# Breast Cancer Multimodality Scalable Interactions

Hugo Lencastre
hugo.lencastre@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

December 2020

**Abstract**

Computer-Aided Diagnosis (**CADx**) systems are essential when diagnosing patients with cancer. Medical Imaging Multimodality Breast Cancer Diagnosis User Interface (**MIMBCD-UI**) is a Computer-aided Detection (**CADe**) system that allows to open, view and manipulate medical images in order to diagnose patients with breast cancer. In this work, we aim to improve this system, thus allowing a faster medical image manipulation, by creating automated processes. With Human-Computer Interaction (**HCI**) techniques, such as Focus Groups, Affinity Diagrams, Interviews, Questionnaires and Scales, we developed functionalities based on the specialists' opinions. The three functionalities created were focused on reducing steps in the medical image manipulation, without reducing its quality and while making the analysis effortless and faster. It was proven that these functionalities enabled us to improve the usability, by increasing its value from 86.935 to 91.(1); the workload, by decreasing its value from 29.1(4) to 15.037; and the time of a diagnosis process by reducing the number of clicks by half, when compared with the previous iteration. All the purposed goals in our Design Goals and Research Questions were achieved and proven with the results obtained from the tests. With a full base system, the upcoming developments will start by refine our functionalities or the creation of the functionalities that are desired. The ultimate goal is to have this system merging with iterations that are being developed at this instant, Artificial Intelligence (**AI**) and eXplainable Artificial Intelligence (**XAI**), which will allow the system to become a complete **CADx** that could be applied in real scenarios and help to save lives.

**Keywords:** Breast Cancer, Human-Computer Interaction, Computer-Aided Diagnosis, User Interface, Design Thinking.

## 1. Introduction

Breast cancer is one of the most common cancers, especially in women affecting around 21%, and is the fifth with a higher mortality overall. However, if the lesion is detected in early stages, depending on the country health system where the patient is treated, the survival rate can increase [3, 22]. In Portugal, the disease surveillance is done regularly, every two years, after the age of 50 years old [19], using Mammography (**MG**) and/or Ultrasound (**US**), and the diagnosis is done following the Breast Imaging-Reporting and Data System (**BI-RADS**) [4] classification.

Medical Imaging Multimodality Breast Cancer Diagnosis User Interface (**MIMBCD-UI**) [8] project aims to develop a system to help physicians with the breast cancer diagnosis. Our work is the ninth iteration of this project and has the objective of improving the base system, by developing new functionalities not yet available in many hospitals systems and refining those existent. With this iteration and with the junction of other iterations, the Artificial Intelligence (**AI**) [9] and the eXplainable AI (**XAI**) [21], we will be able to have a Computer-Aided Diagnosis (**CADx**), a program that enables a better lesion understanding, and where the **AI** is able to do a pre-diagnosis or give a second opinion after the physicians' analysis. With the explainability, we intent to give the user an understanding of the **AI**'s response on the second opinion.

This work will focus on the development of the capabilities, by using a Human-Computer Interaction (**HCI**) approach, where several **HCI** techniques will be used to identify and resolve problems in the several stages of the design process that we chose.

## 2. Background

The breast cancer area domain is very complex, with several clinical terms that could affect the outcome of the lesion classification, and that need to be explained along with other definitions that will be used in this document.

### 2.1. Lesions

There are two types of lesions that can occur in a breast, **Masses** [2] and **Calcifications** [1]. These are very different both in the process of being find and classified, although it is used the same **BI-RADS** [4] classification for the diagnosis.

A **Mass** is a 3D lesion that can be seen from two different projections, and described by three categories: Shape; Margins; Density [2, 4]. Depending on its characteristics, it can lead to a malignant or benign mass.

**Calcifications** are small calcium deposits that calcify on the Terminal Ductal Lobular Unit (**TDLU**) [1, 4], Figure 1, in the *Terminal Duct* or in the *Acini* [1, 4]. The position and the amount of calcifications can lead to a malignant lesion, Figure 2.

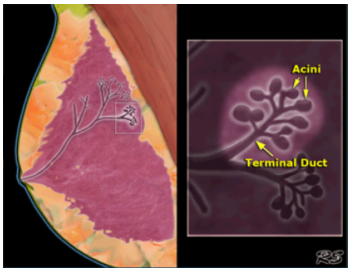


Figure 1: The basic functional unit in the breast, also called the TDLU, where the "leafs" are the *Acini* and the "branch" is the *Terminal Duct* [1].

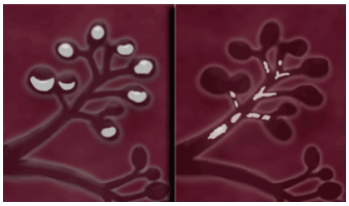

Figure 2: Calcifications, are small calcium deposits founded in the *Terminal Duct* and the *Acini*.[1]

### 2.2. Definitions

All images in the medical imaging field are called Modalities, which represent all types of medical images that can be taken when using medical imaging devices, such as **MG**. This work can accept all these types of images and even allow all to be manipulated through it.

As it was previously mentioned, this work was done using **HCI** techniques. **HCI** is a multidisciplinary field of study, focused on the creation and design of computer technology, with a special attention to the users' needs and their interactions with a computer [5]. This field of study has developed techniques allowing the identification and construction of a full test, and also explaining or grading, if the system meets the users' needs [5].

## 3. Related Work

With this work, we explored systems that have a similar domain of application or a very strong **HCI** approach, that could help us guide our path or take ideas that could be important to explore.

Hatscher, B. *et al.* developed a prototype that translates touchless hand gestures into functions of a special-purpose software for **MRI**-guided interventions [14]. From this work we took the idea of having a similar path of research but with more defined and separated steps in the **Design Process**. The scales used in this work were also inspired by them, both the System Usability Scale (**SUS**)[6, 11] and NASA Task Load Index (**NASA-TLX**) [10, 13, 23] scales.

Li, L. *et al.* developed an interactive online patient decision aid, called *ANSWER-2*, that reduces patient decision conflict and improves their medication-related knowledge and self-management capacity [18]. Regarding this work, we did not use their scales, but several **Design Methods**, such as **Interviews**, to ensure that our work meets the expectations of its users.

Stuijfzand, B. *et al.* aimed to measure the cognitive load of medical students, when interpreting volumetric images such as Computed Tomography (**CT**) or Magnetic Resonance Imaging (**MRI**), by applying **HCI** techniques and an eye tracking system [24]. With this work, we understood that using **Metrics** to analyze our data would be important, such as **Number of Errors** and **Time**. Unfortunately, it was not possible to use the eye tracker technology to understand what is being viewed by the physician given the restrictions that happened due to the pandemic (COVID-19).

## 4. Objectives

We divided the objectives that we have for this work into **Design Goals** and **Research Questions**.

### 4.1. Design Goals

**Design Goals** are specific objectives in terms of design that we want our system to achieve. Overall, we chose three design goals: the systems' **Usability**, where we want to see if our system is easy to use; **Efficiency**, in the realization of the actions necessary to complete a task; and **Produtivity**, where less workload a user has, the more productive they can be.

### 4.2. Research Questions

On the other hand, the **Research Questions** are developed to understand some parts of the system, but can, in some ways, be related to the **Design Goals**. These are our research questions:

[RQ.1] When and who will use the features?

[H1.1] The functionalities were rejected;

[H1.2] The functionalities were used, only in cases of doubt;

[H1.3] The functionalities were used, but only by inexperienced physicians;

[H1.4] The functionalities were used by all physicians.

[RQ.2] What is the impact of the features, in the clinic workflow?

[H2.1] The usability of the system increased;

[H2.2] The workload impact was affected in a positive way;

[H2.3] Diagnostic time per patient was reduced.

[RQ.3] What is the best method to represent the lesion evolution?

[H3.1] Old annotations on top of the more recent image, using a time bar;

[H3.2] With a time bar in the left viewport, and a more recent image in the right viewport;

[H3.3] With a time bar in the right viewport, and a more recent image in the left viewport;

[H3.4] A time bar in each viewport;

[H3.5] The lesion evolution functionality was rejected.

## 5. Evaluation

In **HCI** it is common to have a process of identification, creation, testing and analysis of a problem, which is called a **Design Process**. We wanted that our design process could allow us to re-think and re-design the ideas created to resolve the problems found, without a large cost to the researchers. With these requirements, we chose the **Design Thinking** [7] process, that is divided into 5 design stages [20]: **Defining the Problem**; **Needfinding and Benchmarking**; **Bodystorm; Prototype; Test**.

Only one of our three features made it through the whole design process, given that the others were already considered future work in previous interactions of the project [8] or expressed as a desired between informal interviews.

The **Recorded View**, was identified by us, in **Defining the Problem** [7] stage, when analyzing the videos from previous works. Here, we noticed that in each image loaded, if changes were already made, the values would be erased, which would force the physician to repeat them in order to obtain the same state.

After discovering this problem, we had several meetings, in the **Needfinding and Benchmarking** [7] stage, between researchers to determine what would be the best method to implement in this functionalities.

In the third stage, **Bodystorm**, all features were treated as equal in the design process. Several **Focus Groups** were done during this period, where domain and features were discuss between physicians and researchers. It was used a technique called **Affinity Diagrams**, that consists in creating several notes with ideas to solve the problems found/discussed. With these techniques we had a first understanding on how to proceed with the functionalities. At the end of this stage, we gave to each physician that agreed to participate in the test, a questionnaire that asked about the system currently used in hospitals and functionalities desired to be developed. With the information gathered from these stages, we started to develop the three functionalities: **Recorded View**; **Coordinated View** and **Temporal View**. During the **Prototype** stage, other several small system improvements were done to clean some bugs or to improve the information given to the physician.

With the functionalities developed we started the **Test** stage, where ten tests were performed, one for each user, and where four studies were available to be tested. In each test, the user was asked to make several actions in the system, which corresponded to simple actions or to the functionalities. This was done using a technique called **Talk-aloud**, where the user says what he/she is doing/thinking at the moment. During the test, several metrics were recorded and, at the end, a questionnaire and two scales were given, SUS [11, 6] and NASA-TLX [10, 13, 23]. The final procedure at this stage, is collecting the information obtained and understand if our objective were reached.

## 6. Implementation

Three functionalities, chosen by the users as the most necessary in a medical imaging manipulation system, were developed:

**Recorded View** functionality aims to keep the system state while images are swapped, manipulated and analyzed. This is done by recording each image state when any manipulation is made, Figure 3.

**Coordinated View** functionality opens images that are normally viewed side by side in order to see breast asymmetry and, at the same time, allow the manipulation of one image to be reproduced on the others, Figure 4.

Finally, **Temporal View** allows to compare the

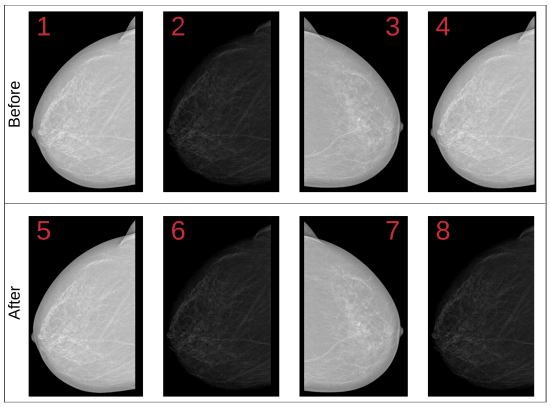


Figure 3: The **Recorded View** functionality effect. The images in the line Before represent how the system worked, whereas the ones in the line After, represent how the process is now conducted.

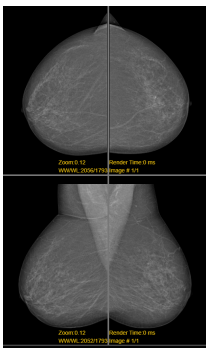

Figure 4: **Coordinated View** final result with four images that were open automatically at their locations after choosing the 2 x 2 viewport and with one contrast manipulation, all others made the same change.

images side by side, like the **Coordinated View**, regarding the same modality, projection and laterality from different time periods, and with a time bar for fast swap between images, Figure 5.

## 7. Results

This work is an evolution from the previous iteration [8], therefore, it is expected the comparison between our results and the ones from that iteration. However, since the tests were different, we were not able to compare exactly the two iterations, but rather the usability, workload and some metrics.

One of the first values that were recorded from the tests, were the metrics, more specifically the **Time**, **Count use of a tool** and **Errors**.

### 7.1. Metrics

Regarding the **Time** metric, this was discarded given that there were different conditions per user during the tests. While some users were direct, when performing the task and responding to the

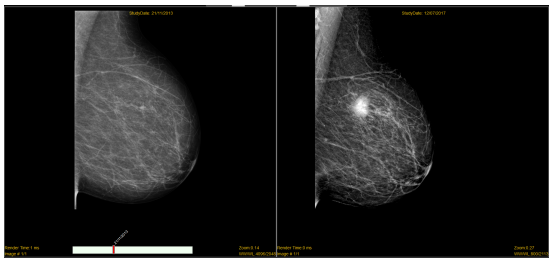

Figure 5: **Temporal View** Comparison effect, the high-fidelity prototype, after understanding the need that exists when analyzing the evolution of the breast. Here, we have the oldest image on the left, with no lesion, and on the right, the same breast architecture that enable us to spot quickly the new lesion

questionnaires, others stayed for longer periods of time giving feedback about the system or the operations that are necessary to be done when analyzing an exam.

In each test, the number of clicks necessary to produce a full action with and without our functionalities were counted, **Count use of a tool** metric. This metric allowed us to understand that each interaction takes up to two seconds to be executed and also to record the operation made in system by the user, generating a report of the test.

In all functionalities, there was an improvement from the previous iteration. The **Recorded View** enabled the storing of the system state, allowing the user to re-visit old images and analyze them, the way they were left, hence saving all the clicks necessary to reproduce the previous state. Regarding the **Coordinated view**, that is used to view asymmetry, it would be necessary 4 clicks and a search without our functionality whereas, with our them, it is only necessary 1 click or one search plus 1 click. For the **Temporal View**, that allows the comparison of two images from different time periods, without the functionalities, it would take 2 clicks, 2 drags and 2 searches and when changing to the other image, 1 more drag and 1 more search, whereas with our functionalities, it would take 1 click, 1 search and 1 drag and for changing the image, just 1 click in the time bar for another time period or drag a new projection. With these results it was possible to reduce several actions, allowing a final time reduction per patient.

The **Errors** metric can be characterized as one of two types, **Non-Critical** and **Critical**. In our tests we had 3 critical errors, from those, 1 made us change some aspects of the test (*e.g.*, changing the browser), the others, though critical, allowed us to continue the test with some of the features. The **Non-Critical** errors are more important for us, since they gave us the ability to understand if

the perfect path was made for an action or if the user encountered any bug in the functionalities. In this regard, we had an average of 1.9 non-critical errors in the entire system for each physician, a good result given that we had several functionalities tested during this process. Surprisingly, the most bug functionality was **Zoom**, a functionality out of our focus, that occurred with 6 physicians, but with a total 8 non-critical errors.

7.2. Scales

In our work, two scales were used, one to measure the usability of the system, the System Usability Scale (SUS) [6, 11] and the other to measure the workload of the system, the NASA Task Load Index (**NASA-TLX**)[10, 13, 23].

From the data collected in this test and the data from previous iterations, we started by clean both of any outlier present. This process required the use of the **Tukey Fences**[15] test that used the *Interquartile range* to identify that every data point outside the range will be considered an outlier, equations 1, 2, 3, 4, 5.

$$Q1 = First\ Quartile;\ Q3 = Third\ Quartile, \quad (1)$$

$$IQR = Q3 - Q1, \quad (2)$$

$$Below\ Outlier < Q1 - 1.5 \times IQR, \quad (3)$$

$$Above\ Outlier > Q3 + 1.5 \times IQR, \quad (4)$$

$$[\ Below\ Outlier\ \ ;\ \ Above\ Outlier\ ] \quad (5)$$

Where:

*First Quartile* ($Q1$) = the middle value between the smallest value and the median of the data set;

*Third Quartile* ($Q3$) = the middle value between the median and the highest value of the data set;

*Interquartile Range* ($IQR$) = measure of variability based on dividing a data set into quartiles;

*Below Outlier* = minimum value accepted in a data set;

*Above Outlier* = maximum value accepted in a data set;

#### 7.2.1 System Usability Scale

System Usability Scale (**SUS**) [6, 11] is characterized by having ten statements, five positives and five negatives, using a likert-scale of five points, from *strongly disagrees* to *strongly agrees*.

In our work, it was found an outlier with the score of 67.5 out of 100, and no outliers were found in the previous iteration [8]. By removing this data point, we are left with the following results:

| Our work, Iteration 9 | Iteration 4: |
|---|---|
| Median = 92.5; | Median = 87.5; |
| Mean = 91.(1); | Mean = 86.935; |
| $\sigma$= 7.648. | $\sigma$= 9.811. |

As it is possible to see in the Figure 6, our data, orange columns, is more condensed in the higher value ranges. The outlier is represented in this graph by an orange column with a red border. The previous iteration, columns in blue, are scattered through the ranges. With this SUS [6, 11] characterization, both results are considered *Excellent* given that both have **Mean** above the 80.3, however, our data had better results overall.

#### 7.2.2 NASA Task Load Index

NASA Task Load Index (**NASA-TLX**) [10, 13, 23] was design to measure the workload, with six questions, five regarding several difficulties that could be experienced and one question about the performance the participants think they had. All questions are responded by choosing a mark, which represent a five step interval from 0 to 100, in a total of twenty steps. On contrary to the SUS[6, 11], the lower the result the better in the end of the analysis.

Once again, the outlier test was performed and were found in both iterations. In our work the outlier had a score value of 46 out of 100, whilst in the previous iteration [8], the outlier had a score value of 84.(3) out of 100. By removing these outliers data points, the final results are the following:

| Our work, Iteration 9 | Iteration 4: |
|---|---|
| Median = 11; | Median = 21.5; |
| Mean = 15.037; | Mean = 29.1(4); |
| $\sigma$= 13.186. | $\sigma$= 19.035. |

Once again, this work results are an improvement over the last iteration [8], Figure 7, where our data is once again the orange columns and the blue columns are the data from the previous iteration [8]. In both cases, the outliers will have a red border around their columns. The data samples in our work are again more condensed around the *Low* $[0; 9]$ and *Medium* $[10; 29]$ categories. On the other hand, the previous iteration [8], has the data scattered in the different ranges, with the majority of their data in the *Medium* and *Somewhat High* categories.

#### 7.2.3 Advanced Statistics

Our data, from both SUS [6, 11] and NASA-TLX [10, 13, 23] scales are non-normal distributions, so we performed a test that is specific to this type of data. We chose the **Kruskal Wallis** [16] test, that has the objective of understanding if our groups have equal median values or not. This test has two hypothesis, $H0$ that refers if the population medians are equal and the $H1$ the opposite. If $H0$ is rejected, it is necessary to make a *Post-Hoc* test, where we chose the **Dunn's**[12] test that measures how similar or different they are.

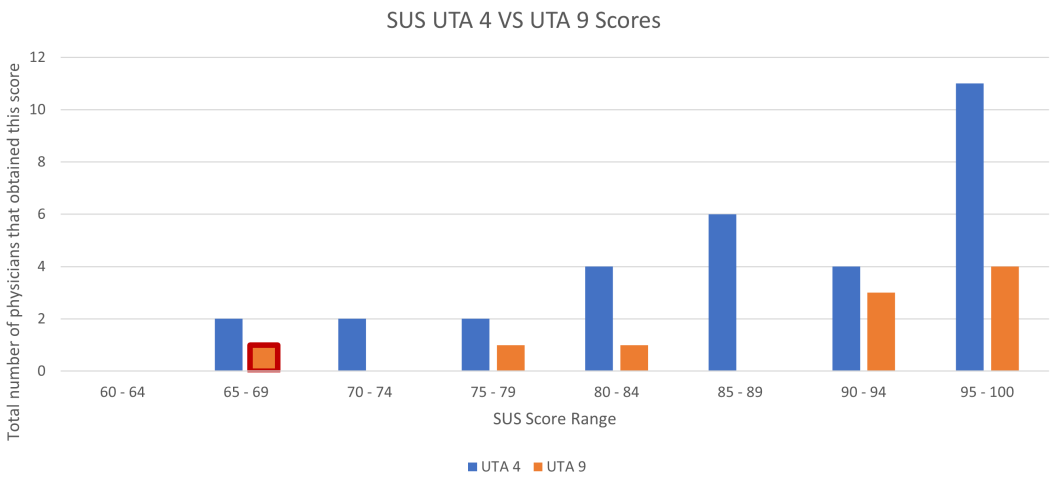


Figure 6: Comparison between fourth iteration[8] and this work. In this graph we can see both data points using the same scale, in the vertical axis, the total number of users that obtain a score in that score range. In the horizontal axis, the SUS[6, 11] score ranges with five score distance. The previous iteration[8] is represented by blue columns and this work score is represented by orange columns. Important to notice the columns with a red border are the data points that were considered an outlier. Both graphs are not-normal distributions and demonstrate the satisfactory result present in both iterations.

**Kruskal Wallis** [16] test had the following results for each scale and **UTA**:

$chi - square = 16.919$;

SUS UTA9 - $H$=1.407;

SUS UTA4 - $H$=3.213;

NASA-TLX UTA9 - $H$=3.652;

NASA-TLX UTA4 - $H$=0.659;

The **Kruskal Wallis** [16] test says that if the $H$ is higher than the $chi - square$, it means that the median value is equal. In our case this is the opposite, we can not prove that hypothesis, so the $H1$ is accepted and we need to do a *Post-Hoc* test to be able to understand how similar or different they are.

**Dunn's** [12] test, concludes that our groups and the groups from the previous iteration [8] are completely different in the majority of the scales, except in two occasions, in the **NASA-TLX** [10, 13, 23] in our work, where the **Intern** and the **Senior** group are more similar, with a value of 0.056; and in **SUS** [6, 11] from the previous iteration [8], where the same groups are similar with a value of 0.096. If the value was equal to zero, it would mean that the groups were similar.

With these tests, we can conclude that, although our groups are not similar in general, they can have very similar opinions in some situations and, therefore, it is important to keep the tests with all groups.

7.3. Questionnaires

In our tests, we used one questionnaire prior to the test, **Interaction tools** questionnaire, to understand what exists in current hospital systems and what are the functionalities desired; and one questionnaire after the test, **Interaction tools Post-Task** questionnaire, where we asked the user to classify our system using a 5 step likert-scale from *dislike*, value 1, to *like*, value 5, and an open comment box for each functionalities, so the users could give us feedback about their experience.

#### 7.3.1 Interaction Tools

The first questionnaire was focused on understanding the state of current hospital systems, from which we concluded that the majority of the systems have the same original tools. In the desired functionalities, we had five hypothesis, being two of them present in the same **Coordinated View** functionality; one of the others was discarded in the **Focus Group**, given that it aimed to resolve the same problem that the **Temporal View**; and the **MRI 3D Space Awareness** that was also discarded for lack of time to its development, however, 8 out of 10 users would want to have it. The **Recorded View** was not present at the questionnaire given that it was chosen exclusively in the **Focus Groups**.

#### 7.3.2 Interaction Tools Post-Tasks

The after test questionnaire was focused on obtaining a feedback from the functionalities made, espe-

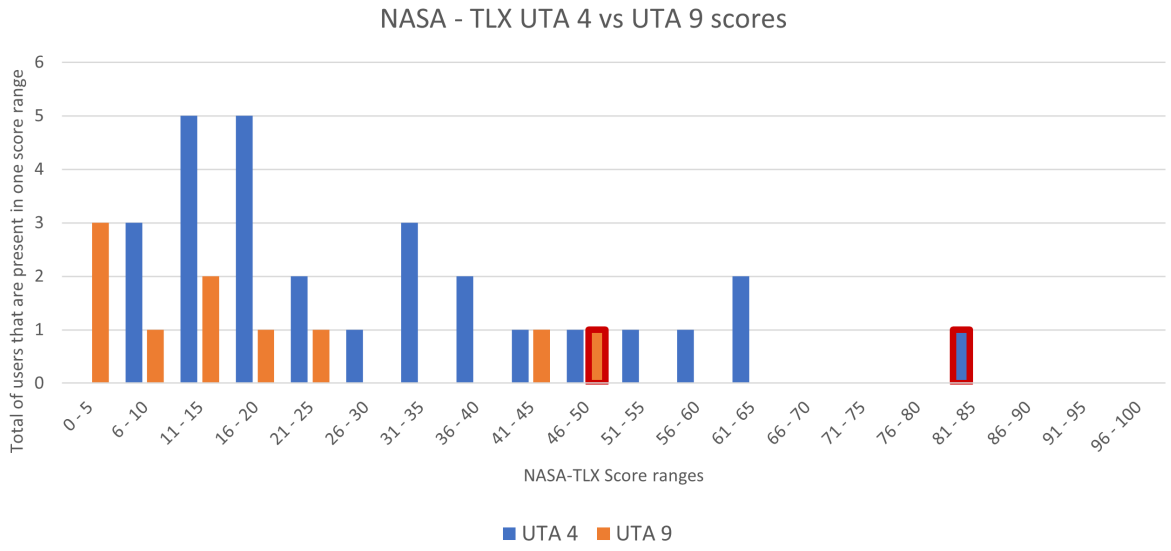


Figure 7: Results of the application of the NASA-TLX [10, 13, 23] scale in our work and the previous iteration [8].In the vertical axis, it is shown the total number of users that obtained a score in each score range, and the horizontal axis represents the score ranges by 5 score steps. In this graph we can observe that we have again a non-normal distribution with two outliers, columns with a red border.

cially from those users who did not gave any feedback during the test. In this questionnaire all physicians voted 4 or 5, from 1 to 5 in a likert-scale from *dislike* to *like*, and 70% of those responses were 5.

## 8. Discussion

With the data gathered during the **Design Process**, we were able to respond to the objectives purposed in the beginning of the this work. Those goals were defined as **Design Goals** e **Research Questions**.

### 8.1. Design Goals

As previously mentioned, three **Design Goals** were chosen that would be answered by the results obtained. Those three goals are the following: **Usability**; **Efficiency**; and **Productivity**. For the first goal, the **SUS** [6, 11] was used, where we aimed to obtain a value above 86.9, that would mean an improvement from the fourth iteration[8]. Since our results showed a total score of 91.(1), it means that our system is in the *Excellent* category and above our target.

The **Efficiency** goal was explored with two approaches, first reducing the times by using a **Time** and **Count Use of a tool** metric, and second, the total number of **non-critical errors** which had to be less than 6 for each physician. For the first approach, we aimed to reduce the time of interactions, where any type of improvement would be sufficient to reached it. We demonstrate that with our new functionalities this was possible, since the number of clicks necessary are reduced to half of what existed, consequently, reducing the overall diagnosis time. Regarding the second approach, our mean value is equal to 1.9 in **non-critical errors**, a better result that was proposed as a goal. With these two approaches proven, we can conclude that another goal was met.

Lastly, the **Productivity** goal was explored with two approaches as well, being one the **Time** and **Count Use of a tool** metrics, and the second the data from the **NASA-TLX** [10, 13, 23] to evaluate the workload of the system. For this goal the set a score lower than 29.1(4), the fourth iteration[8] score. The data collected from our tests, where we achieved a score of 14.593, indicated a *Medium* workload, proving, once again, our goal.

### 8.2. Research Questions

Here we explore how we evaluated the research questions mentioned in the beginning of this document, basing our conclusions on the data gathered during the **Design Process**.

The first research question, **RQ.1**, "When and who will use the functionalities", will be answered with the results obtained in the questionnaires, **Interaction Tools** and **Interaction Tools Pos-Tasks**, and based on the comments given by the participants. With these questionnaires we understood which functionalities the users wanted to see implemented in the system and what was their opinion after the test. The results from the post-task

questionnaire showed us that our functionalities were well received, since all of them registered a 4 or 5 classification (in a scale form 1 to 5). However, it is still impossible to determine if our functionalities would indeed be used, if they were available. We can prove that with this information, our scale results and with the opinions given by the physicians, it is probable that the hypothesis **H1.1** would be rejected. Hypothesis **H1.2** is also discarded because all functionalities developed can be used in any situation, since they are not specific to any type of difficulty. We can also conclude that, regardless of the level of the physicians' expertise, the functionalities could be used, discarding the hypothesis **H1.3**. Finally, the hypothesis **H1.4** is the one accepted for this research question, since all the functionalities were, as previously said, well received.

The hypothesis from the second research question, **RQ.2**, "What is the impact of the functionalities, in the clinic workflow?", can be validated with the responses given in both scales and metrics, **Time** and **Count use of a tool**. Regarding the hypothesis **H2.1**, the results from the **SUS** [6, 11] showed a condensed data in the higher ranges, categories *Good* and *Excellent*, proving that the usability has increased from the previous work[8] thus accepting this hypothesis. The hypothesis **H2.2**, was meant to understand how the workload was affected with the introduction of the new functionalities. The **NASA-TLX** [10, 13, 23] showed that we obtained better results than in previous **UTA**s[8], with the majority of the data points in the first five ranges, *Low* or *Medium* workload, which represents a good evolution. With these results we proved the reduction of the workload present in the system, also accepting this hypothesis. The hypothesis **H2.3**, focused on the time of a task by reducing the clicks necessary to make an action which is demonstrated by the automation that the functionalities provide, proving that the reduction of some steps can reduce the total time.

The third and last research question, **RQ.3**, "What is the best method to represent the lesion evolution?", explores the best representation for the **Temporal View** feature. We presented in the MCs project [17] a low-fidelity prototype, where we took a screenshot of the system and added annotations on top of the lesion with a timeline that could change the annotation in order to compare the lesion, however, this low-fidelity prototype showed us that this was not the right way to proceed. In the **Focus Groups** and **interviews**, physicians told us that they prefer to see the images side by side from different dates. With this type of configuration, even breast asymmetry over-time could be seen, rejecting hypothesis **H3.1**. The time bar hypothesis, **H3.2**, **H3.3** and **H3.4**, are opposite to each other, so accepting one will reject the others. Physicians gave their opinions while doing the test, and some of them were the following quotes:

> ***"I would prefer to have the most recent image in the left side, I have this in my daily system "***.
> Intern Physician

> ***"It is essential to have the most recent at the right side"***.
> Senior Physician

> ***"It is not common to compare two past images of breasts but could be a necessity to have that possibility."***.
> Senior Physician

We can see, with these quotes, that physicians have different ways to approach the problem, however, in the end, all of them agreed that having the possibility of two time bars, one in each side, could resolve the problem. Thus, the hypothesis **H3.2** and **H3.3** were rejected and the hypothesis **H3.4** was accepted. Also, hypothesis **H3.5**, is rejected by the acceptance of the **H1.4**.

## 9. Conclusions

This thesis purpose was to create a set of functionalities aiming to help physicians in the diagnosis process, by allowing them to do the same basic operations with a more automated system, making it faster and effortless. These improvements, had the users' needs as the center of our focus, by using **HCI** techniques to understand what would be the best path in the users' perspective. This work will be the base for future developments, along with an **AI** [9] and **XAI** [21] systems, that will give and explain results from the medical image analysis.

In this work we were able to improve all areas that we aimed for, making a system with better usability, workload and reducing the time necessary to complete a task. For this it was essential to use **HCI** techniques such as **interviews**, **Focus Groups**, **Affinity Diagrams**, **questionnaires** and **scales**, focused on the users' opinions of the functionalities necessary to complete a diagnosis.

Future ideas were also documented, some are small developments in the interface to make it simpler, others are functionalities that could help understand lesion characteristics and position in the body. Although one of the next developments would be merging the several iterations at work [8, 21] into a complete **CADx** system to be able to be used in real scenarios to help save lives.

## References

[1] Breast Calcifications Differential Diagnosis. `https://radiologyassistant.nl/breast/`

breast-calcifications-differential-diagnosis, 2008.

[2] BIRADS For Mammography. https://radiologyassistant.nl/breast/bi-rads-for-mammography-and-ultrasound-2013, 2014.

[3] C. Allemani, H. K. Weir, H. Carreira, R. Harewood, D. Spika, X.-S. Wang, F. Bannon, J. V. Ahn, C. J. Johnson, A. Bonaventure, R. Marcos-Gragera, C. Stiller, G. Azevedo e Silva, W.-Q. Chen, O. J. Ogunbiyi, B. Rachet, M. J. Soeberg, H. You, T. Matsuda, M. Bielska-Lasota, H. Storm, T. C. Tucker, and M. P. Coleman. Global surveillance of cancer survival 1995–2009: analysis of individual data for 25676887 patients from 279 population-based registries in 67 countries (concord-2). *The Lancet*, 385(9972):977 – 1010, 2015.

[4] C. Balleyguier, S. Ayadi, K. Van Nguyen, D. Vanel, C. Dromain, and R. Sigal. Birads™ classification in mammography. *European journal of radiology*, 61(2):192–194, 2007.

[5] M. Beaudouin-Lafon. An overview of human-computer interaction. *Biochimie*, 75(5):321–329, Jan. 1993.

[6] J. Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[7] T. Brown et al. Design thinking. *Harvard business review*, 86(6):84, 2008.

[8] F. M. 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 Tool for a Multimodality use of Breast Cancer Diagnosis on a User Interface.

[9] F. M. Calisto. Assistant introduction: User testing guide for a comparison between multimodality and ai-assisted systems. Technical Report 7, Instituto Superior Técnico, 04 2019.

[10] F. M. Calisto and J. C. Nascimento. Medical imaging multimodality breast cancer diagnosis user interface: Nasa-tlx survey template file. http://rgdoi.net/10.13140/RG.2.2.26978.79044, 2018.

[11] F. M. Calisto and J. C. Nascimento. Medical imaging multimodality breast cancer diagnosis user interface: Sus survey template file. http://rgdoi.net/10.13140/RG.2.2.24758.86088, 2018.

[12] O. J. Dunn. Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64, 1961.

[13] S. G. Hart. Nasa-task load index (nasa-tlx); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9):904–908, 2006.

[14] B. Hatscher, A. Mewes, E. Pannicke, U. Kägebein, F. Wacker, C. Hansen, and B. Hensen. Touchless scanner control to support MRI-guided interventions. *International Journal of Computer Assisted Radiology and Surgery*, 15(3):545–553, Sept. 2019.

[15] D. C. Hoaglin. John w. tukey and data analysis. *Statistical Science*, 18(3):311–318, Aug. 2003.

[16] W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, Dec. 1952.

[17] H. Lencastre. Master project: Breast cancer multimodality scalable interactions, 2020.

[18] L. C. Li, C. D. Shaw, D. Lacaille, E. Yacyshyn, C. A. Jones, C. Koehn, A. M. Hoens, J. Geldman, E. C. Sayre, G. G. Macdonald, J. Leese, and N. Bansback. Effects of a web-based patient decision aid on biologic and small-molecule agents for rheumatoid arthritis: Results from a proof-of-concept study. *Arthritis Care & Research*, 70(3):343–352, Feb. 2018.

[19] L. Marques. Cancro da mama. *Revista Portuguesa de Medicina Geral e Familiar*, 19:463–468, 2003.

[20] C. Meinel, L. Leifer, and H. Plattner. *Design Thinking : Understand, Improve, Apply*, volume 86. Springer, Berlin, Heidelberg, 2011.

[21] N. Mourão. Master project: 2d breast cancer diagnosis explainable visualizations. Master's thesis, Instituto Superior Técnico, Avenida Rovisco Pais 1, 1049-001 Lisboa - Portugal (EU), 01 2020. Introduction of explainable methods on medical decision making.

[22] W. H. Organization et al. *WHO position paper on mammography screening*. World Health Organization, 2014.

[23] A. Ramkumar, P. J. Stappers, W. J. Niessen, S. Adebahr, T. Schimek-Jasch, U. Nestle, and Y. 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.

[24] B. G. Stuijfzand, M. F. van der Schaaf, F. C. Kirschner, C. J. Ravesloot, A. van der Gijp, and K. L. Vincken. Medical students' cognitive load in volumetric image interpretation: Insights from human-computer interaction and eye movements. *Computers in Human Behavior*, 62:394–403, Sept. 2016.