., Palmer, L. J., 2017. Detecting hip fractures with radiologist-level performance using deep neural networks. arXiv preprint arXiv:1711.06504.

Gambino, A., Kim, J., Sundar, S.S., 2019. Digital doctors and robot receptionists: user attributes that predict acceptance of automation in healthcare facilities. Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. LBW0287:1–LBW0287:6. https://doi.org/10.1145/3290607.3312916.

Ghahramani, Z., 2015. Probabilistic machine learning and artificial intelligence. Nature 521 (7553), 452–459.

Graffy, P.M., Liu, J., O'Connor, S., Summers, R.M., Pickhardt, P.J., 2019. Automated segmentation and quantification of aortic calcification at abdominal CT: application of a deep learning-based algorithm to a longitudinal screening cohort. Abdominal Radiol. 1–8.

Greenspan, H., van Ginneken, B., Summers, R.M., 2016. Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. Trans. Med. Imaging 35 (5), 1153–1159.

Grier, R.A., 2015. How high is high? A meta-analysis of NASA-TLX global workload scores. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 59. SAGE Publications Sage CA: Los Angeles, CA, pp. 1727–1731.

Gunning, D., 2017. Explainable Artificial Intelligence (XAI). Defense Advanced Research Projects Agency (DARPA), nd Web.

Harboe, G., Minke, J., Ilea, I., Huang, E.M., 2012. Computer support for collaborative data analysis: augmenting paper affinity diagrams. Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work. ACM, New York, NY, USA, pp. 1179–1182. https://doi.org/10.1145/2145204.2145379.

Harrington, H.J., 2016. Affinity diagrams. The Innovation Tools Handbook, Vol. 2. Productivity Press, pp. 45–54.

Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, M., Schnabel, J.A., 2012. MIND: modality independent neighbourhood descriptor for multi-modal deformable registration. Med. Image Anal. 16 (7), 1423–1435.

Høiseth, M., Giannakos, M.N., Alsos, O.A., Jaccheri, L., Asheim, J., 2013. Designing healthcare games and applications for toddlers. Proceedings of the 12th International Conference on Interaction Design and Children. ACM, New York, NY, USA, pp. 137–146. https://doi.org/10.1145/2485760.2485770.

Høiseth, M., Giannakos, M.N., Jaccheri, L., 2013. Research-derived guidelines for designing toddlers' healthcare games. CHI '13 Extended Abstracts on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 451–456. https://doi.org/10.1145/2468356.2468436.

Holzinger, A., 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inf. 3 (2), 119–131.
Holzinger, A., Carrington, A., Müller, H., 2020. Measuring the quality of explanations: the system causability scale (SCS): comparing human and machine explanations. Kunstliche Intelligenz 34 (2), 193–198.
Holzinger, A., Kieseberg, P., Weippl, E., Tjoa, A.M., 2018. Current advances, trends and challenges of machine learning and knowledge extraction: from machine learning to explainable AI. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (Eds.), Machine Learning and Knowledge Extraction. Springer International Publishing, Cham, pp. 1–8.
Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H., 2019. Causability and explainability of artificial intelligence in medicine. WIREs Data Min. Knowl. Discov. 9 (4), e1312. https://doi.org/10.1002/widm.1312.
Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H., Aerts, H.J., 2018. Artificial intelligence in radiology. Nat. Rev. Cancer 18 (8), 500.
Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708.
Jodogne, S., 2018. The Orthanc ecosystem for medical imaging. J. Digit. Imaging 31 (3), 341–352. https://doi.org/10.1007/s10278-018-0082-y.
Ker, J., Wang, L., Rao, J., Lim, T., 2018. Deep learning applications in medical image analysis. Access 6, 9375–9389.
Khairat, S., Marc, D., Crosby, W., Al Sanousi, A., 2018. Reasons for physicians not adopting clinical decision support systems: critical analysis. JMIR Med. Inf. 6 (2), e24.
Khan, S., Islam, N., Jan, Z., Din, I.U., Rodrigues, J.J.C., 2019. A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. Pattern Recognit. Lett. 125, 1–6.
Kocielnik, R., Amershi, S., Bennett, P.N., 2019. Will you accept an imperfect ai?: Exploring designs for adjusting end-user expectations of ai systems. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 411:1–411:14. https://doi.org/10.1145/3290605.3300641.
Kohli, A., Jha, S., 2018. Why cad failed in mammography. J. Am. College Radiol. 15 (3), 535–537.
Kooi, T., Litjens, G., Van Ginneken, B., Gubern-Mérida, A., Sánchez, C.I., Mann, R., den Heeten, A., Karssemeijer, N., 2017. Large scale deep learning for computer aided detection of mammographic lesions. Med. Image Anal. 35, 303–312.
Kumar, D., Chung, A.G., Shaifee, M.J., Khalvati, F., Haider, M.A., Wong, A., 2017. Discovery radiomics for pathologically-proven computed tomography lung cancer prediction. In: Karray, F., Campilho, A., Cheriet, F. (Eds.), Image Analysis and Recognition. Springer International Publishing, Cham, pp. 54–62.
Lakhani, P., Sundaram, B., 2017. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 284 (2), 574–582.
Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R.G., Granton, P., Zegers, C.M., Gillies, R., Boellard, R., Dekker, A., et al., 2012. Radiomics: extracting more information from medical images using advanced feature analysis. Eur. J. Cancer 48 (4), 441–446.
Leung, L., Chen, C., 2019. E-health/m-health adoption and lifestyle improvements: exploring the roles of technology readiness, the expectation-confirmation model, and health-related information activities. Telecommun. Policy.
Lewis, J.R., 2018. The system usability scale: past, present, and future. Int. J. Hum.-Comput.Interact. 34 (7), 577–590.
Liang, S., Tang, F., Huang, X., Yang, K., Zhong, T., Hu, R., Liu, S., Yuan, X., Zhang, Y., 2019. Deep-learning-based detection and segmentation of organs at risk in nasopharyngeal carcinoma computed tomographic images for radiotherapy planning. Eur. Radiol. 29 (4), 1961–1967.
Liikkanen, L.A., 2017. The data-driven design era in professional web design. Interactions 24 (5), 52–57.
Lim, B., Rogers, Y., Sebire, N., 2019. Designing to distract: can interactive technologies reduce visitor anxiety in a children's hospital setting&. ACM Trans. Comput.-Hum. Interact. 26 (2), 9:1–9:19. https://doi.org/10.1145/3301427.
Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.
Maicas, G., Nguyen, C., Motlagh, F., Nascimento, J. C., Carneiro, G., 2019. Unsupervised task design to meta-train medical image classifiers. arXiv preprint arXiv:1907.07816.
Mathews, A., Marc, D., 2017. Usability evaluation of laboratory information systems. J. Pathol. Inf. 8.
McKay, D., Makri, S., Chang, S., Buchanan, G., 2020. On birthing dancing stars: the need for bounded chaos in information interaction. Proceedings of the 2020 Conference on Human Information Interaction and Retrieval. Association for Computing Machinery, New York, NY, USA, pp. 292–302. https://doi.org/10.1145/3343413.3377983.
McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.C., Darzi, A., et al., 2020. International evaluation of an ai system for breast cancer screening. Nature 577 (7788), 89–94.
Mehta, S., Mercan, E., Bartlett, J., Weaver, D., Elmore, J.G., Shapiro, L., 2018. Y-Net: joint segmentation and classification for diagnosis of breast biopsy images. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Springer International Publishing, Cham, pp. 893–901.
Middleton, B., Sittig, D., Wright, A., 2016. Clinical decision support: a 25 year retrospective and a 25 year vision. Yearbook Med. Inf. 25 (S 01), S103–S116.
Miglioretti, D.L., Smith-Bindman, R., Abraham, L., Brenner, R.J., Carney, P.A., Bowles, E. J.A., Buist, D.S., Elmore, J.G., 2007. Radiologist characteristics associated with interpretive performance of diagnostic mammography. J. Natl. Cancer Inst. 99 (24), 1854–1863.
Miller, T., 2018. Explanation in artificial intelligence: insights from the social sciences. Artif. Intell.
Mohamed, A.A., Berg, W.A., Peng, H., Luo, Y., Jankowitz, R.C., Wu, S., 2018. A deep learning method for classifying mammographic breast density categories. Med. Phys. 45 (1), 314–321.
Mullie, L., Afilalo, J., 2019. CoreSlicer: a web toolkit for analytic morphomics. BMC Med. Imaging 19 (1), 15.
Murtaza, G., Shuib, L., Wahab, A.W.A., Mujtaba, G., Nweke, H.F., Al-garadi, M.A., Zulfiqar, F., Raza, G., Azmi, N.A., 2019. Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges. Artif. Intell. Rev. 1–66.
Pacheco, A.C., Martinho, C., 2019. Alignment of player and non-player character assertiveness levels. Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment. AAAI, Georgia Institute of Technology, Atlanta, Georgia, USA, pp. 181–187.
Paradeda, R., Ferreira, M.J., Oliveira, R., Martinho, C., Paiva, A., 2019. The role of assertiveness in a storytelling game with persuasive robotic non-player characters. Proceedings of the Annual Symposium on Computer-Human Interaction in Play. Association for Computing Machinery, New York, NY, USA, pp. 453–465. https://doi.org/10.1145/3311350.3347162.
Park, S.Y., Chen, Y., Rudkin, S., 2015. Technological and organizational adaptation of EMR implementation in an emergency department. ACM Trans. Comput.-Hum. Interact. 22 (1), 1:1–1:24. https://doi.org/10.1145/2656213.
Raghu, M., Zhang, C., Kleinberg, J., Bengio, S., 2019. Transfusion: Understanding transfer learning for medical imaging. In: Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc., pp. 3347–3357
Raja, K., Patrick, M., Elder, J.T., Tsoi, L.C., 2017. Machine learning workflow to enhance predictions of adverse drug reactions (ADRs) through drug-gene interactions: application to drugs for cutaneous diseases. Sci Rep 7 (1), 3690.
Ramkumar, A., Stappers, P.J., Niessen, W.J., Adebahr, S., Schimek-Jasch, T., Nestle, U., Song, Y., 2017. Using GOMS and NASA-TLX to evaluate human–computer interaction process in interactive segmentation. Int. J. Hum.-Comput.Interact. 33 (2), 123–134.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.
Rosset, A., Spadola, L., Ratib, O., 2004. OsiriX: an open-source software for navigating in multidimensional DICOM images. J. Digit. Imaging 17 (3), 205–216.
Ruddle, R.A., Thomas, R.G., Randell, R., Quirke, P., Treanor, D., 2016. The design and evaluation of interfaces for navigating gigapixel images in digital pathology. ACM Trans. Comput.-Hum. Interact. 23 (1), 5:1–5:29. https://doi.org/10.1145/2834117.
Saadatmand, S., Bretveld, R., Siesling, S., Tilanus-Linthorst, M.M., 2015. Influence of tumour stage at breast cancer detection on survival in modern times: population based study in 173 797 patients. Bmj 351, h4901.
Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Muller, K.-R., 2019. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, first ed. Springer Publishing Company, Incorporated.
Sarcevic, A., Marsic, I., Burd, R.S., 2012. Teamwork errors in trauma resuscitation. ACM Trans. Comput.-Hum. Interact. 19 (2), 13:1–13:30. https://doi.org/10.1145/2240156.2240161.
Savage, N., 2019. Digital assistants aid disease diagnosis. Nature 573 (7775), S98–S99. https://doi.org/10.1038/d41586-019-02870-4.
Schaekermann, M., Beaton, G., Habib, M., Lim, A., Larson, K., Law, E., 2019. Capturing expert arguments from medical adjudication discussions in a machine-readable format. Companion Proceedings of The 2019 World Wide Web Conference. Association for Computing Machinery, New York, NY, USA, pp. 1131–1137.
Schaekermann, M., Beaton, G., Habib, M., Lim, A., Larson, K., Law, E., 2019. Understanding expert disagreement in medical data analysis through structured adjudication. Proc. ACM Hum.-Comput. Interact. 3 (CSCW) https://doi.org/10.1145/3359178.
Schaekermann, M., Beaton, G., Sanoubari, E., Lim, A., Larson, K., Law, E., 2020. Ambiguity-aware ai assistants for medical data analysis. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 1–14. https://doi.org/10.1145/3313831.3376506.
Schaekermann, M., Cai, C.J., Huang, A.E., Sayres, R., 2020. Expert discussions improve comprehension of difficult cases in medical image assessment. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 1–13.
Schaekermann, M., Hammel, N., Terry, M., Ali, T.K., Liu, Y., Basham, B., Campana, B., Chen, W., Ji, X., Krause, J., Corrado, G.S., Peng, L., Webster, D.R., Law, E., Sayres, R., 2019. Remote tool-based adjudication for grading diabetic retinopathy. Translational Vision Science and Technology 8 (6). https://doi.org/10.1167/tvst.8.6.40.40–40
Schaekermann, M., Law, E., Larson, K., Lim, A., 2018. Expert disagreement in sequential labeling: a case study on adjudication in medical time series analysis. SAD/CrowdBias HCOMP, pp. 55–66.
Schaekermann, Mike, 2020. Human-ai interaction in the presence of ambiguity: from deliberation-based labeling to ambiguity-aware AI.
Seifabadi, R., Cheng, A., Malik, B., Kishimoto, S., Wiskin, J., Munasinghe, J., Negussie, A. H., Bakhutashvili, I., Krishna, M.C., Choyke, P., et al., 2019. Correlation of ultrasound tomography to MRI and pathology for the detection of prostate cancer.

Medical Imaging 2019: Ultrasonic Imaging and Tomography, Vol. 10955. International Society for Optics and Photonics, p. 109550C.
Shah, P., Kendall, F., Khozin, S., Goosen, R., Hu, J., Laramie, J., Ringel, M., Schork, N., 2019. Artificial intelligence and machine learning in clinical development: a translational perspective. NPJ Digit. Med. 2 (1), 69.
Shin, H.-C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M., 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging 35 (5), 1285–1298.
Sonntag, D., Schulz, C., Reuschling, C., Galarraga, L., 2012. Radspeech's mobile dialogue system for radiologists. Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces. ACM, New York, NY, USA, pp. 317–318. https://doi.org/10.1145/2166966.2167031.
Sorace, A.G., Partridge, S.C., Li, X., Virostko, J., Barnes, S.L., Hippe, D.S., Huang, W., Yankeelov, T.E., 2018. Distinguishing benign and malignant breast tumors: preliminary comparison of kinetic modeling approaches using multi-institutional dynamic contrast-enhanced MRI data from the international breast mr consortium 6883 trial. J. Med. Imaging 5 (1), 011019.
Subramonyam, H., Drucker, S.M., Adar, E., 2019. Affinity lens: data-assisted affinity diagramming with augmented reality. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 1–13. https://doi.org/10.1145/3290605.3300628.
Sultanum, N., Brudno, M., Wigdor, D., Chevalier, F., 2018. More text please! Understanding and supporting the use of visualization for clinical text overview. Conf. Human Factors in Computing Systems (CHI). ACM, New York, NY, USA, pp. 1–13. https://doi.org/10.1145/3173574.3173996.
Szolovits, P., 2019. Artificial Intelligence in Medicine. Routledge.
Tjoa, E., Guan, C., 2020. A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans. Neural Netw. Learn.Syst. 1–21. https://doi.org/10.1109/TNNLS.2020.3027314.
Topol, E.J., 2019. High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. 25 (1), 44.
Torre, L.A., Bray, F., Siegel, R.L., Ferlay, J., Lortet-Tieulent, J., Jemal, A., 2015. Global cancer statistics, 2012. CA Cancer J. Clin. 65 (2), 87–108.
Tyllinen, M., Kaipio, J., Lääveri, T., Nieminen, M.H., 2016. We need numbers!: Heuristic evaluation during demonstrations (HED) for measuring usability in it system procurement. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 4129–4141. https://doi.org/10.1145/2858036.2858570.
Urban, T., Ziegler, E., Lewis, R., Hafey, C., Sadow, C., Van den Abbeele, A.D., Harris, G.J., 2017. LesionTracker: extensible open-source zero-footprint web viewer for cancer imaging research and clinical trials. Cancer Res. 77 (21), 119–122.
Venkatesh, V., Thong, J.Y., Xu, X., 2016. Unified theory of acceptance and use of technology: a synthesis and the road ahead. J. Assoc. Inf. Syst. 17 (5), 328–376.
Wagner, S., Beckmann, M.W., Wullich, B., Seggewies, C., Ries, M., Bürkle, T., Prokosch, H.-U., 2015. Analysis and classification of oncology activities on the way to workflow based single source documentation in clinical information systems. Med. Inf. Decis. Making 15 (107), 1–13.
Waite, S., Kolla, S., Jeudy, J., Legasto, A., Macknik, S.L., Martinez-Conde, S., Krupinski, E.A., Reede, D.L., 2017. Tired in the reading room: the influence of fatigue in radiology. J. Am. Coll. Radiol. 14 (2), 191–197.
Wang, J., Yang, X., Cai, H., Tan, W., Jin, C., Li, L., 2016. Discrimination of breast cancer with microcalcifications on mammography by deep learning. Sci. Rep. 6, 27327.
Weese, J., Lorenz, C., 2016. Four challenges in medical image analysis from an industrial perspective. Med. Image Anal. 33, 44–49.
Welch, H.G., Prorok, P.C., O'Malley, A.J., Kramer, B.S., 2016. Breast-cancer tumor size, overdiagnosis, and mammography screening effectiveness. N. Engl. J. Med. 375 (15), 1438–1447.
Wernli, K.J., Ichikawa, L., Kerlikowske, K., Buist, D.S., Brandzel, S.D., Bush, M., Johnson, D., Henderson, L.M., Nekhlyudov, L., Onega, T., et al., 2019. Surveillance breast MRI and mammography: comparison in women with a personal history of breast cancer. Radiology 182475.
Wilde, D., Vallgårda, A., Tomico, O., 2017. Embodied design ideation methods: analysing the power of estrangement. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 5158–5170. https://doi.org/10.1145/3025453.3025873.
Wobbrock, J.O., Findlater, L., Gergle, D., Higgins, J.J., 2011. The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 143–146. https://doi.org/10.1145/1978942.1978963.
Wolf, I., Vetter, M., Wegner, I., Böttger, T., Nolden, M., Schöbinger, M., Hastenteufel, M., Kunert, T., Meinzer, H.-P., 2005. The medical imaging interaction toolkit. Med. Image Anal. 9 (6), 594–604.
Yang, Q., 2019. Profiling artificial intelligence as a material for user experience design. Microsoft Research. Ph.D. thesis.
Yang, Q., Banovic, N., Zimmerman, J., 2018. Mapping machine learning advances from HCI research to reveal starting places for design innovation. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 1–11. https://doi.org/10.1145/3173574.3173704.
Yang, Q., Zimmerman, J., Steinfeld, A., Carey, L., Antaki, J.F., 2016. Investigating the heart pump implant decision process: opportunities for decision support tools to help. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 4477–4488. https://doi.org/10.1145/2858036.2858373.
Yu, C.-S., Lin, Y.-J., Lin, C.-H., Lin, S.-Y., Wu, J.L., Chang, S.-S., 2020. Development of an online health care assessment for preventive medicine: a machine learning approach. J. Med. Internet Res. 22 (6), e18585. https://doi.org/10.2196/18585.
Zhang, X., Pina, L.R., Fogarty, J., 2016. Examining unlock journaling with diaries and reminders for in situ self-report in health and wellness. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 5658–5664. https://doi.org/10.1145/2858036.2858360.