MIDA
User Testing Guide
Multi-Modality vs *AI-Assisted*
Assistant Introduction

Francisco Maria Calisto
francisco.calisto@tecnico.ulisboa.pt

07/03/2019

Prototype: prototype-multi-modality **Version:** v1.0.1-alpha
Milestone: 1.2.0-beta **Release:** v1.2.0-beta

Prototype: prototype-multi-modality-assistant **Version:** v1.2.0-alpha
Milestone: 1.2.0-alpha **Release:** v1.2.0-alpha

DICOM: dicom-server
Commit: 80191e9941c24043c7f612b2dadcd415c060bf96

Depl. Env.: Localhost **Depl. Server:** Localhost
Link: breastscreening.io/dashboard

Main Server: Localhost **Port:** 8486
Private IP: localhost **Public IP:** localhost
Private Domain: localhost

DICOM Server: Localhost **Port:** 8448
Private IP: localhost **Public IP:** localhost
Private Domain: localhost **From:** 8448

1 Introduction

This document [15] aims to describe the protocol and guidelines of the presented information. We perform a set of tests in the scope of v1.0.1-alpha, v1.2.0-alpha and v1.2.0-alpha versions from *Multi-Modality (MM)*, *Assistant* and *Heatmap* prototypes, respectively. The repositories are part of the MIDA project using traditional devices (mouse and keyboard). The goal of the test is to compare each prototype, measuring the user performance, efficiency and efficacy metrics. The sessions will be recorded via video on a computer and using an interaction record, while triggering event tools. It is guaranteed the confidentiality of the recordings, which will be used only for academic purpose. Also, we will use an eye-tracking device to track the participant's eye movements during the breast cancer diagnosis.

Dividing the activity session into three distinct phases (**Pha*N.***, where *N.* is the nth number of a limited series, $n \in \mathbb{N} \ \forall \ N = \{1, 2, 3\}$) per each three activities representing two different scenarios: *MM* (**Sce1.**) vs *Assistant* (**Sce2.**). The first two phases took place on an early stage of the *User Tests*, while we were focus to publish the results on a near future. The third phase, will cover the hereby *User Testing Guide*. Still, we will describe, as follows, the overall of the three phases to give higher contextualization.

Each scenario (**Sce*M.***, where *M.* is the mth number of a limited series, $m \in \mathbb{N} \ \forall \ M = \{1, 2\}$) will have three random patients (*i.e.*, **Pat1.**, **Pat2.** or **Pat3.**) from a set of 74 total number of patients. For each patient (**Pat*P.***, where *P.* is the pth number of a limited series, $p \in \mathbb{N} \ \forall \ P = \{1, 2, 3\}$), we will choose it from the total set of patients randomly. We do it as follows, let R be a random variable (**Rdm**) following the discrete uniform distribution as $r_1, r_2, r_3 \in \mathbb{N} \ \forall \ R = \{1, ..., 74\}$. While $Pat1. = Rdm_{r_1} \wedge Pat2. = Rdm_{r_2} \wedge Pat3. = Rdm_{r_3}$ as far as $r_1 \neq r_2 \neq r_3$ is *True*. In both two scenarios, *i.e.*, *MM* (**Sce1.**) and *Assistant* (**Sce2.**), by supporting our traditional devices, the interaction is made with mouse and keyboard. Participants will classify each patient by using the BIRADS [5]. We will do several small questionnaires at the end of each scenario using NASA-TLX [56], SUS [54], DOTS [48] and measuring the breast severity, as well as patients' pathologies.

Describing each phase, the first phase is nominated as **Pha1. User Characterization**. It is the Demographic Questionnaire, supporting our several characterizations of the participant profile. For the major number of participants, this phase was committed on an early stage, as stated above. The **Pha1.** phase has three activities (**Act*A.***, where *A.* is the ath number of a limited series, $a \in \mathbb{N}$) which are described as follows. The first activity, named as **Act1. Consent Form**, serves the purpose of providing participants information about privacy of the data and accept to proceed the test. The second activity, named as **Act2. Study Introduction**, serves the purpose of giving participants project contextualization and task awareness. Last but not least, the third activity of **Pha1.** phase, called as **Act3. Demographic Questionnaires**, is where participants fill a set of surveys regarding their characterization as a participant.

The second phase, nominated as **Pha2. Improving Visualization**, is also related with our later User Test Evaluations, corresponding to the tests done for the *MM* (**Sce1.**) scenario. At this **Pha2.** we divided it into two activities: (1) **Act4. Scenario Introduction** activity; and (2) **Act5. Scenario Evaluation** activity. Both applied for the **Sce1.** scenario.

The first **Act4.** activity, aims at providing participants information about what tasks they will do. Each task represents the diagnostic of each patient, while we named it as **Act5. Scenario Evaluation** activity of the **Sce1.** scenario. An intermediate **Act6. Scenario Exploration** was created, aiming at providing participants a free opportunity to explore the system and give opinion in regard for **Pha2.** phase. It will be important information for the *Qualitative Analysis (QA)*. Finally, the **Act7. Post-task Scenario Questions**, of the **Sce1.** scenario, will be a set of *post-task* questions for **Pha2.** phase.

Finally, and most importantly, the third phase, nominated as **Pha3. Assistant Establishment** phase, is where participants proceed for the diagnosis of the respective three, *i.e.*, **Pat1.**, **Pat2.** and **Pat3.** patients. It is here, where we will also verify if our proposed designs impact [2, 41] on user expectations, as intended, specifically as outlined by our several *Hypotheses* for the respective *Research Questions* (Section 6). We did it in regard to the novel introduction of an *AI-Assisted* system, calling it as *BreastScreening* [10, 61].

On the **Pha3.** phase, we will measure the participants expectations concerning the *AI-Assisted* system applying an **Act8. Scenario Expectations** activity for the **Sce2.** scenario. The next **Act4. Scenario Introduction** activity of the **Sce2.** scenario, will be the introductory information about *AI-Assisted* system and what feature are covered by the system. The **Act5. Scenario Evaluation** activity of the **Sce2.** scenario, represents the diagnostic of each patient, but this time with support of our novel *AI-Assisted* recommendations and explainability. Another intermediate **Act6. Scenario Exploration** of the **Sce2.** scenario, in regard for the **Pha3.** phase, where also, it will generate important information for our *QA*. Finally, the **Act7. Post-task Scenario Questions** of the **Sce2.** scenario, will be a set of *post-task* questions.

For this *User Testing Guide*, we used three prototype repositories, *i.e.*, *MM*, *Assistant* and *Heatmap* prototypes. The prototypes are similar mirrors of the prototype-breast-screening with major changes. It is further described.

The first, *i.e.*, *MM* prototype, aims at providing clinicians a *MM* strategy view. The *MM* view, gives clinicians the possibility for visualizing three modalities: (i) MammoGraphy (MG), both CranioCaudal (CC) and MedioLateral Oblique (MLO) views; UltraSound (US); and Magnetic Resonance Imaging (MRI). It corresponds to the **Sce1.** scenario of both **Act4.**, **Act5.** and **Act6.** activities at **Pha2.** phase.

Second and third repositories, *i.e.*, from both prototype-multi-modality-assistant and prototype-heatmap repositories, aims at providing clinicians a recommendation system regarding our *AI-Assistive* techniques. Those techniques, will provide clinicians a twofold: (a) the opportunity of receive automatic recommendations concerning breast severities (BIRADS) of the patients; and (b) giving clinicians explainability (XAI) [31, 32] of those results.

The automatic recommendations will be covered by the prototype-multi-modality-assistant repository, while the explainability will be covered by the prototype-heatmap repository. Both techniques are corresponding to the **Sce2.** scenario of **Act4.**, **Act5.**, **Act6.**, **Act7.** and **Act8.** activities at the **Pha3.** phase.

2 Description

This document describes our test plan for conducting our user tests during the development of the BreastScreening project and systems. The goals of the user testing phases include establishing a baseline of participant performance, establishing and validating participant performance measures, and identifying potential design concerns to be addressed in order to improve the efficiency, productivity, and end-user satisfaction within the development and introduction of *AI-Assistive* methods inside the Radiology Room (RR), between others, for Medical Imaging (MI), or more precisely, the breast cancer diagnosis.

The user test objectives are:

1. To determine design inconsistencies and issues within the UI and content areas;
   (a) **Navigation Errors**;
   (b) **Presentation Errors**;
   (c) **Control Usage Problems**;
2. Exercise the prototype under controlled test conditions with representative users;
3. Establish baseline user performance and user-satisfaction levels of the user interface for future usability evaluations;

Potential sources of error may include: (a) **Navigation Errors:** failure to locate functions, excessive keystrokes to complete a function, failure to follow recommended screen flow; (b) **Presentation Errors:** failure to locate and properly act upon desired information in screens, selection errors due to labeling ambiguities; and (c) **Control Usage Problems:** improper toolbar or entry field usage. Data will be used to access whether usability goals regarding an effective, efficient, and well-received user interface have been achieved.

To verify our work, we identified measurable and explicit targets. By having several goals, including that a value percentage of the users should be able to operate the tasks without the need of help. On the same rate value, the user should be able to start and complete the medical diagnosis tasks over the system with little errors or mitigating those errors. Measuring the expected number of errors with a relation between pilot early tests. On the laboratory pilot tests we aim to test our prototypes with researchers.

Researchers are in the context of the system, and know well the functionalities, so that, we need to expect a percentage value over their results compared to clinicians and not the same benefits. Last but not least, both users (researchers and clinicians) should be able to understand in a similar time amount the meaning of all visible controls. By the similar amount of time, it is expected to have a variance of the percentage value between researchers and clinicians, as well as values of early tests, of the same value percentage of the early goals described in this paragraph.

We tested each objective in early laboratory and field tests, so that we could take the appropriate corrective actions. Also, we expect to run early field tests with researchers and clinicians to highlight issues that we overlooked and ignored during the prototyping phase. To support interaction use by the participants, we will try to emphasize several key factors on our user tests. The tasks must be simple, low intrusive, support for natural interaction and the system must always give visibility and the task current-state.

3 Methodology

The hereby used prototypes are both v1.2.0-alpha and v1.2.0-alpha versions of our *Assistant* and *Heatpmap* prototypes, respectively. The purpose of these prototypes is to involve an *AI-Assisted* tool (*Assistant*) for medical imaging at a breast screening diagnosis level[40]. This *Assistant* was created with a front-end and back-end architecture utilizing common programming languages, libraries, frameworks and tools including JavaScript (JS) [29], NodeJS [67], HammerJS, CornerstoneJS [33] and Orthanc [36]. For the Machine Learning (ML) and Deep Learning (DL) [58, 59] component we will use several MATLAB technologies [65], promoting and feeding our Convolutional Neural Networks (CNN) [21] and Deep Reenforcement Learning (DRL) [45] techniques. Other central component of this prototype is a web-based PACS [23] pairwise with ubicous web technologies and based on the **Open Source (OS)** CornerstoneJS library [28, 33].

3.1 Environments

This section describes the user environment over interaction, the so called **Radiology Room (RR)** (Figure 1). This guide is based on soft-copy diagnosis using computer workstations in their current reading room environment. It will be here where we take impressions regarding the efficacy of radiologists, and their recommendations based on their experience for improvements on the soft-copy reading environment. Several studies demonstrated [66] that radiologist fatigue levels and performance are related to environmental factors such as number of False-Negative (FN) and False-Positives (FP), in addition to workstation enhancements. Supported by this guide, our research aims to conduct an investigation for the several environmental variables and improvements regarding the potentially enhancement that an *AI-Assisted* diagnosis could take in the **RR** [26, 50].

Figure 1: Radiology Room

3.2 Participants

The participants' responsibilities will be attempting to complete a set of representative task scenarios (Section 7) presented to them in as efficient and timely a manner as possible, and to provide feedback regarding the usability and acceptability of an *AI-Assisted* diagnosis. The participants will be directed to provide honest opinions regarding the user tests of the interacted systems, and to participate in post-session subjective questionnaires and debriefing.

3.3 Procedure

Participants will take part in the tests at our formed institution protocols (*e.g.*, Hospital Fernando Fonseca (HFF)) with both v1.2.0-alpha and v1.2.0-alpha versions of our prototype-multi-modality-assistant and prototype-heatmap repositories, respectively. The interaction with the system will be used in a typical **RR** environment. Note takers and data logger(s) will monitor the sessions for observation in the **RR**, connected by screen recording feed. The test sessions will be recorded and further analyzed.

The facilitator will brief the participants on the system features and instruct the participant that they are evaluating the system, rather than the facilitator evaluating the participant. Participants will sign an informed consent (**Pha1. - Act1.**) that acknowledges: the participation is voluntary, that participation can cease at any time, and that the session will be videotaped and eye tracked but their privacy of identification will be safeguarded. The facilitator will ask the participant if they have any questions.

Participants will complete a pre-test demographic (**Pha1.** - **Act3.**) and background information (**Pha1.** - **Act2.**) questionnaires. The facilitator will explain that the amount of time taken to complete the test task, will be measured and that exploratory behavior outside the task flow should not occur until after task completion. At the start of each task, the participant will listen the task description from the printed copy and begin the task. Time-on-Task (ToT) measurement [57] begins when the participant starts the task.

The facilitator will instruct the participant to "think aloud" [6, 39] so that a verbal record exists of their interaction with the system. The facilitator will observe and enter user behavior, user comments, and system actions. Before each task, participants will complete a pre-task (**Pha2.** - **Act8.**) questionnaire. During each task, participants will elaborate on the task session (**Pha2.** - **Act5.** and **Act6.**) and complete the post-task (**Pha2.** - **Act7.**) questionnaires with the facilitator. After all task scenarios are attempted, the participant will complete several *open-ended questions* [1, 49].

3.4 Briefing

A presentation of the *Assistant* and its use and capabilities will be made. Participants will be presented to the available interactions and will be explained how to interact with the prototype, underlining the limitations. The facilitator will brief the participants on the *Assistant* system and instruct the participant that they are evaluating the system, rather than the facilitator evaluating the participant. Participants will sign an informed consent that acknowledges: the participation is voluntary, that participation can cease at any time, and that the session will be videotaped and eye gaze monitorization but their privacy of identification will be granted. The facilitator will ask the participant if they have any question.

3.5 Training

The participants will receive and overview of the user test procedure, equipment and system. The facilitator will show how to interact with the system and what features are available. We choose this approach, as it provide clinicians the most important concepts to understand and interact with our system. Also, it is of chief importance to give clinicians information of what is and is not available analysis of our *Assistant* and what it can do.

4 Roles

The roles involved in our user tests are as follows. An individual may play multiple roles, as well as the test may not require all roles.

4.1 Trainer

- Provide training overview prior to user testing phases;

4.2 Facilitator

- Provides overview of study to participants;
- Defines tasks and purpose of the user testing to participants;
- Assists in conduct of participant and observer debriefing sessions;
- Responds to participant's requests for assistance;

4.3 Data Logger

- Records participant's actions and comments;

4.4 Test Observers

- Silent observer;
- Assists the data logger in identifying problems, concerns, coding bugs and procedural errors;
- Serve as note takers;

4.5 Ethics

All persons involved with the Usability (Usa.) test are required to adhere to the following ethical guidelines:

- The performance of any test participant must not be individually attributable;
- Individual participant's name should not be used in reference outside the testing session;
- A description of the participant's performance should not be reported to his or her superior;

5 Apparatus

For the material and apparatus, it is essential to capture the session apprehending the user interactions. In our case, we will record this interaction by using the QuickTime Player Version 10.4 (928.5.1) to obtain all interactions. We will pair this video tool with a user watch tool called Hotjar and an eye tracker. The Hotjar tool serves the purpose of using several logs of the interaction and gives us visualization over it. The eye tracker will be the Tobii Eye Tracker 4C (SN: IS404-100107008875) device. All instruments will help us to capture where and when are users interacting. By looking at the test participant's reactions, we find a lot of information regarding the prototype design.

The tools that we choose for the material and apparatus of this User Testing Guide are low-cost and easy to use. Our equipment is a cost-effective and, by using our laboratory materials, bringing it to the RR, we enable to capture not only what the user is doing on the screen, but on the body language supported by the interviews and observations.

The material used in the test sessions for the user interface consists of:

- MacBook Pro: it will allow the user to interact with the keyboard and a wireless mouse;
- Wireless Mouse: it will allow the user to interact with a mouse and will complement the keyboard;

5.1 Technical Details

To produce this traditional environment, and since we can simulate with a laptop, the mouse and keyboard interaction, we are using a Microsoft Mobile Mouse 4000 together with the MacBook Pro (Retina, 13-inch, Early 2016) with a standard integrated keyboard on the laptop.

5.2 Devices

Traditional interaction remains the most common way to interact with user interfaces in a clinical environment. Unfortunately, most of this interaction is made by low profile equipment that makes users produce more errors and take more time interacting with those User Interfaces (UI) [43, 55, 63].

For this *User Testing Guide*, we will use the Tobii Eye Tracker 4C (Tobii, Sweden). This device, is a remote eye tracking device that provides an estimate of the point-gaze and 3D eye position for each eye at 90Hz [25, 46, 47]. We will attach the eye tracker under the display area with a magnetic mounting bracket, following the product's instructions. A 9-point calibration [22] will be performed for each user. Our protocol was implemented in Python (*versions* $\geqslant$ v2.7) and Processing (*versions* $\geqslant$ v3.5.3). The setup repository was the eye-tracker-setup, while the repository to measure and calculate the setup gazing information was the eye-tracker-naive repository.

5.3 User Interactions

The systems have several buttons (Figure 2) that allows the user to interact or access to a set of user interface features. Each item of the following list represents each metaphoric icon of Figure 2.

Figure 2: Toolbar of the System available features.

The buttons are (from left to right of Figure 2) as follows:

- WW/WC
- Invert
- Zoom
- Pan
- Stack Scroll
- Freehand
- Probe
- Save
- Window Controller

On Figure 3, the user can select the list of patients. The list has a table with several patient information. The first column is the *Patient ID*; we used it as an identifier of the patient. In that way, we can have anonymized information with no reference to the patient name. The second column is the *Study Date*, the third column is the *Modality* of the used **DICOM** image, the fourth column is the *Study Description* of the used study and the last column is the number of *Images*.

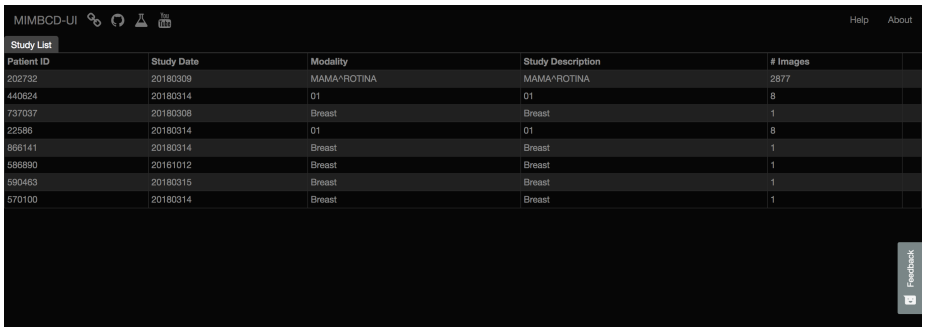

Figure 3: List of Patients.

As we can see in Figure 4, it shows the first task in our User Interface (UI), where the patient's breasts are on a small left column. The options are in a short row near of the viewport and described below. We also have the tabs where the user can change the patient. The centre viewport shows the **DICOM** image, and it can be configured to display a number up to four **DICOM** images at the same time. The viewport has some text information on it (yellow) with the details of the metadata. Nevertheless, the *Assistant* suggestions are shown on the top-right corner of the system. Our *Assistant* system allows a doctor-bot to provide a second opinion about the image severity of each patient. In other words, the *Assistant* is an Artificial Intelligence (AI) engine to interact with clinicians. It has an interface, wherein a representation of a doctor, presenting with a text box. The *Assistant* system is, therefore, made to diagnose and provide a second opinion for lesion severities of the breast cancer diseases.

Manual annotation [19] is adopted by us thanks to Freehand ROI and Probe annotation features, both from CornerstoneJS. According to the CornerstoneJS Library, the user can create an annotation by setting up consecutive landmarks around a Region of Interest (ROI). The markers finish a lesion annotation when it interconnects the last bullet point. Additional features, available in our User Interface (UI), includes on-demand increment of the number of landmarks, and throw transformations of the shape of an annotation. For the *Assistant* (**Sce2.**), we provide the recommendations of our *bot-like* system. This *bot-like* will give clinicians information regarding the patient's achieved severity of the breast (BIRADS), and the respective interpretation by text. The interpretation can be simple analysis of the patient's co-variables. A further analysis, when we show the heatmap answer (prototype-heatmap repository), will provide explainability to clinicians concerning the lesion severities across each image.

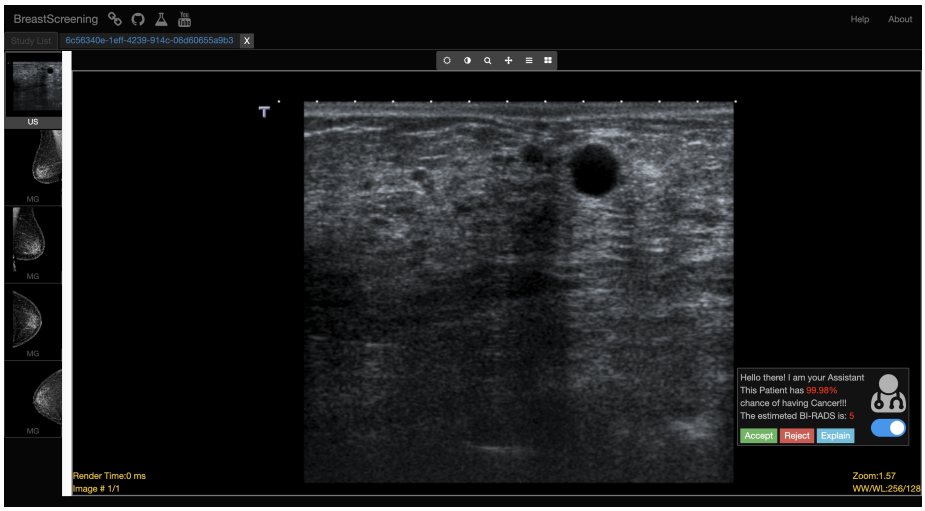

Figure 4: Viewer of the **DICOM** images.

5.4 Software

To track our user interactions across our system, we are using Hotjar. This tool is an analytic package allowing us to follow our users remotely. It also provides two critical pieces of functionality, among others, that can aid in remote user testing. First of all, the frequency areas allow us to see where users are clicking, tapping and scrolling on our system. Second, it records a video playback of the entire user session. The tool shows evidence of being useful for our studies while we successfully used it in the past. To record the task activities and the interview, we used QuickTime [60]. The QuickTime (Apple Computer) tool is available on MacBook Pro for movie, audio and screen recording. Despite of have an overall of features, we just used it for our user's screen recording. It provides this functionalities at minimum requirements and compatible to our apparatus. Finally, we will take advantage of the Tobii Pro SDK [22], providing us the gaze information of the eye tracking device.

6 Evaluation

Introduction of *AI-Assistive* agents are significant factors which can naturally affect the performance of a medical workflow. While some prior studies [9, 10, 11] have investigated the functionality of healthcare systems, the *AI-Assisted* acceptability has mostly been overlooked in the existing Health Informatics (HI) literature regarding a Human-Computer Interaction (HCI) research.

The following Table 1 is presenting three main *Research Questions* to have in mind during evaluation. The purpose of this questions is to facilitate systematic user studies [53] regarding our novel *Assistant* in a clinical environment and support user stimulation for the introduction of *AI-Assisted* methods. The proposed issues, involve various aspects of workflow combined with, either need for satisfaction, nor division of attention.

Number	Research Questions
RQ1.	What is the impact of an *AI* system for avoiding different types of errors on clinician perception?
RQ2.	What are the design techniques for setting appropriate clinician expectations of *AI* systems?
RQ3.	What is the impact of expectation-setting intervention techniques on satisfaction and acceptance of *AI* topic?

Table 1: Research Evaluation Questions

The influence of *AI-Assisted* [30] is an important variable for our empirical analysis. In fact, we expect that the trust of the user will increase when the user perceived that the *Assistant* is giving the right inputs and that there will be a consequent increase of the clinician trust in our system. We also want to measure (Section 8) that our *Assistant* can operate at the same level of overall accuracy, *i.e.*, total number of correct predictions divided by all possible predictions. Measuring predictions is typically quantified as precision in contrast with recall. We therefore explore the above *Research Questions* and associated each to a set of *Hypotheses* following this [2, 41] authors instructions. The first work [2], describes a set of 18 guidelines for Human-AI Interaction (HAII) being highly useful to answer the **RQ2.** question, and respective hypothesis mapping each with the guidelines. The second work [41], developed by almost the same team, provide us an exploratory study of an *Assistant* to study the impact of several methods of expectation-setting, also answering the **RQ2.** question. In both studies, the authors show that different focus on avoiding types of errors lead to a vastly different subjective perceptions (*i.e.*, the **RQ1.** question) of accuracy and acceptance (*i.e.*, the **RQ3.** question).

List of associated *Research Questions* to respective set of *Hypotheses*:

1. **RQ1.** What is the impact of an *AI* system for avoiding different types of errors on clinician perception ?
   (a) **H1.1.** An *AI* system focused on *High Precision* will result in higher perceptions of accuracy.
   (b) **H1.2.** An *AI* system focused on *High Precision* will result in higher acceptance?
2. **RQ2.** What are the design techniques for setting appropriate clinician expectations of *AI* systems ?
   (a) **H2.1.** An *AI* system that directly is communicating it accuracy to clinicians will reduce the lack between system accuracy and user perception.
   (b) **H2.2.** Providing clinicians explanations (XAI) [31, 32] will lead to higher perception of understanding how the *AI* system works.
   (c) **H2.3.** A first clinician contact with the system will lead to higher perceived level of control over the *AI* results.
3. **RQ3.** What is the impact of expectation-setting intervention techniques on satisfaction and acceptance of *AI* ?
   (a) **H3.1.** In the mediation of an imperfect *AI* system providing clinicians the power of prior interventions will lead to higher acceptance and satisfaction in comparison to a lack of such interventions.

For the first question, enumerated as **RQ1.**, we want to explore the impact of our *AI* system avoinding errors in regard with clinicians' perception. We will explore how the system, focused on *High Precision*, will result in higher perceptions of accuracy (**H1.1.**) and higher acceptance (**H1.2.**). We mapped [2] the **H1.1.** with **G1**, **G2** and **G3** guidelines. On the other hand, we mapped the **H1.2.** with **G5**, **G6** and **G7** guidelines.

The second question, enumerated as **RQ2.**, the prior work of the authors [41] show us three major contributions to clinician's expectations: (1) information from external sources (**H2.1.**); (2) reasoning and understanding (**H2.2.**); and (3) first hand experience (**H3.3.**). From here, our second *Research Question* explores design techniques for achieving these mechanism pairwise with the work done by the same team on another one [2]. Again, we also associated each of the three *Hypotheses* with the set of guidelines [2]. First of all, the **H2.1.** was mapped with **G2**, **G12**, **G15**, **G16** and **G18** guidelines. Second, the **H2.2.** was mapped with **G2**, **G3**, **G4** and **G11** guidelines. And thirdly, the **H2.3.** was mapped with **G2**, **G13**, **G14** and **G17** guidelines.

Finally, for the third question, enumerated as **RQ3.**, we will expect that a more accurate expectations of an *AI* system's capabilities should result in clinicians being better prepared for *AI* system imperfections and, therefore, result in higher satisfaction and acceptance. For this question *Hypotheses*, enumerated as **H3.1.**, we mapped it with **G8**, **G9** and **G10** guidelines.

Several other questions are in our mind, and could be addressed on both current and future work. However, the following questions will not be our focus for this work. For instance, we can ask about *How would the user describe the potential adoption of AI-Assisted methods on the Health Institution?* to obtain answers in regard. For a second question, the *What are the user oppositions for AI-Assisted methods?* question, we aim to understand what are the user constrains regarding an *AI* adoption the the user's current workflow. Third, we could intend to filter possible examples of the clinical applications of *AI* on the Health Institutions by asking *What examples of AI-Assisted methods does the user know regarding the Health Institution?* directly to the clinician. A fourth question, could underline the reasons why several obstacles are present on the Health Institution, with the question *What are the obstacles of the user's Health Institution?* we can understand the challenges of achieving those issues and what are the solutions for surpass it. On a fifth question, where we could ask *What is more important for the AI-Assisted information, the BIRADS or Pathology?*, we aim to understand what is more important for the user, the BIRADS or the Pathology [44] of the patient [27]. Almost last, a six question, where we could ask for *Is it important for the user to have the feature of* ***Approve***, ***Reject*** *and* ***Explain*** *options?* is a valuable question to understand the feature needs and options. Last but not least, on a seven question, *For the user's opinion, what are the aspects that influence the decision?*, is where we could understand what is the most important information to show to the clinicians, therefore, we can more effectively and efficiently give more accurate information to the users.

To conclude this section, by answering this questions, we aim to support our user studies giving our users, the clinicians, an opportunity of improving our empirical analysis regarding user's *open answers*. However, the results should be treated with caution. Several bias exists since we are doing here an ambiguous approach.

7 Tasks

During our user tests, we need to ask participants to provide a subjective assessment of their experience using our *Assistant*. There are several widely used questionnaires giving us different prons-and-cons. However, in most cases, a single question instrument [62] is the right method for a quantitative usability testing. By taking less time and effort to answer, participants are pursuing to this phase after task, while it is minimally disruptive.

The tasks were derived from test scenarios (*i.e.*, **Sce1.** and **Sce2.**) developed from use cases and/or with the assistance of a subject-matter expert. Due to the range and extent of functionality provided in the system, and the short time for which each participant will be available, the tasks are the most common and relatively complex for the available functions. The tasks are identical for all participants of a given user role in the study.

At this stage, each participant will interact with our UI. The UI show to each participant the set of patients chosen randomly (Section 1). The set of patients will provide participants with 50 patients, while all patients much have at least one of the three available modalities (Figure 5). Each participant will open each patient (*e.g.*, **Pat1.**, **Pat2.** or **Pat3.**), of the three patients chosen randomly, and will examine the set of images. During examination, the participant will interact with the available features (Section 5) of the system.

We will try to understand if, with the *AI-Assisted* techniques, the clinicians will encounters the most accurate severity (BIRADS) of the breast lesions [52] and patient's prognostic. For this purpose, we have several patients (Figure 5); each patient has several images in the respective modalities: (i) MG; (ii) UltraSound (US); and (iii) MRI. The clinicians will proceed to the activity of diagnosing three random patients within the support of our *Assistant* by the observation of all images.

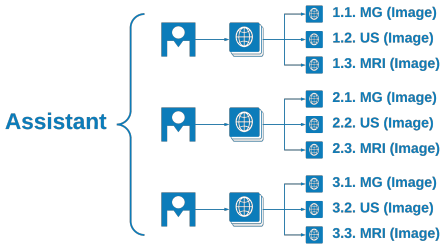

Figure 5: Diagram representing the use of the *Assistant* by clinicians. Each patient can have, and is not limited to have more images per modality.

In our *User Testing Guide* a set of tasks is necessary and carefully crafted. Our test studies involve asking participants to perform a set of tasks. By looking at what our user need to do with our system, our tasks are realistic as possible. We are not describing the exact steps participants need to take. We achieve that by avoiding the precise language used as labels in our system. The tasks are emotionally neutrals. And we did several pilot tests to prevent misleading situations saving us from wasting resources by accidentally use a lousy task or from getting bad data. The tasks are as follows.

The task descriptions below are required to be reviewed by all researchers and facilitators (Section 4) to ensure that the content, format, and presentation are representative of real use and substantially evaluate the total system. Their acceptance is to be documented prior to the user test. Each task, is related to the set of *Phases*, *Scenarios* and *Activities* (Section 1) on the next section (Section 8) further explained.

List of stand alone tasks for both **Pha1.**, **Pha2.** and **Pha3.** phases:

Task 1.1.1: Fill the Consent Form (Section 1) and accept the user test;

Task 1.1.2: Fill the User Characterization Form (Section 1) and proceed;

Task 2.1.1: Classify *Patient 1* (**Pat1.**) on the *MM* condition;

Task 2.1.2: Classify *Patient 2* (**Pat2.**) on the *MM* condition;

Task 2.1.3: Classify *Patient 3* (**Pat3.**) on the *MM* condition;

Task 2.2.1: Freely explore *Patient 1* (**Pat1.**) on the *MM* condition;

Task 2.2.2: Freely explore *Patient 2* (**Pat2.**) on the *MM* condition;

Task 2.2.3: Freely explore *Patient 3* (**Pat3.**) on the *MM* condition;

Task 2.3.1: Work. & Usa. questions for the *MM* condition;

Task 2.3.2: *Post-task* questions for the *MM* condition;

Task 3.1.1: Classify *Patient 1* (**Pat1.**) on the *Assis.* condition;

Task 3.1.2: Classify *Patient 2* (**Pat2.**) on the *Assis.* condition;

Task 3.1.3: Classify *Patient 3* (**Pat3.**) on the *Assis.* condition;

Task 3.2.1: Freely explore *Patient 1* (**Pat1.**) on the *Assis.* condition;

Task 3.2.2: Freely explore *Patient 2* (**Pat2.**) on the *Assis.* condition;

Task 3.2.3: Freely explore *Patient 3* (**Pat3.**) on the *Assis.* condition;

Task 3.3.1: Work., Usa. & Trust questions for the *Assis.* condition;

Task 3.3.2: *Post-task* questions for the *Assis.* condition;

8 Metrics

Our user test metrics refers to user performance measured against specific performance goals necessary to satisfy the test requirements. Scenario completion success rates, adherence to dialog scripts, error rates, and subjective evaluations will be used. Time-to-Completion (TtC) [35] of scenarios will also be collected. From the set of tasks (Section 7), each task corresponds to the set of *Phases*, *Scenarios* and *Activities* (Section 1), meaning that we first need to explain it relations.

The first two tasks, *i.e.*, **Task 1.1.1** and **Task 1.1.2**, are related to the **Pha1.** phase, as well as with **Act1.**, **Act2.** and **Act3.** activities. The **Pha2.** phase focus on testing and analyzing the *MM* condition *MM*, *i.e.*, corresponding to **Sce1.** scenario. The six next tasks, *i.e.*, **Task 2.1.1**, **Task 2.1.2**, **Task 2.1.3**, **Task 2.2.1**, **Task 2.2.2** and **Task 2.2.3**, are related to **Pha2.** phase. Also, the **Pha2.** phase is related with **Act4.**, **Act5.** and **Act6.** activities, of the **Sce1.** scenario, for diagnosing **Pat1.**, **Pat2.** and **Pat3.** patients. The second next two tasks, *i.e.*, **Task 2.3.1** and **Task 2.3.2**, are related with **Act7.** activity of **Pha2.** phase. Now, on the **Pha3.** phase, we will have a relation between testing and analyzing the *Assistant (Assis.)* condition, *i.e.*, corresponding to **Sce2.** scenario. The six next tasks, *i.e.*, **Task 3.1.1**, **Task 3.1.2**, **Task 3.1.3**, **Task 3.2.1**, **Task 3.2.2** and **Task 3.2.3**, are therefore related to **Pha3.** phase. Also, the **Pha3.** phase is related with **Act4.**, **Act5.**, **Act6.** and **Act8.** activities, but this time of the **Sce2.** scenario, for diagnosing **Pat1.**, **Pat2.** and **Pat3.** patients. At the end, the last two tasks, *i.e.*, **Task 3.3.1** and **Task 3.3.2**, are related with **Act7.** activity for **Pha3.** phase.

8.1 Patient Classification

For the patient classification, we will use the well known scale for classifying the breast cancer disease called BIRADS [5]. The BIRADS scale is a scheme for putting the findings from breast into a small number of well-defined categories [51].

The BIRADS assessment categories are:

- 0 - Incomplete;
- 1 - Negative;
- 2 - Benign Findings;
- 3 - Probably Benign;
- 4 - Suspicious Abnormality;
- 5 - Highly Suspicious of Malignancy;
- 6 - Known Biopsy Proven Malignancy;

For each participant, we will ask the respective examination and respective BIRADS value. From here, we will register the respective value per each scenario, both *i.e.*, **Sce1.** and **Sce2.** scenarios. At the end, we can compare the values provided between **Sce1.** and **Sce2.** scenarios. On **Sce1.** scenario, it is where we just improve the visualization technique. Now, with **Sce2.** scenario, *i.e.*, an *AI-Assistance* diagnosis, we want to understand if the given severity value changed and improved. Also, on **Sce2.** scenario, we want to understand where, *i.e.*, on *Assistant* or *Heatmap* prototype, did the participant took the final improved answer. Nevertheless, we will also compare several other patients' variables, like pathology, to address several other clinical issues.

8.2 Workload

To measure the workload, we used the NASA Task Load Index (NASA-TLX) [56] scale. The scale is a subjective workload assessment tool that will allow us to perform subjective workload assessments on our participants. For the purpose, we created a repository [17, 13] to cover this need of content.

By incorporating a multi-dimensional rating procedure, NASA-TLX derives an overall workload score based on a weighted average of ratings on six subscales:

- Mental Demand
- Physical Demand
- Temporal Demand
- Performance
- Effort
- Frustration

At the end, each participant will provide answers regarding the workload information during **Act5.** activity, of **Sce1.** and **Sce2.** scenarios, on both **Pha1.** and **Pha2.** phases respectively. This will also cover both **Task 2.3.1** and **Task 3.3.1** tasks.

8.3 Usability

To measure the usability, we used the System Usability Scale (SUS) [54]. The SUS provides a "quick and dirty", reliable tool for measuring the usability. It consists of a 10 item questionnaire with ten response options for respondents; from *Strongly Agree* to *Strongly Disagree*. Originally created by John Brooke in 1986, it allows you to evaluate a wide variety of products and services, including hardware, software, mobile devices, websites and applications. For the purpose, we created a repository [16, 12] to cover this need of content.

When using SUS, participants are asked to score the following 10 items with one of ten responses that range from **Strongly Agree** to **Strongly Disagree**:

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

Again, each participant will provide answers regarding the system's usability information during **Act5.** activity, of **Sce1.** and **Sce2.** scenarios, from both **Pha1.** and **Pha2.** phases respectively. This will also cover both **Task 2.3.1** and **Task 3.3.1** tasks.

8.4 Trust

The Dimensions Of Trust Scale (DOTS) [14, 18] was introduced on a recent work [7, 8], introducing the concept of measuring trust across *AI* systems. For that, we created a repository [14] supporting our user tests. DOTS is a scale to measure the trustworthiness of our *AI-Assisted* system. Therefore, we created a three items list of questions on a 20-point scale of Likert-style [37]. The following list, represents the questions adapted from this model [48] addressing each of the three items, *i.e.*, *understanding*, *capability* and *benevolence* [7].

1. I understand what the system is thinking. (**Understanding**)
2. The system seems capable. (**Capability**)
3. The system seems benevolent. (**Benevolence**)

Each participant will provide answers regarding the system's trustworthiness during **Act5.** activity, of **Sce2.** scenario, from **Pha2.** phase. This will cover the **Task 3.3.1** from the list of tasks.

8.5 Predictions

To measure system predictions with purpose of comparing participants acceptance, we applied our own computational method. The computational method is as follows, while we defined several variables to it, defined next to this information and further explained. Let the *Overall Accuracy* [4, 42] be Ø, a variable following the discrete uniform distribution as $Ø \in \mathbb{R}$. The accuracy is used by us to measure how accurate is the overall performance of our solution, considering both positive and negative classes without worrying about data imbalance. Let *Total Number of Correct Predictions* [4, 42] be τ, a variable following the discrete uniform distribution as $\tau \in \mathbb{R}$. Let *All Possible Predictions* [4, 42] be α, a variable following the discrete uniform distribution as $\alpha \in \mathbb{R}$. As follows, we report our computational method.

Computational method to measure the *Overall Accuracy* of our solution:

$$Overall\ Accuracy = \frac{Total\ Number\ of\ Correct\ Predictions}{All\ Possible\ Predictions}$$

8.6 Eye Tracking

Eye movement data was collected for several groups of subjects recruited from the same Portuguese institutions, both public and private, with breast domain-expertise levels, while participants inspected (Section 7) breast images. Medical images will be presented to participants on a monitor (Section 5) attached to a 90Hz eye tracking device, called Tobii Eye Tracker 4C with reported accuracy for the collection of eye movement data.

For the eye tracking measurements, we will use the work done by Vaidyanathan et al. [64], titled as "*Recurrence Quantification Analysis Reveals Eye-Movement Behavior Differences between Experts and Novices*". The work uses eye movement data from medical experts and novices, while they inspected several medical images. Most importantly, the work describe and demonstrate how Recurrence Quantification Analysis (RQA) [3], and the associated measures, can be used to differentiate eye movement behavior during different viewing conditions and image type finding significant differences. From this work, we aim to use their RQA method to measure and quantify certain eye movement aspects, defined as: (1) Recurrence (REC); (2) Determinism (DET); (3) Laminarity (LAM); and (4) Center Of Recurrence Mass (CORM). We are not yet sure if will use all the presented information, however, the hereby *User Testing Guide* serves the purpose of presenting all options for the tests. At this point, we are concern with addressing and collecting the maximum user data as possible.

8.7 Qualitative Evaluation

Qualitative and subjective evaluations regarding ease of use and satisfaction will be collected. This collection will be done via *open-ended questions* [1, 49], and during debriefing at the conclusion of the session.

The *open-ended questions* will utilize free-form responses and feedback, when possible. Whenever possible, it's best to ask *open-ended questions* so we can find out more than we can anticipate. We will test our questions by trying to answer them with short answers, and rewrite those to find out more about *how* and *what*. In some cases, we won't be able to accommodate free-form or write-in answers, though, and then it is necessary to limit the possibilities.

8.8 Scenario Completion

Each scenario, *i.e.*, **Sce1.** and **Sce2.** scenarios, will require, or request, that the participant obtains, or inputs, specific data. This data would be used in course of a typical task. The scenario is completed when the participant indicates the scenario's goal has been obtained. Whether successfully or unsuccessfully. Or the scenario is completed when the participant requests and receives sufficient guidance as to warrant scoring the scenario as a critical error.

8.9 Time Completion

The time to complete (ToT) [24, 34] each scenario, *i.e.*, **Sce1.** and **Sce2.** scenarios, not including qualitative and subjective evaluation durations, will be recorded. From this measure, it will be also possible to collect more specific metrics, such as the percentage of time that participants follow an optimal path or the number of times participants need to backtrack.

8.10 Critical Errors

Critical Errors are deviations at completion from the targets of the scenario. Obtaining or otherwise reporting of the wrong data value due to participant workflow is a Critical Error. Participants may or may not be aware that the task goal is incorrect or incomplete.

An example of a Critical Error, could be a situation where the participant is not able to open a patient. From this error, we can not even proceed to the next tasks and complete the user test. Despite of the independent completion of the scenario is the goal, we need to guarantee the execution of the test, however, when this errors occur, the facilitator must act.

Critical Errors can also be assigned when the participant initiates, or attempts to initiate, an action that will result in the goal state becoming unobtainable. In general, Critical Errors are unresolved errors preventing completion of the task or errors that produce an incorrect outcome.

8.11 Non-Critical Errors

Non-Critical Errors, are errors that are recovered from and by the participant. Or, if not detected, do not result in processing problems or unexpected results. Although Non-Critical Errors can be undetected by the participant, when they are detected they are generally frustrating to the participant.

These errors may be procedural, in which the participant does not complete a scenario in the most optimal means (*e.g.*, excessive steps and keystrokes). These errors may also be errors of confusion (*e.g.*, initially selecting the wrong function, using a UI control incorrectly such as attempting to edit an un-editable field).

Non-Critical Errors can always be recovered from during the process of completing the scenario. Exploratory behavior, such as opening the wrong menu while searching for a function, will be coded as a non-critical error.

9 Goals

The next sections will describe the goals for *MM*, *Assistant* and *Heatmap* prototype expectations. We will try to assess performance-related metrics such as time and correctness of participants completing *tasks* for our expectations. Our expectations are based of the results obtained at the lab as pilot tests.

9.1 Completion Rate

Completion Rate is the percentage of test participants who successfully complete the task without critical errors. A critical error is defined as an error that results in an incorrect or incomplete outcome. In other words, the completion rate represents the percentage of participants who, when they are finished with the specified task, have an "output" that is correct.

*A **Completion Rate** of **90%** is the goal for each task in this usability test.*

Note: If a participant requires assistance in order to achieve a correct output then the task will be scored as a critical error and the overall completion rate for the task will be affected.

9.2 Error-Free Rate

Error-Free Rate is the percentage of test participants who complete the task without any errors (critical or non-critical errors). A non-critical error is an error that would not have an impact on the final output of the task but would result in the task being completed less efficiently.

*An **Error-Free Rate** of **80%** is the goal for each task in this tests.*

9.3 Time on Task (ToT)

The time to complete a scenario is referred to as "Time on Task" (ToT). It is measured from the time the participant begins the scenario to the time which the participant signals completion.

9.4 Subjective Measures

Subjective opinions about specific tasks, time to perform each task, features, and functionality will be surveyed. At the end of the test, participants will rate their satisfaction with the overall system. Combined with the interview/debriefing session, these data are used to assess attitudes of the participants.

Measuring subjective outcomes based on participants' experiential goals can pose challenges (Section 10) from which an *open-ended* flexible approach is catered to personally meaningful goals. On the other hand advocates of formalized Clinician-Centred Design (CCD) goal exploration condemn such informal interviewing as ineffective and we should take it into consideration.

9.5 Case Studies

The functionality of the prototype will be best demonstrated by a series of case studies. By describing the expected workflow and capabilities of the research study at the **RR** specific environment and changes of the workflow by using our system prototype. The study implies the evaluation of medical imaging *AI-Assisted* features on several breast lesions. The primary goal of this case studies analysis is to generate a receiver operating characteristic to evaluate the performance and validation of our *Assistant*. Let us consider a list of hypothetical use cases for the research investigation that evaluates the interaction and usability performance of the *Assistant*. Therefore, the following list will show the preliminary case studies.

List of case studies to analyse our solution prototype:

- *MM AI-Assisted* of a Breast Cancer Diagnosis;
- Priority & Minimal Information Visualisation;
- Performance & Response Measurement Values Acquisition;
- Radiologist Validation;

We expect to demonstrate several uses through a series of case studies, including implementation of our research prototype using an *AI-Assisted* technique and features for several view studies and other imaging research, as well as creation of a novel *Assistant* for the purpose. By creating a set of *Research Questions* (Section 6), we will try to achieve and feed this case studies. It is automatically associated with all cases. The radiologist may interact with our *Assistant* while manipulating the medical images and report to us difficulties and improvements. The number of questions is not restricted to the present document, since the interview will be *open-ended* and suggestive [38]. There is no limit to the number of questions that can be asked per case but it should fit the amount of expected time per each.

The data will be collected from the study video and observations into a spreadsheet for further analysis [20]. We will report the results of this tests and conclusions. This guide and respective use cases will be iteratively improved.

10 Challenges

In addition to the challenges already highlighted in the presented document, we must accomplish the participation test issues. The difference in knowledge and expertise levels between the participants will inhibit communication and participation of participants in different ways. Moreover, the factor that posed challenges to participants are involving them to a nominal adoption of consequences in the perceptions and practice, related ethical and self conflicts in presence of results. Challenges to the test and for practitioners to improve both study and research.

11 Results

A Test Report will be provided at the end of this tests. It will consist of a report and/or a presentation of the results; evaluation of the metrics against the preapproved goals, subjective evaluations, and specific issues of the system, as well as, recommendations for resolution. The recommendations will be categorically sized by development to aid in implementation strategy. The results will be translated to a spreadsheet (view only). Also, more related information can be found at Test 4: Assistant.

12 Acknowledgements

A special thanks for the support provided by Hugo Lencastre. We would like to thank Doctor Clara Aleluia, Doctor Gisela Andrade, Doctor Willian Schmitt, Doctor Ana Sofia Germano and Doctor Pedro Marques from the HFF for the generous support and medical expertise. Also, an immense thank for Doctor Cristina Ribeiro da Fonseca. My appreciation goes also to Bruno Cardoso and Bruno Dias for help and above all for the good companionship. Thanks to Professor Daniel Gonçalves, Professor Daniel Simões Lopes and Daniel Mendes for the technical inputs and network. Last but not least, thank to my advisors Professor Jacinto C. Nascimento and Professor Nuno Jardim Nunes. We also want to provide a special acknowledgment to Professor Ramtin Zargari Marandi who, among others, gave us important information and comments regarding the presented report. This work was partially supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013 and Instituto Superior Técnico (IST-ID) through the FCT/UID/EEA/50009/2013 project, BL89/2017-IST-ID grant. We would like to convey Hospital Fernando Fonseca (HFF) for the collaboration.

Acronyms

AI Artificial Intelligence.

Assis. Assistant.

BIRADS Breast Imaging Reporting and Data System.

CC CranioCaudal.

DICOM Digital Imaging and Communications in Medicine.

DOTS Dimensions Of Trust Scale.

MG MammoGraphy.

MI Medical Imaging.

MIDA Medical Imaging Diagnosis Assistant.

MIMBCD-UI MI Multimodality Breast Cancer Diagnosis UI.

MLO MedioLateral Oblique.

MM Multi-Modality.

MRI Magnetic Resonance Imaging.

NASA-TLX NASA Task Load Index.

SS Single-Modality.

SUS System Usability Scale.

UI User Interface.

US UltraSound.

Usa. Usability.

Work. Workload.

References

[1] Julia Abelson, Kathy Li, Geoff Wilson, Kristin Shields, Colleen Schneider, and Sarah Boesveld. Supporting quality public and patient engagement in health system organizations: development and usability testing of the p ublic and p atient e ngagement e valuation t ool. *Health Expectations*, 19(4):817–827, 2016.

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. 2019.

[3] Nicola C Anderson, Walter F Bischof, Kaitlin EW Laidlaw, Evan F Risko, and Alan Kingstone. Recurrence quantification analysis of eye movements. *Behavior research methods*, 45(3):842–856, 2013.

[4] Nabeela Ashraf, Waqar Ahmad, and Rehan Ashraf. A comparative study of data mining algorithms for high detection rate in intrusion detection system. *Annals of Emerging Technologies in Computing (AETiC)*, 2(1), 2018.

[5] Corinne Balleyguier, Salma Ayadi, Kim Van Nguyen, Daniel Vanel, Clarisse Dromain, and Robert Sigal. Birads classification in mammography. *European journal of radiology*, 61(2):192–194, 2007.

[6] Sifra Bolle, Geke Romijn, Ellen MA Smets, Eugene F Loos, Marleen Kunneman, and Julia CM van Weert. Authors' reply: Response to "older cancer patients' user experiences with web-based health information tools: A think-aloud study". *Journal of medical Internet research*, 18(11):e289, 2016.

[7] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pages 258–262, New York, NY, USA, 2019. ACM.

[8] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 4:1–4:14, New York, NY, USA, 2019. ACM.

[9] Francisco M. Calisto, Alfredo Ferreira, Jacinto C. Nascimento, and Daniel Gonçalves. Towards touch-based medical image diagnosis annotation. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces*, ISS '17, pages 390–395, New York, NY, USA, 2017. ACM.

[10] Francisco M Calisto, Pedro Miraldo, Nuno Nunes, and Jacinto C Nascimento. Breastscreening: A multimodality diagnostic assistant.

[11] Francisco Maria Calisto. Medical imaging multimodality breast cancer diagnosis user interface. Master's thesis, Instituto Superior Técnico, Avenida Rovisco Pais 1, 1049-001 Lisboa - Portugal (EU), 10 2017. A Medical Imaging for Multimodality of Breast Cancer Diagnosis by an User Interface.

[12] Francisco Maria Calisto. Mimbcd-ui/nasa-tlx: v1.0.0-alpha, September 2018.

[13] Francisco Maria Calisto. Mimbcd-ui/sus: v1.0.0-alpha, September 2018.

[14] Francisco Maria Calisto. mida-project/dots: v1.0.1-alpha, May 2019.

[15] Francisco Maria Calisto. mida-project/testing-guide-breast: v1.0.1-alpha, April 2019.

[16] Francisco Maria Calisto and Jacinto C. Nascimento. Nasa-tlx survey, 2018.

[17] Francisco Maria Calisto and Jacinto C. Nascimento. Sus survey, 2018.

[18] Francisco Maria Calisto, Nuno Jardim Nunes, and Jacinto C. Nascimento. Dimensions of trust scale (dots) survey. 2019.

[19] Hông-An Cao, Tri Kurniawan Wijaya, Karl Aberer, and Nuno Nunes. A collaborative framework for annotating energy datasets. In *Big Data (Big Data), 2015 IEEE International Conference on*, pages 2716–2725. IEEE, 2015.

[20] Pascale Carayon, Sarah Kianfar, Yaqiong Li, Anping Xie, Bashar Alyousef, and Abigail Wooldridge. A systematic review of mixed methods research on human factors and ergonomics in health care. *Applied ergonomics*, 51:291–321, 2015.

[21] Gustavo Carneiro, Jacinto Nascimento, and Andrew P Bradley. Unregistered multiview mammogram analysis with pre-trained deep learning models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 652–660. Springer, 2015.

[22] Pierre Chatelain, Harshita Sharma, Lior Drukker, Aris T Papageorghiou, and J Alison Noble. Evaluation of gaze tracking calibration for longitudinal biomedical imaging studies. *IEEE transactions on cybernetics*, (99):1–11, 2018.

[23] Robert E Cooke Jr, Michael G Gaeta, Dean M Kaufman, and John G Henrici. Picture archiving and communication system, June 3 2003. US Patent 6,574,629.

[24] Cheryl Delgado and Linda Wolf. Time on task: Perceived and measured time in online courses for students and faculty. *Journal of Nursing Education and Practice*, 7(5), 2017.

[25] Leandro L Di Stasi, Michael B McCamy, Stephen L Macknik, James A Mankin, Nicole Hooft, Andrés Catena, and Susana Martinez-Conde. Saccadic eye movement metrics reflect surgical residents' fatigue. *Annals of surgery*, 259(4):824–829, 2014.

[26] Ruth Ann Ehrlich and Dawn M Coakes. *Patient Care in Radiography-E-Book: With an Introduction to Medical Imaging*. Elsevier Health Sciences, 2016.

[27] Eda Elverici, Ayşe Nurdan Barça, Hafize Aktaş, Arzu Özsoy, Betül Zengin, Mehtap Çavuşoğlu, and Levent Araz. Nonpalpable bi-rads 4 breast lesions: sonographic findings and pathology correlation. *Diagnostic and Interventional Radiology*, 21(3):189, 2015.

[28] Joseph Feller, Brian Fitzgerald, et al. *Understanding open source software development*. Addison-Wesley London, 2002.

[29] David Flanagan. *JavaScript: the definitive guide*. " O'Reilly Media, Inc.", 2006.

[30] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[31] David Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2017.

[32] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.

[33] Jason Hostetter, Nishanth Khanna, and Jacob C Mandell. Integration of a zero-footprint cloud-based picture archiving and communication system with customizable forms for radiology research and education. *Academic radiology*, 2018.

[34] Jue Huang, Christine Ulke, Christian Sander, Philippe Jawinski, Janek Spada, Ulrich Hegerl, and Tilman Hensch. Impact of brain arousal and time-on-task on autonomic nervous system activity in the wake-sleep transition. *BMC neuroscience*, 19(1):18, 2018.

[35] John PA Ioannidis. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *Jama*, 279(4):281–286, 1998.

[36] S. Jodogne, C. Bernard, M. Devillers, E. Lenaerts, and P. Coucke. Orthanc – A lightweight, RESTful DICOM server for healthcare and medical research. In *Biomedical Imaging (ISBI), IEEE 10th International Symposium on*, pages 190–193, San Francisco, CA, USA, April 2013.

[37] Ankur Joshi, Saket Kale, Satish Chandel, and DK Pal. Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, 7(4):396, 2015.

[38] Ger Joyce, Mariana Lilley, Trevor Barker, and Amanda Jefferies. From healthcare to human-computer interaction: Using framework analysis within qualitative inquiry. In *International Conference on Applied Human Factors and Ergonomics*, pages 93–100. Springer, 2017.

[39] Ellen Kilsdonk, LW Peute, Rinke J Riezebos, Leontien C Kremer, and Monique WM Jaspers. Uncovering healthcare practitioners' information processing using the think-aloud method: From paper-based guideline to clinical decision support system. *International journal of medical informatics*, 86:10–19, 2016.

[40] Y Kobashi, Y Munetomo, A Baba, S Yamazoe, and T Mogami. Evaluation of the ossification of the cervical poste-rior longitudinal ligament utilizing x-ray, ct and mr imaging. *Orthop Res Traumatol Open J*, 2(1):35–39, 2017.

[41] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. 2019.

[42] Tong Li, Michael Luke Marinovich, and Nehmat Houssami. Digital breast tomosynthesis (3d mammography) for breast cancer screening and for assessment of screen-recalled findings: review of the evidence. *Expert review of anticancer therapy*, 18(8):785–791, 2018.

[43] Daniel Simões Lopes, Daniel Medeiros, Soraia Figueiredo Paulo, Pedro Brasil Borges, Vitor Nunes, Vasco Mascarenhas, Marcos Veiga, and Joaquim Armando Jorge. Interaction techniques for immersive ct colonography: A professional assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 629–637. Springer, 2018.

[44] Gabriel Maicas, Andrew P Bradley, Jacinto C Nascimento, Ian Reid, and Gustavo Carneiro. Pre and post-hoc diagnosis and interpretation of malignancy from breast dce-mri. *arXiv preprint arXiv:1809.09404*, 2018.

[45] Gabriel Maicas, Gustavo Carneiro, Andrew P Bradley, Jacinto C Nascimento, and Ian Reid. Deep reinforcement learning for active breast lesion detection from dce-mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 665–673. Springer, 2017.

[46] Ramtin Zargari Marandi, Pascal Madeleine, Øyvind Omland, Nicolas Vuillerme, and Afshin Samani. Eye movement characteristics reflected fatigue development in both young and elderly individuals. *Scientific reports*, 8(1):13148, 2018.

[47] Ramtin Zargari Marandi, Pascal Madeleine, Øyvind Omland, Nicolas Vuillerme, and Afshin Samani. Reliability of oculometrics during a mentally demanding task in young and old adults. *Ieee Access*, 6:17500–17517, 2018.

[48] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.

[49] Rajan Merchant, Rubina Inamdar, Kelly Henderson, Meredith Barrett, Jason G Su, Jesika Riley, David Van Sickle, and David Stempel. Digital health intervention for asthma: patient-reported value and usability. *JMIR mHealth and uHealth*, 6(6):e133, 2018.

[50] Diana L Miglioretti, Rebecca Smith-Bindman, Linn Abraham, R James Brenner, Patricia A Carney, Erin J Aiello Bowles, Diana SM Buist, and Joann G Elmore. Radiologist characteristics associated with interpretive performance of diagnostic mammography. *Journal of the National Cancer Institute*, 99(24):1854–1863, 2007.

[51] S Obenauer, KP Hermann, and E Grabbe. Applications and literature review of the bi-rads classification. *European radiology*, 15(5):1027–1036, 2005.

[52] American College of Radiology. BI-RADS Committee. *Breast imaging reporting and data system*. American College of Radiology, 1998.

[53] Bruno Oliveira, Francisco Maria Calisto, Lilian Gomes, and José Borbinha. Adaptive q-sort matrix generation: A simplified approach.

[54] Konstantina Orfanou, Nikolaos Tselios, and Christos Katsanos. Perceived usability evaluation of learning management systems: Empirical evaluation of the system usability scale. *The International Review of Research in Open and Distributed Learning*, 16(2), 2015.

[55] Soraia Figueiredo Paulo, Nuno Figueiredo, Joaquim Armando Jorge, and Daniel Simões Lopes. 3d reconstruction of ct colonography models for vr/ar applications using free software tools.

[56] Anjana Ramkumar, Pieter Jan Stappers, Wiro J Niessen, Sonja Adebahr, Tanja Schimek-Jasch, Ursula Nestle, and Yu Song. Using goms and nasa-tlx to evaluate human–computer interaction process in interactive segmentation. *International Journal of Human–Computer Interaction*, 33(2):123–134, 2017.

[57] Carrie Reale, Ross Speir, Kurt Ruark, Jennifer Herout, Jason Slagle, Matthew B Weinger, and Shilo Anders. Using scenarios throughout the user-centered design process in healthcare. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 62, pages 610–614. SAGE Publications Sage CA: Los Angeles, CA, 2018.

[58] David Ribeiro, André Mateus, Pedro Miraldo, and Jacinto C Nascimento. A real-time deep learning pedestrian detector for robot navigation. In *Autonomous Robot Systems and Competitions (ICARSC), 2017 IEEE International Conference on*, pages 165–171. IEEE, 2017.

[59] David Ribeiro, Andre Mateus, Jacinto C Nascimento, and Pedro Miraldo. A real-time pedestrian detector using deep learning for human-aware navigation. *arXiv preprint arXiv:1607.04441*, 2016.

[60] Melissa R Rowell, Frank M Corl, Pamela T Johnson, and Elliot K Fishman. Internet-based dissemination of educational audiocasts: a primer in podcasting-how to do it. *American Journal of Roentgenology*, 186(6):1792–1796, 2006.

[61] Carlos Santiago, Jacinto C Nascimento, and Jorge S Marques. Combining an active shape and motion models for object segmentation in image sequences. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3703–3707. IEEE, 2018.

[62] J Sauro. 10 things to know about the single ease question (seq). *Measuring U, 2012*, 2012.

[63] Maurício Sousa, Daniel Mendes, Soraia Paulo, Nuno Matela, Joaquim Jorge, and Daniel Simões Lopes. Vrrrroom: Virtual reality for radiologists in the reading room. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 4057–4062. ACM, 2017.

[64] Preethi Vaidyanathan, Jeff Pelz, Cecilia Alm, Pengcheng Shi, and Anne Haake. Recurrence quantification analysis reveals eye-movement behavior differences between experts and novices. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 303–306. ACM, 2014.

[65] Andrea Vedaldi and Karel Lenc. Matconvnet: Convolutional neural networks for matlab. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 689–692. ACM, 2015.

[66] Stephen Waite, Srinivas Kolla, Jean Jeudy, Alan Legasto, Stephen L Macknik, Susana Martinez-Conde, Elizabeth A Krupinski, and Deborah L Reede. Tired in the reading room: the influence of fatigue in radiology. *Journal of the American College of Radiology*, 14(2):191–197, 2017.

[67] Jim Wilson. *Node. js 8 the Right Way: Practical, Server-side Javascript that Scales*. Pragmatic Bookshelf, 2018.