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 aims to describe the protocol and guidelines of the presented information. We perform a set of tests in the scope of both v1.0.1-alpha, v1.2.0-alpha and v1.2.0-alpha versions from both prototype-multi-modality, prototype-multi-modality-assistant and prototype-heatmap repositories, 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 a record heat-map, 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 clinician'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, $n \in \mathbb{N} \; \forall \; N = \{1, ..., 10\}$) per each three activities representing two different scenarios: *Multi-Modality* (**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 50 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 $r_1, r_2, r_3 \in \mathbb{N} \; \forall \; R = \{1, ..., 50\}$. While $Pat1. = Rdm_{r_1} \wedge Pat2. = Rdm_{r_2} \wedge Pat3. = Rdm_{r_3}$ as soon as $r_1 \neq r_2 \neq r_3$ is *True*. In both two scenarios, *i.e.*, *Multi-Modality* (**Sce1.**) and *Assistant* (**Sce2.**), by supporting our traditional devices, the interaction is made with mouse and keyboard. Clinicians will classify each patient by using the BIRADS [2]. We will do several small questionnaires at the end of each scenario using NASA-TLX [30], System Usability Scale (SUS) [29] and measuring the BIRADS [2].

Describing each phase, the first phase is nominated as **Pha1. User Characterization**. It is the Demographic Questionnaire, supporting our several characterizations of the clinician profile. For the major number of clinicians, 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 the survey regarding their characterization as a user.

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

Finally, and most importantly, the third phase, nominated as **Pha3. Assistant Establishment** phase, is where clinicians 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 [1, 23] on user expectations, as intended, specifically as outlined by our several *Hypotheses* (Section 6) 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* [5]. On this phase, we will measure the participants expectations concerning the *AI-Assisted* system that we called **Act7. Second Scenario User Expectations** activity. The next **Act8. Second Scenario Introduction** activity, will be the introductory information about *AI-Assisted* system and what feature are covered by the system. The **Act9. Second Scenario Evaluation** activity represents the diagnostic of each patient (same as the **Act5.** activity), but this time with support of our novel *AI-Assisted* recommendations and explainability. And finally, the **Act10. Post-task Second Scenario Questions** activity will be a set of *post-task* questions.

For the user tests we used a three distinct prototype repositories, *i.e.*, the prototype-multi-modality repository, the prototype-multi-modality-assistant repository and the prototype-heatmap repository. The three are similar mirrors of the prototype-breast-cancer with major changes. The first repository, *i.e.*, the prototype-multi-modality repository, aims at providing clinicians a *Multi-Modality* strategy view. The *Multi-Modality* 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). This part corresponds to the **Sce1.** scenario of both **Act4.** and **Act5.** activities at the **Pha2.** phase. The second and third repositories, *i.e.*, the prototype-multi-modality-assistant repository and the prototype-heatmap repository, 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) [16, 17] 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 **Act7.**, **Act8.**, **Act9.** and **Act10.** 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 clinician performance, establishing and validating clinician 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) for 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;

The 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. The 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 clinicians, 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 prototype-multi-modality-assistant and prototype-heatmap repositories, respectively. The purpose of these prototypes is to involve an *AI-Assisted* tool (*Assistant*) for medical imaging at a breast screening diagnosis level. This *Assistant* was created with a front-end and back-end architecture utilising common programming languages, libraries, frameworks and tools including JavaScript (JS) [14], NodeJS [38], HammerJS, CornerstoneJS [18] and Orthanc [19]. For the Machine Learning (ML) and Deep Learning (DL) [32, 33] component we will use several MATLAB technologies [36], promoting and feeding our Convolutional Neural Networks (CNN) [8] and Deep Reenforcement Learning (DRL) [25] techniques. Other central component of this prototype is a web-based PACS [10] pairwise with ubicous web technologies and based on the **Open Source (OS)** CornerstoneJS library [13, 18].

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 [37] that radiologist fatigue levels and performance are related to environmental factors such as number of false-negative and false-positives, 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**. We expect to analyze the needed information and best solution to improve the workstation results.

Figure 1: Radiology Room

3.2 Participants

The participants' responsibilities will be to attempt to complete a set of representative task scenarios 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 [31] begins when the participant starts the task.

The facilitator will instruct the participant to "think aloud" [3, 21] 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.** - **Act7.**) questionnaire. During each task, participants will elaborate on the task session (**Pha2.** - **Act8.**) and complete the post-task (**Pha2.** - **Act9.**) questionnaires with the facilitator. After all task scenarios are attempted, the participant will complete the post-test (**Pha2.** - **Act10.**) questionnaire.

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* application 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 the 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 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 Material

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 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 heatmaps 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 [34]. The QuickTime (Apple Computer) tool is available for MacBook Pro to 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 [9], 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 [4, 5, 6] 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).

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 [28] 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* ?

Table 1: Research Evaluation Questions

The influence of *AI-Assisted* [15] 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 over 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 [1, 23] authors instructions. The first work [1], 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 [23], 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) [16, 17] 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 **RQ2.**, 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 [1] 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 [23] 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 [1]. Again, we also associated each of the three *Hypotheses* with the set of guidelines [1]. 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. For instance, we can ask about *How would the user describe the potential adoption of AI-Assisted methods on the Health Institution?* question. 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 [24] of the patient [12]. 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 an important 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 will 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 by 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 [35] 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.

We will try to understand if, with the *AI-Assisted* techniques, the clinicians will encounters the most accurate severity (BIRADS) of the breast lesions [27] and patient's prognostic. For this purpose, we have three patients (Figure 2); each patient has three images in the respective modalities: (i) MG; (ii) US; and (iii) MRI. The clinicians will proceed to the activity of diagnosing the three patients within the support of our *Assistant* by the observation of ALL images.

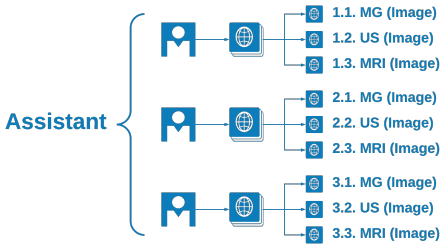

Figure 2: Diagram representing the use of the *Assistant* by clinicians.

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.

List of stand alone tasks:

Task 1.1: Classify *Patient 1* on the *Assistant*;

Task 1.2: Classify *Patient 2* on the *Assistant*;

Task 1.3: Classify *Patient 3* on the *Assistant*;

Task 2.1: Freely explore *Patient 1* on the *Assistant*;

Task 2.2: Freely explore *Patient 2* on the *Assistant*;

Task 2.3: Freely explore *Patient 3* on the *Assistant*;

8 Metrics

9 Goals

The next sections will describe the goals for prototype-multi-modality-assistant 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.

10 Challenges

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 pre-approved 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

12.1 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:

- Multi-Modality *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 questions,

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 imaging and report to us difficulties and improvements. The number of questions is not restricted to the present document, since the interview will be open and suggestive [20]. 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 collect from the study video and observations into a spreadsheet for further analysis. We will report the results of this tests and conclusions. This guide and respective use cases will be iteratively improved.

12.2 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).

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. We will attach the eye tracker under the display area with a magnetic mounting bracket, following the product's instructions. A 9-point calibration [9] will be performed for each user. Our protocol was implemented in Python (*versions* >= v2.7) and Processing (*versions* >= 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.

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*.

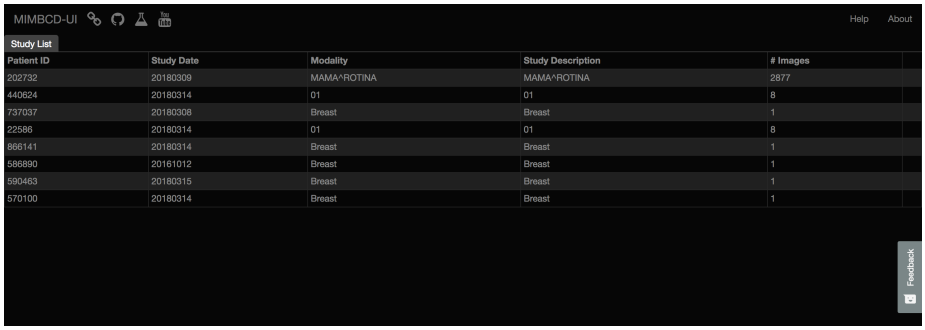

Figure 3: List of Patients.

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.

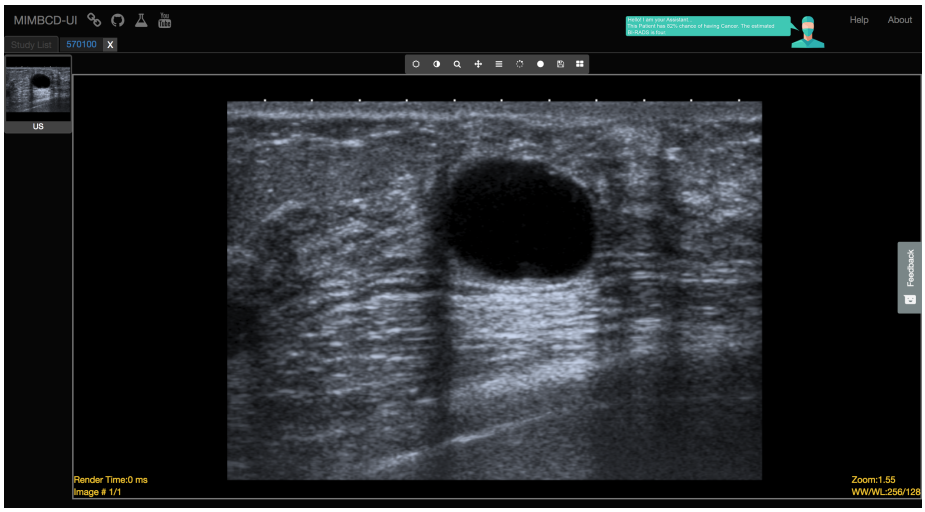

Figure 4: Viewer of the **DICOM** images.

Manual annotation 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 informatiuon 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.

12.3 User Interactions

The systems have several buttons (Figure 5) 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 5.

Figure 5: Toolbar of the System available features.

The buttons are (from left to right of Figure 5) as follows:

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

12.4 Post-Task Questionnaire

Our metrics will refer the *AI-Assisted* involvement against specific goals necessary to satisfy several requirements of our *Assistant*. For our **Post-Task Questionnaire** we will use *Quantitative Analysis* (QtA) and *Qualitative Analysis* (QlA) in response to our questions (Section 3) to measure the acceptability of our *Assistant* each time a *scenario* is completed (Section 3). From a set of tasks (Section 6) we aim to cover our main scenario, an *AI-Assisted* that we call *Assistant*. Therefore, both QtA and QlA will allow the facilitator to quickly and easily assess the requirements of the given scenario. Our QtA and QlA requirements will have several attributes [20] that make it a good choice for our clinical participants. Those attributes are as follows.

List of the scale attributes:

- The requirements are technology agnostic, making it flexible enough;
- The requirements are relatively quick and easy to answer;
- The requirements provide a single score on a scale that is easily understood;
- The requirements are nonproprietary, making it a cost effective tool;

The facilitator will explain that the amount of time taken to complete the *tasks* will be measured and that exploratory behaviour outside the *task* flow should not occur until after task completion. At the beginning of each task, the participant will listen the *task* description from the facilitator and begin the task. *Time-on-Task* (ToT) measurements begins when the participant starts the *task*, measured until the end of each *task*.

12.5 Training Session

The participant will receive and overview the test procedure. However, the user will not receive information how to annotate and interact in all degrees of freedom. With the aim of disabling users to get their work done before the test tasks. It will take advantage of a "surprise" acknowledgement.

12.6 Execution of Tasks

The *tasks* were derived from test scenarios developed from **Case Studies**. Due to the range and extent of functionality provided by our *Assistant*, and the short time from which each participant will be available, the *tasks* are the most common and relatively complex of available functions. The *tasks* are the identical for all participants of a given user role in the study.

The *tasks* will be performed by several classes of radiology experience. Professionals from Radiology Seniors, Middles, Juniors and Interns will be performing these *tasks*. On the **RR** the Radiologist is characterised [11, 26] as a physician who examines and interpret Medical Imaging (MI) [22], such as X-Rays, CT Scans or MRIs.

12.7 Post-Activity Questionnaire

After completing all *tasks* and scenarios, participants will be asked to complete a questionnaire to classify the *Assistant* according to various parameters regarding the several features. To measure this, we will use an open session *observation* and *interview* [7]. We will use this techniques to identify participants' requirements during the various stages of the workflow.

13 Measurements

Our measurements refers to user performance measured against specific performance goals necessary to satisfy requirements. *Task* completion success rates, adherence to dialog scripts, error rates and subjective evaluations will be used. *Time-to-Completion* (TtC) of *tasks* will also be collected. The measures are as follows.

The tests are intended to achieve the following measures:

- BIRADS Classification;
- Pathology Classification;
- Time measurement;
- Number of clicks;
- Number of errors;
- Efficiency;
- Difficulty;
- Experience;

To prioritise recommendations, a method for problem difficulty and degree severity classification, as well as, pathology importance, will be used in the analysis of the collected data during evaluation process. The approach treats problem severity has a combination of several factors. Those factors are measuring the impact of the problem and the frequency of users experiencing issues during the evaluation. Nevertheless, the opinion will also be of chief importance and we will also register the received ones.

Through the questionnaire after the test session, we intend to obtain the answers to the following questions for each *task*:

- Acceptability of the *Assistant* lesion classification;
- Acceptability of the *Assistant* interaction;
- Acceptability of the *Assistant* translation;
- Acceptability of the *Assistant* available features;
- *Assistant* degrees of classification;
- *Assistant* degrees of interaction;
- *Assistant* degrees of information visualisation;

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

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

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

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

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

[37] 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.

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