A Multimodal, Usable, and Flexible Clinical Decision-Support System for Breast Cancer Diagnosis and Reporting

Suzanne Kieffer[1*], Annabelle Gouze[2] and Jean Vanderdonckt[3]

[1*]Institute for language and communication (ILC), Université catholique de Louvain, Ruelle de la Lanterne Magique, 14, Louvain-la-Neuve, B-1348, Belgium.

[2]EVS Broadcast Equipment, Rue Bois Saint-Jean, 13, Seraing, B-4102, Belgium.

[3]Louvain Research Institute in Management and Organizations (LouRIM), Université catholique de Louvain, Place des Doyens, 1, Louvain-la-Neuve, B-1348, Belgium.

*Corresponding author(s). E-mail(s): suzanne.kieffer@uclouvain.be;
Contributing authors: annabelle.gouze@gmail.com; jean.vanderdonckt@uclouvain.be;

Abstract

To support radiologists in their decision-making regarding women breast cancer diagnosis based on mammographies, this paper addresses three challenges posed by a Clinical Decision Support System (CDSS) for breast cancer screening, diagnosis, and reporting: multimodality (via textual, graphical, and two-dimensional gestural interaction), usability (via human-computer interactions methods such as knowledge elicitation interviews, scenario-focused questionnaires, multi-fidelity prototyping and user testing), and flexibility (via selected image processing techniques ensuring manual and semi-automatic annotation of breast cancer findings based on radiologists' gestures).

Keywords: Clinical decision-support system, breast cancer, user-centred design, image processing, stroke gesture-based interaction, relevance feedback.

1 Introduction

Breast Cancer (BC) is the most prevalent cancer among women worldwide [51]. In 2020, 2.3 million women were diagnosed with breast cancer and 685,000 deaths according to the lastreportoftheWorldHealthOrganization. At the end of 2020, 7.8 million women had breast cancer in the last five years. Women are affected by breast cancer regardless of country, race, and age [14]. Early detection of BC improves treatment and recovery, and several programs are carried out worldwide and nationwide for this purpose. For instance, such a program named "Mammotest" started in Belgium in 2001. Screening mammography programs are promoted to facilitate the early detection of breast cancer for women 50 up to 70 of age. Currently, 80% of women with cancer diagnosed early are cured compared to 30% otherwise.

There are at least two promising ways to help improving BC treatment: the screening mammography examination and the usage of a Clinical Decision Support System (CDSS) for breast cancer [33]. The mammography screening examination aims to discriminate BC findings into benign and malignant lesions. On the other hand, CDSS helps detect missed findings in the screening procedure [41]. Automatic detection tools based on image processing are called *Computer-Aided Diagnosis* (CAD) systems.

Mammography screening is the most reliable examination, as it allows the early detection of BC disease and its treatment. The screening workflow consists of the radiologist in image acquisition, image analysis, lesion detection, interpretation, and reporting. The report includes a complete and structured description of any significant findings, comparison with previous studies, general diagnosis, and follow-up recommendations. Typically, image acquisition results in four X-ray images, or four standard projections of the breast: left/right craniocaudal (CC) and left/right mediolateral oblique (MLO) projections (Fig. 1). Recent advances in acquisition system allow to directly produce digital mammography instead of X-ray film. The benefit of digital mammography over screen-film mammography is already demonstrated [48, 54].

As soon as digital mammography replaced screen-film mammography, information technology has been progressively introduced in BC screening and diagnosis. Both the number of devices and the number of interactive tools have increased significantly over the past decades not only to improve the reliability of the diagnosis (*e.g.*, with digital mammography, the magnification, orientation, brightness, and contrast of the image may be altered after

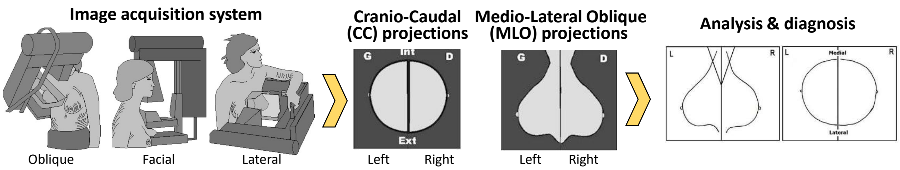


Fig. 1: Image acquisition system and its projections.

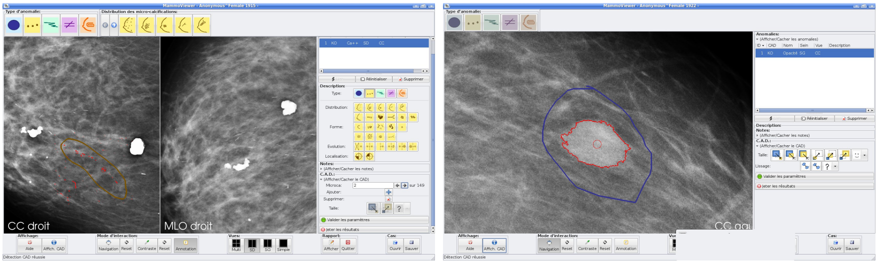


Fig. 2: CAD system user interface.

the examination is completed to help the radiologist more clearly see certain areas), but also in order to improve the productivity of the radiologist (*e.g.*, new information and communication technologies have improved and facilitated the access and the sharing of medical information [44]).

However, radiologists still deplore several shortcomings of the existing CDSS for BC, which we synthesize under the umbrella of three software quality factors: multimodality, usability, and flexibility.

Multimodality or multimodal interaction refers to the "interaction with the virtual and physical environment through natural modes of communication such as speech, body gestures, handwriting, graphics, or gaze" [5]. Beyond textual and graphical modalities, radiologists express their wish to also interact with gestures, as this technique enables them to perform a direct manipulation of mammography elements. For example, gestures captured by a pen or a similar pointing device directly on the regions of interest of the mammography would be more appreciated than a "point-and-click" interaction that forces them to alternate constantly between the menus and the regions of interest, with little or no support for direct manipulation.

Usability is defined as the "degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" [22]. Several aspects should concur towards an overall usability of the CDSS. Radiologists need a unique and integrated system that supports their activity to maximize the duration of image viewing and minimize any distractions from both the equipment and software [2]. Recent efforts have been made to develop and improve digital mammography equipment from an image acquisition system (Fig. 1, left) to image viewing and analysis systems (Fig. 1, right). The setting for screening procedures has to be compliant with the European guidelines for quality assurance in BC screening and diagnosis [1] and is most often composed of two screens, a task-oriented keypad and a mouse. Furthermore, due to the specificity and the complexity of each task involved in mammogram analysis, research efforts have focused mainly on the implementation of task-oriented interactive systems such as image viewers, radiology information systems (which are used to store, manipulate, and distribute the patient radiological data and imagery), CAD software, and digital case databases. Hence, radiologists distribute their work between a growing number of interactive tools (*e.g.*, case retrieval databases, viewers, and CAD tools), workstations, and media (*e.g.*, screen, mouse, keyboard, and handheld recorder). This equipment heterogeneity decreases its productivity.

Furthermore, radiologists require that the BC findings be compliant with the domain standard. Although many research efforts have been carried out separately on the implementation of task-oriented systems, much less effort has been invested in the design and development of technologies that comply with the domain standard *i.e.*, the Breast Imaging Reporting And Data System (BI-RADS). The BI-RADS is an approved quality assurance system of descriptive terms and reporting guidelines [2]. Such tools not only facilitate reporting, providing radiologists with homogeneous, structured, and standardized reports, but also lead to data accessibility. They enable data exchange and storage, interpretation monitoring [56], retrieval of useful and interesting cases for teaching and research purposes [57].

Flexibility or Adaptability is defined as the "degree to which a product or system can be effectively and efficiently adapted to different or evolving hardware, software, or other operational or usage environments" [22]. For this purpose, CAD systems (Fig. 2) should enable radiologists to modify any detected area, in particular when its identification is not accurate. Mammography is the most commonly used method for BC screening since it enables early detection and treatment. Two important and early signs of the disease in breast tissues are a cluster of microcalcifications (small calcium deposits) and masses. The current CAD tools provide radiologists with the automatic detection of possibly missed findings, such as cluster of microcalcifications (CA++) or masses [9]. However, radiologists deplore the limited adaptability of CAD tools. For instance, BC-oriented interactive tools should support both the association of numerical data from the CAD to the standard annotation of BC findings and the modification of these numerical data if required, with the ultimate goal to include these data to the final report. A prototype combining manual and automatic features to describe breast lesions was already proposed [18]. Based on this combination, our solution integrates the complete standard lexicon of mammography terms defined by the BI-RARDS for the manual characterization of the findings [2]. Furthermore, this work has taken into account human abilities, skills, and requirements when designing the entire system in a new usability engineering method that combines several levels of fidelity and goes beyond the simple consideration of usability guidelines [52].

In summary, a CDSS for BC should primarily satisfy three quality properties: multimodality, usability, and flexibility, *e.g.* by integrating domain-oriented functionalities within a unique platform that supports manual and semi-automated annotation of findings through various modalities while preserving the overall usability of the system. Other quality properties are addressed elsewhere. For example, Esmaeili *et al.* [15] showed that a k-nn classifier reaches a prediction accuracy of findings of up to 84%. Massafra *et al.* show how to ensure a prediction of cancer recurrence based on a CDSS [36].

To this end, this paper combines Human-Computer Interaction (HCI) and image processing methods to develop a multimodal, usable, and flexible domain-oriented system for domain-expert users. HCI methods include elicitation of user requirements, design of multi-fidelity prototypes, and usability testing. Image processing methods include lesion detection, image segmentation [23] and relevance feedback. We present how the different methods and techniques borrowed from both fields were selected, customized, and combined to achieve the CDSS for BC. While other CDSS focus on detection accuracy [15], an important quality property, our CDSS is aimed at satisfying multimodality, usability, and flexibility as the first software quality criteria, which, to the best of our knowledge, were not explicitly addressed in related work (see [33] for a systematic literature review of CDSSs for BC).

The remainder of this paper is structured as follows: we first present how the combination of selected HCI and image processing techniques contributed to our CDSS for BC; we then describe the main features of the system emerging; we conducted three rounds of usability testing to assess the aforementioned quality factors, and we finally discuss the results and conclude by suggesting some future avenues to this work.

2 Combining HCI and Image Processing

2.1 Involvement of HCI in our CDSS

The HCI methods involved in this work concern the achievement of usability through the adoption of a user-centred design approach to develop our CDSS. The development of computer technologies provides nowadays a means to support and facilitate the daily activities of potentially all users. This may be of particular importance to experts in BC screening and diagnosis, under the condition that the system fits user needs, expectations, and requirements. Interest in user-centered approaches stems precisely because of the need to design and implement interactive systems supporting the activities of domain-expert users, who are not necessarily experts in computer science. Throughout the software development life cycle, important care must be devoted to the study of the specific needs and expectations of such domain-expert users [10]. Furthermore, attention must be paid to the utility and usability of the system to develop meaningful and usable interfaces [21], systems that people actually want [40]. We achieved utility by integrating both equipment and software into a unique platform that supports the entire screening mammography workflow from image acquisition to reporting. We achieved usability by adopting a user-centred design approach to develop our CDSS early in the software development life, specifically by employing knowledge elicitation interviews, scenario-focused questionnaires with paper mock-ups, and user tests.

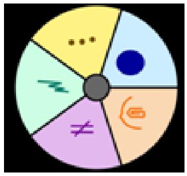
Screening analysis activity	Scenario #1	Scenario #2
Description of the lesion type Masses, Calcifications (Ca++), Architectural distortion, Special cases, Associated findings	Array of buttons 	Pie menu 

Table 1: Scenario-focused questionnaire. The end-user activity is the description of the lesion type (col. 1). Two interactive scenarios support this activity: using an array of buttons or using a pie-menu (resp., col. 2 and 3).

2.1.1 Knowledge Elicitation Interviews

We collected domain- and task-relevant knowledge early in the life cycle thanks to knowledge elicitation interviews. Five domain-expert users were questioned thoroughly about the BC domain, the task series involved in their activity, their needs and expectations with the goal to implement the collected information in the system. The equipment used was paper notes and video recording. The information collected during these interviews includes the description and explanations of: the terminology to be used, the setting of the equipment, the acquisition procedure, and the screening analysis procedure. The screening analysis procedure can be decomposed as follows:

1. Select case from list of patients
2. Analyse patient information, such as previous examinations if available
3. Visualise projections on reading viewers, by following this sequence: global view, 2 CC views, 2 MLO views, and global view again
4. Visualise zoomed-in parts of projections to detect suspicious findings
5. Interpret
6. Synthesise screening results and report.

2.1.2 Scenario-focused Questionnaire and Paper Mock-ups

Based on the information collected during knowledge elicitation interviews, we developed a scenario-focused questionnaire [55] and paper mock-ups (Fig. 3) in order to identify which scenarios would best support human work. For example, Table 1 shows two potential scenarios that are intended to support the description of the type of lesion. We presented the paper mock-ups with different scenarios to six domain-expert users and used the questionnaire as visual aid during the sessions. This allowed us to identify potential interactive scenarios to develop later in the software development life cycle and to validate informally the overall spatial organization of the user interface, the terminology and jargon, and specific components such as icons. We used the same method to develop the semi-automatic detection tool and the final report. The material used for this purpose is available in [19].

Fig. 3: Mock-up during design.

Paper mock-ups (*i.e.*, low-fidelity prototypes) were preferred to coded prototypes (*i.e.*, high-fidelity prototypes) since the evaluation should lead to a lot of drawings, direct manipulation of paper components, and discussions between designers and domain-expert users [46]. Each low-fidelity prototype progressively evolved to a high-fidelity prototype by mapping [38], thus resulting in multifidelity prototyping [11]. Due to the participation of the end user, this design method has been shown to reduce development time and cost and improve usability [42, 55].

2.1.3 User Testing

User testing is recognized as a first essential step to conduct user-centered design [35, 37]. We adopted a formative approach that consists in carrying out user testing in several rounds to detect and fix usability issues with the prototype. As such, we carried out three rounds user testing. The first round (section 4) aimed to compare the usability of the pie-star menus with the array. The second round (section 5) aimed at evaluating the usability of the mass segmentation and the relevance of the features chosen for user feedback. The third round (section 6) aimed to evaluate the usability of microcalcification detection and segmentation within the mammography viewer. Such incremental and iterative process allowed us to test-and-refine one feature before moving to developing the next.

2.2 Involvement of Image Processing in our CDSS

The main contribution of image processing to this work is the integration of both manual and semi-automatic means to characterize any BC finding in compliance with the approved standard of the domain. Specifically, we improve image processing techniques such as mass segmentation thanks to a relevance feedback about the results performed by domain-expert user. CAD systems were initially installed to attract radiologists' attention to features they might have overlooked or dismissed by highlighting potential BC lesions with markers such as crosses, triangles, or squares. However, the numerical data arising from the CAD are restricted to the spatial location of the lesions and need to be enhanced by complementary features from segmentation to provide useful information for reporting. Complementary features include

- For masses: segmented area (*i.e.*, area surrounding the mass), location, size, circularity, local contrast, perimeter, mean, and luminance;
- For clusters of calcification's: centre location, number of calcifications, length of the convex hull, size of the convex hull, circularity of the convex hull, density, and local contrast;
- For microcalcifications: location, size, circularity, and local contrast.

Furthermore, the segmentation accuracy depends on different factors such as breast composition, missing lesion edge due to the presence of a more marked edge, and missed/false findings. Therefore, segmentation can provide outputs that do not correctly represent the lesions and may not be meaningful. Consequently, it is mandatory to propose a new interactive method which allows exploiting the radiologist knowledge inside the segmentation process: relevance feedback [43].

2.2.1 Segmentation of Masses

Mass segmentation is performed applying the Kupinski and Giger algorithm [28], which is based on a region growth process and on probabilistic models. The region growing process starts from a source point, located inside of the ROI where the segmentation is computed. Then, this region iteratively grows towards the edge of the object to be segmented, while the pixels bordering the current region are integrated inside the region upon condition that the luminance variation is lower than a given threshold. Different thresholds are applied to obtain different partitions in two regions: the object (lesion) and the background. The probability distribution of the luminance inside and outside the object is computed. The probability of the image partition is equal to the product of probabilities at each point. The final mass segmentation corresponds to the partition that maximizes the probability of the partitions.

2.2.2 Detection of Microcalcifications

Microcalcifications are detected by applying Dengler's algorithm [13] and grouped into clusters, considering the distance between them. Dengler's algorithm combines the Weighted Difference of Gaussian (WDoG) and the Top

Hat morphological transform. The size of the Gaussian filters used in the WDoG method depends on the expected size of the microcalcifications and the local contrast. Some modifications were made in the application of the WDoG method by combining several Gaussian sizes to facilitate the detection of microcalcifications of various sizes. This approach improves the detection results. Clusters are defined by both the ROI delineated by the radiologist and the distance of microcalcifications to each other and are delimited by computing the convex hull.

2.2.3 Relevance Feedback

Relevance feedback techniques focus on adjusting features, using the feedback provided by the radiologist about the previous results in order to improve or validate the system performance [43]. To improve segmentation, relevance feedback requires, for each finding, the selection of the features that best suit the acquisition of radiology knowledge and segmentation. Then, for each chosen feature, the predefined feedback as well as the actions associated with each feedback are introduced.

3 Resulting CDSS Features

3.1 Two-dimensional Stroke Gestural Interaction

Gesture interaction is an effective method of providing end users with a natural, intuitive, and convenient interaction [55], in particular 2D stroke gestures in which unistrokes and multistrokes are issued on a surface [34]. Due to its high naturalness and mainly to its convenience in satisfying the lesion characterisation requirement, gesture interaction with a graphics tablet and a pen was chosen as the modality to interact with the system: navigate in a clinical case (*i.e.*, among mammograms), navigate in a specific mammogram (*i.e.*, zoom-in/out), sketching a region of interest (ROI) [25], annotating findings and reporting.

3.2 Iconic Representation of the BC Domain Standard

The BI-RADS [2] provides a standardised terminology for the description of BC findings. Any finding is described according to a lesion type (*i.e.*, mass, calcification, architectural distortion, special case or associated finding), and type-related characteristics. Beyond the specific characteristics related to a lesion type, the breast imaging report contains the finding location and the comparison with previous studies, whatever the type. We chose the BI-RADS as it is common knowledge among BC radiologists.

We created an icon framework according to this standard to allow any findings to be fully described (see the complete icon catalog in Appendix 7). Every single term of the BI-RADS lexicon is represented by a unique icon. The complete set comprises about 150 different icons to characterize the lesion in a straightforward and unambiguous manner. A color code was adopted

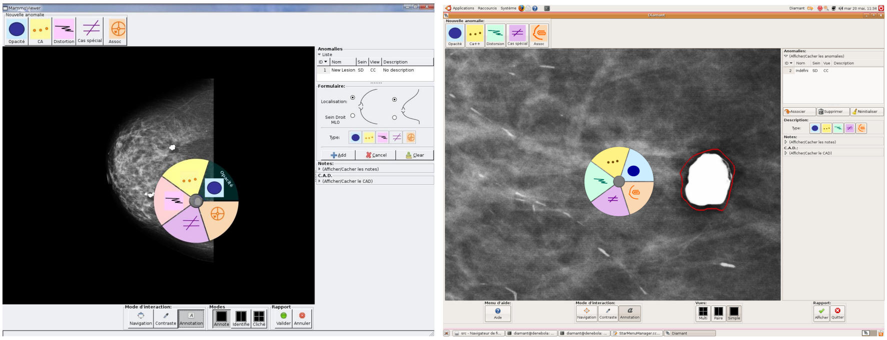


Fig. 4: Pie menu in the mammography viewer. Also, see video.

to facilitate discrimination between the findings: masses in blue, calcifications in yellow, architectural distortions in green, special cases in violet, and associated findings in orange. The icons' schemes related to finding location and the comparison with previous studies are common to all lesion types; only colors are different. For example, Fig. 15 shows the icons related to the specific characteristics of the mass. Masses are characterised by basic shape (round, oval, lobular, or irregular), margin (circumscribed, microlobulated, obscured, indistinct, or spiculated) and density (high-density, equal density, low-density or fat-containing radiolucent). The same reasoning holds for calcification (Fig. 16), architectural distortion (Fig. 17), special case (Fig. 18) or associated finding (Fig. 19). Furthermore, all corresponding subsets are defined in a consistent way. For example, the six cases of evolution, *i.e.*, "Statu quo", "New", "Change", "Enlarging", "Contracting", and "Not available" represent their actions in the same way to preserve consistency, but on different targets, *e.g.*, mass icons in Fig. 15 and calcification icons in Fig. 16. Consistency is similarly preserved for other aspects, like "Location". Since each icon corresponds to a unique term of the BI-RADS lexicon, which is assumed to be a pre-requisite learned by radiologist in their background, they are familiar with the terminology and the icons are therefore considered as self-explanatory. Radiologists should therefore not learn them while using them for the first time, but instead recognize them in their background. The icon structure in the menus make them rapidly selectable and accessible.

3.3 Manual Annotation with the Pie and the Star Menus

Providing experts in BC imaging with an interactive tool that supports their activity is a difficult problem for HCI considering its utility and usability [21, 40]. Utility is ensured by the compliance of the system with the domain standard and by the integration within a single interactive tool of the functionalities supporting the entire screening mammography workflow.

To achieve usability, especially during the manual annotation of the BC findings, we paid sustained attention to the graphical representation of the mammographic BI-RADS terminology (by the design of an exhaustive set of

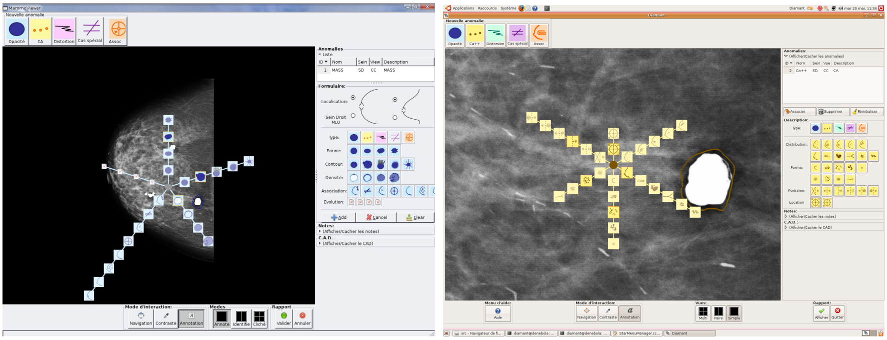


Fig. 5: Star menu in the mammography viewer. Also see a video.

icons), to the spatial organization of multimedia data (not only the overall spatial organisation of the user interface, but the specific location of widgets as well), and to the design of new interactive solutions suited to the finding annotation with a pen on a graphics tablet. Therefore, two menus were implemented to support the gesture-based annotation of the BC findings: a pie menu [6] for the gesture-based selection of the type of lesion and a star menu for the gesture-based description of the type-related characteristics of the lesion.

The pie menu (Fig. 4) was implemented to facilitate the pen-based annotation of the type of lesion. This format was chosen because it reduces the target seek time and improves the accuracy of target selection [6, 39]. The star menu (Fig. 5) was implemented to facilitate the pen-based annotation of complementary characteristics by grouping related icons. This format was chosen because this display layout was shown to be very efficient and accurate for visual inspection or visual detection compared to matrix, elliptic, and random spatial structures [47]. Manual annotation of any significant finding can be performed as follows. First, the suspicious finding is surrounded with a pen. This action corresponds to the ROI drawing. Then, the lesion is annotated thanks to the pie-menu (type of lesion) and the star-menu (characteristics of the specific lesion). Finally, the complete description of the lesion is automatically stored in the final report.

3.4 Semi-automatic Annotation and Relevance Feedback

While radiologists manually annotate a finding, segmentation computation and numerical feature extraction are automatically initiated inside the ROI. The segmented findings are displayed in red (highlighted edge) together with relevant numerical features of the segmented objects. The system enables, if required, to “highlight” or to “hide” the edge of the detected mass, the mask of detected microcalcifications or the convex hull of clusters, by simply pressing a button. It also provides users with a simple query interface with two modes: (i) validation mode, which enables either the validation or the report of mistakes, and (ii) (interactive) segmentation mode, which enables

Feature	**Icon**	**Relevance Feedback**
Area		Reduce (lightly, moderately, highly): add corrector term to optimal region growing threshold
		Enhance (lightly, moderately, highly) : subtract corrector term to optimal region growing threshold
Location		Displace: recompute with new seed point
		Erase: erase selected mass
Smoothness		Sharpen: no action if user did not smooth over the edge, undo step of smooth filtering
		Smooth over: voting filing filter to smooth edge

Table 2: Predefined features of masses (col. 1), related icon (col. 2) and relevant feedback with associated actions of the system (col. 3) (adapted from [27], used with permission).

Feature	**Icon**	**Relevance Feedback**
Localization		Add: click inside the lesion to select a position and press the icon
		Erase: erase the calcification with the pen or click inside the lesion and press the ´´Delete" icon

Table 3: Predefined features of microcalcifications for a cluster (col. 1), related icon (col. 2) and relevance feedback with associated actions of the system (col. 3) (adapted from [27], used with permission).

radiologists to improve the segmentation by providing new parameters for the segmentation computation. This is called relevance feedback.

Tables 2 and 3 show how we implemented relevant feedback for masses and microcalcifications respectively. We designed a set of icons to enable radiologists to easily interact with the system. Upon validation of the segmentation results, numerical features and manual annotations are included in the report. However, radiologists are unable to evaluate these numerical values themselves or to return any numerical feedback. Radiologists can only provide evaluations of good or erroneous results and return their subjective feedback. In this way, positive feedback implies a validation of the features, while negative feedback launches a new segmentation with new parameters. Since negative feedback involves immediate action from a cognitive point of view, the feedback is designed to suggest an action rather than an assertion. For example, if the radiologist judges the mass area to be incorrect and returns a feedback to "lightly enhance region", the system allows the interactive icon-based specification of a different area of a bigger size. The associated action is to subtract a corrector from the optimal threshold of the region-growing algorithm [28] to obtain a larger area. The validation mode enables radiologists either to validate all segmentation results at once by clicking on the corresponding button, or to validate each feature separately. Validated features are the same as those defined by the interactive segmentation mode. Possible feedback are nominal and numerical values. The nominal value (1-7) is either a validation or a comment to invalidate incorrect features, as the feedback from the validation mode allows the results to be commented on. For instance, the detected area can be commented as to be "too small".

3.5 Reporting

The ACR promotes the BI-RADS standard to solve problems of nonuniform, vagueness and inconclusive reporting [2]. BI-RADS provides a uniform verbal description of any significant finding and clear, concise, and directive reporting to reduce the complexity and variability of interpretation and decision-making. The BI-RADS improves feature analysis agreement with experienced breast imagers [4]. The report incorporates a structured description of any significant finding, comparison with previous studies, overall impression, and follow-up recommendations. Reporting starts with manual and semi-automatic annotation of findings and ends with the filling out a form (Fig. 6). The form displays for each breast the summary of the annotation and the breast density according to four types standard :

- *Type 1*=breast almost entirely fat
- *Type 2*=scattered fibroglandular densities
- *Type 3*=breast tissue heterogeneously dense
- *Type 4*=breast tissue extremely dense.

The European guidelines for quality assurance impose radiologists to evaluate the quality of each view. Then the radiologist's final impression for each breast has to be classified among six assessment categories: need imaging evaluation (ACR0), no finding (ACR1), benign finding(s) (ACR2), probably benign finding(s) (ACR3), suspicious abnormality (ACR4), and highly suggestive of malignancy (ACR5). The final recommendation consists in decision-making, which is simplified thanks to the ACR assessment categories [2].

4 First Round of User Testing: Pie-Star Menus *vs.* Array

We carried out this first round of user testing to compare the usability of the pie-star menu with respect to the array menu.

Fig. 6: Form for reporting (taken from [27], used with permission).

4.1 Participants

Nine volunteers (3 female, 6 male, and 0 identified as gender variant/non-conforming) aged 43 to 58 years participated in this study. All participants were experienced breast radiologists practicing in different hospitals in Belgium. We recruited them based on their experience in BC screening. A background questionnaire allowed us to assess their computer skills: all participants were familiar with computers, especially medical computer-based

Fig. 7: Array of icons: Lesion type (left), Distribution of a calcifications (right).

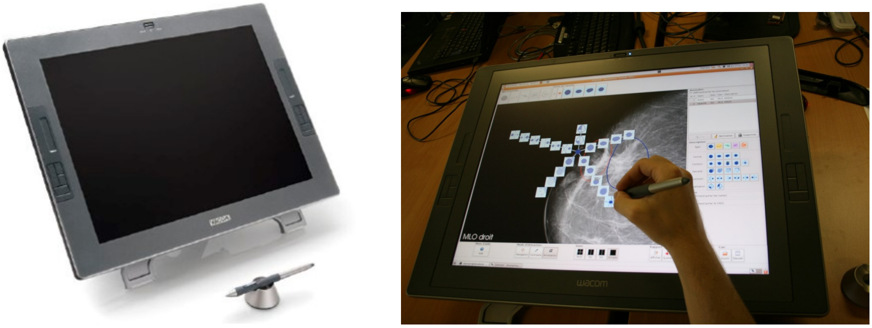


Fig. 8: The Wacom CINTIQ tablet (left) and the viewer loaded (right).

applications; all were experienced in visual search and navigation activities on computer displays; all were average mouse and keypad users with similar motor abilities.

4.2 Tasks

We asked participants to annotate BC findings in clinical cases using pie-star menus or an array of icons located at the top of the user interface. The array icons and the pie-star menu icons are the same in terms of scheme, colour and size. The array allows for the sequential selection of the characteristics found, according to the following order: lesion type, type-related characteristics (each characteristic must be described one at a time), comparison with previous studies, and location. Fig. 7 presents the array of icons displayed to describe the type of lesion (left) and the distribution of the calcifications (right).

4.3 Experimental Setup and Material

We used a 2×5 factorial design with two conditions (pie-star menu *vs.* array) and five tasks (medical cases to characterize). Each participant carried out ten tasks, five per condition. The order of conditions was counterbalanced between participants according to a 2×2 Latin Square design and the order of medical cases was randomized. We used counterbalancing and randomization to neutralize possible task learning effects and to control inter-individual diversity. We carried out the tests in an isolated room in each hospital. Participants were seated approximately 40 cm from the graphic tablet. 2D stroke gesture-based annotation was the input modality, visual display the output modality. The computer system had the following technical characteristics: Intel Core2 Duo E8400 (3GHz) processor, 4 GB of DDR SDRAM and a 9600GT Nvidia graphic card. The gesture tablet was a Wacom CINTIQ 21UX (Fig. 8, left), a set-up that was previously effectively tested for 2D stroke gesture recognition [12]. DICOM images were loaded into the viewer (Fig. 8, right).

Factors	Values	N	M (sec.)	SD (sec.)	CI (sec.)
Condition	Pie-Star	114	17.3478	12.8070	2.3407
	Array	105	20.5524	12.3304	2.3584
View	CC	138	20.5000	1.0629	0.1773
	MLO	81	16.3457	1.3874	0.3021
Finding type	Mass	69	16.2609	1.4949	0.3527
	Calcification	89	20.9438	1.3162	0.2734
	Arch. Distortion	39	17.6410	1.9883	0.6240
	Special case	5	31.6000	5.5531	4.8674
	Associated finding	17	18.8824	3.0116	1.4315

Table 4: Annotation times in seconds per condition, view and finding type: mean (M), standard deviation (SD), and confidence interval of 95% (CI).

Factors	Values	N	M	SD	CI
Finding type	Mass	69	5.4637	0.2133	0.0503
	Calcification	89	4.5842	0.1878	0.0390
	Arch. Distortion	39	4.0256	0.2837	0.0890
	Special case	5	4.0000	0.7924	0.6945
	Associated finding	17	6.5882	0.4297	0.2042

Table 5: Number of clicks per finding type: mean (M), standard deviation (SD), and confidence interval of 95% (CI).

4.4 Procedure

The test sessions involved one volunteer at a time. First, we gave participants an oral presentation of the project, an explanation of their role in the user test and a demonstration of the functionalities of the system. Then, participants started the training session: one clinical case to annotate per experimental condition. Once they felt comfortable enough with the system and the manipulation of the pen, we provided them with paper printed instructions and demographic and background questionnaires to fill prior to the effective test. We logged the task completion time, the number of clicks, and satisfaction, three standard usability metrics [21, 40]. After each condition, we administered a satisfaction questionnaire. After the two conditions, we administered IBM CSUQ [29, 31], a 19-item questionnaire aimed at evaluating the usability of a system in terms of System Usefulness (SysUse), Information Quality (InfoQual) and Interface Quality (InterQual) on a 7-point Likert scale [32]. A debriefing ended the session. The sessions lasted approximately 30 minutes.

4.5 Task Performance

Overall, we collected 219 annotations. We performed an analysis of variance (ANOVA) to determine any potential significant difference in task performance, measured in terms of annotation time (Table 4) and the number of clicks (Table 5). We present and discuss the results per condition (pie-star *vs.* array), per view (CC *vs.* MLO) and finding type (mass, calcification, architectural distortion, special case, and associated finding).

4.5.1 Pie-star Menu *vs.* Array Menu

Data show no statistical differences between pie-star menus and the array, with respect to neither the annotation times (Table 6: F=3.5605; p=.0605, *n.s.*) nor the number of clicks (Table 6: F=.0216; p=.8832, *n.s.*). The pie-star menus neither increase nor decrease annotation time and number of clicks compared to the array. The pie-star menu, though more innovative and less common than the array, does not seem to introduce noise during the annotation task.

However, annotation with pie-star is a bit faster than with array (Table 4 pie-star: 17.5 sec *vs.* array: 20.5 sec). This may be due to the spatial organization of icons: by simultaneously displaying all icons related to a finding type on radial lines, the star menu allows for parallel processing, which enables users to anticipate their next click and to be faster. By contrast, by displaying one icon at a time, the array forces linear processing, which prevents users from anticipating their next click and slows down annotation.

4.5.2 CC *vs.* MLO

Data show a significant view effect with respect to annotation times (Table 6: F=5.6496; p=.0183*) and there is no statistical difference between CC and MLO in terms of the number of clicks (Table 6: F=1.8155; p=.1792, *n.s.*). The annotation is faster in the MLO projection than in the CC projection (Table 4: MLO: 16.34 sec. *vs.* CC: 20.5 sec.). This difference may be explained as follows: in practice, radiologists start the diagnostic by analysing and interpreting the CC view. It seems natural that users adopt the same task order with the interactive tool and it seems normal that the examination lasts longer in the first projection (CC) than in the second one (MLO). Second, masses and calcifications need to be characterized in both views. Thus, a “Duplicate” button was implemented to reduce the number of clicks required to complete the annotation of the findings. Consequently, for masses and calcifications, there were fewer properties to be annotated in the second projection (MLO).

4.5.3 Effect of Finding Type

Data show a significant effect of finding type on annotation times (Table 6: F=2.7884; p=.0274*) and a highly significant effect of finding type on the number of clicks (Table 6: F=9.0073; p<.0001***). As can be seen from Table 4, the annotation of masses (16.26 sec.), architectural distortions (17.64 sec) associated findings (18.89 sec.) or calcifications (20.94 sec.) is on average much faster than the annotation of special cases (31.6 sec.). The important difference observed between special cases and other finding types may be explained as follows: special cases are unusual and therefore may require more time to be examined. The relative difference observed between calcifications and the triplet {mass, architectural distortion, associated finding} may be explained

Factors	*df*	Annotation time (sec.)					Number of clicks				
		F value	*p*-value	Sig.	Cohen's *d*	Inter.	*F* value	*p*-value	Sig.	Cohen's *d*	Inter.
Condition	1	3.5605	0.0605	*n.s.*			0.0216	0.8832	*n.s.*		
View	1	5.6496	0.0183	*	3.3615	small	1.8155	0.1792	*n.s.*		
Finding type	4	2.7884	0.0274	*	3.3250	small	9.0073	0.0001	***	4.3766	small

Table 6: Analysis of variance with factors: experimental condition, view, and finding type. No significance=p>.05, $p\leq.05^{*}$, $p\leq.01^{**}$, $p\leq.001^{***}$.

by the number of characteristics to describe, which differs between the types of finding: 6 for calcifications, 5 for masses, and 3 for architectural distortions, special cases, and associated findings.

The results on the number of clicks are shown in (Table 5). Unsurprisingly, only a few clicks on average are required to annotate special cases (4.00) and architectural distortions (4.03). Further, the annotation requires in average 4.58 for calcifications, 5.46 for masses and 6.59 for associated findings. The fact that masses require more clicks than calcifications (5.46 *vs.* 4.58, respectively) is surprising, considering there are fewer characteristics to describe for masses than calcifications (5 *vs.* 6 respectively). When looking at the number of characteristics to describe, even more surprising is the important number of clicks required to annotate the associated findings (6.59). This results may be explained by hesitations triggered by masses and special cases, which led to more mistakes and auto-corrections.

4.6 Satisfaction and Preferences

During the interviews, participants considered the interaction with the system as natural, intuitive, and reliable. Most of the participants (8) hesitated less than five times and all of the participants were satisfied with the compliance with BI-RADS. Five participants expressed positive judgments about the star menu in terms of information visualization, speed, and comfort. They preferred the star menu because "it enables the parallel visualization of the items due to its spatial organization", "it is more comfortable due to its position close to the center of the screen", and "it is faster (than the array)". Four participants preferred the array because "it is usual" and "the characteristics follow a logical sequence".

These results are consistent with the CSUQ results (Fig. 9): SysUse (M=5.46, SD=0.96), InfoQual (M=5.56, SD=1.11), InterQual (M=5.81, SD=0.88), and Satisfaction (M=5.56, SD=1.89) are all above 5, which is considered the reference threshold for this level. Furthermore, a Wilcoxon signed rank test for a single sample with one tail suggests that all metrics are significantly above the median value of 4 in a positive manner: SysUse (z-score=6.82, $p<0.001^{***}$) with a large effect size (Pearson's r=0.80), InfoQual (z-score=6.18, $p<0.001^{***}$) with a large effect size (Pearson's r=0.78), InterQual (z-score=4.43, $p<0.001^{***}$) with a large effect size (Pearson's r=0.85), and Satisfaction (z-score=2.66, $p=0.0019^{***}$) with a large effect

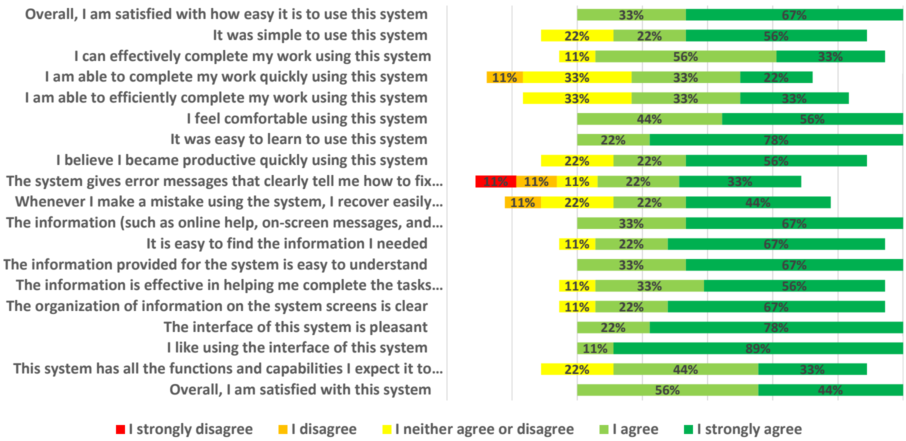


Fig. 9: Distribution of answers to IBM CSUQ for the first round.

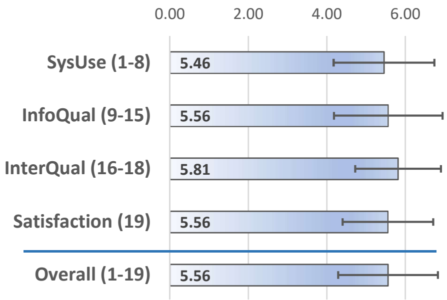


Fig. 10: IBM CSUQ metrics for the first round. Error bars show a confidence interval of 95% (α=0.05).

size (Pearson's r=0.88). The usability of the application is assessed very positively by the participants.

Although the results of the testing are consistent with the results of the IBM CSUQ, the study involved only nine participants, which represents a small sampling of the population. For the dichotomous variables tested (*e.g.*, pie-star menu *vs.* array menu and CC *vs.* MLO), a binomial (P=.5, W=.1, CI=95%) and normal (α=.025, Z_α=1.96) approximation ([20], Appendix 6E, p. 81) to the sample size gives a number of participants expected to be N=402, respectively N=384. Finding and involving such a large number of radiologists is unfortunately impossible. Therefore, the results should be moderated depending on our number of participants and the 219 annotations produced. The small effect size obtained in the testing (Table 6) should be counterbalanced with the large effect size obtained in the CSUQ questionnaire. The high values of Cohen's d coefficients could suggest that the group means differ by 3 or 4 standard deviations.

5 Second Round of User Testing: Mass Segmentation

We carried out this second round to evaluate the usability of the mass segmentation and the relevance of the features chosen for user feedback.

5.1 Methodology

Round 2 took place in the radiology department of a hospital in Belgium. We recruited three radiologists (3 male, who were different from the initial 9 participants involved in the first round to avoid any carry-over effect) who annotated 32 masses each: ten circumscribed masses, seven microlobulated masses, one obscured mass, three indistinct masses, and 11 spiculated masses. The task consisted first in drawing a ROI around each mass to run the segmentation and second in providing a relevance feedback by either rerunning the segmentation or validating the results (Fig. 11). Participants were able to require new segmentation solutions until they were satisfied with the results. Finally, they could return a validation feedback and their assessments of the final segmentation.

Data collection methods included a think-aloud protocol in which participants commented on their actions and thoughts while interacting with the system, a log file with the extracted features from mass segmentation and relevance feedback, and a 7-point Likert scale questionnaire [32]. The think-aloud protocol allowed us to collect participants' auto-declarations regarding their user experience with mass segmentation and relevance feedback. The log-file allowed us to collect the number of relevant feedback required by participants (*i.e.*, the number of clicks on the CAD button panel) to obtain the final solution, and therefore to measure the performance of mass segmentation and relevance feedback [19]. We consider the number of clicks a quality measure for the evaluation of the system performance, since radiologists explicitly requested a system that required a minimum of clicks during the elicitation interviews. The 7-point Likert scale questionnaire allowed us to collect participants' subjective evaluation of the final segmentation (1="the final result is very bad", 7="the final result is very good") regarding both the accuracy of the segmented area and the accuracy of the contour regularity.

5.1.1 Number of Relevance Feedback

Overall, participants intensively tested the user interface in the first cases and gave a large amount of counter-feedback for these cases. The number of counter-feedback decreased after two or three cases. We explain this observation as follows: radiologists needed to understand the reactions of the system to their feedback, especially for weighted feedback on the mass size.

Regarding the mass size, the mean number of relevance feedback per image is "Reduce" (S+M+L: M=1.94), "Enhance" (S+M+L: M=1.91), number of

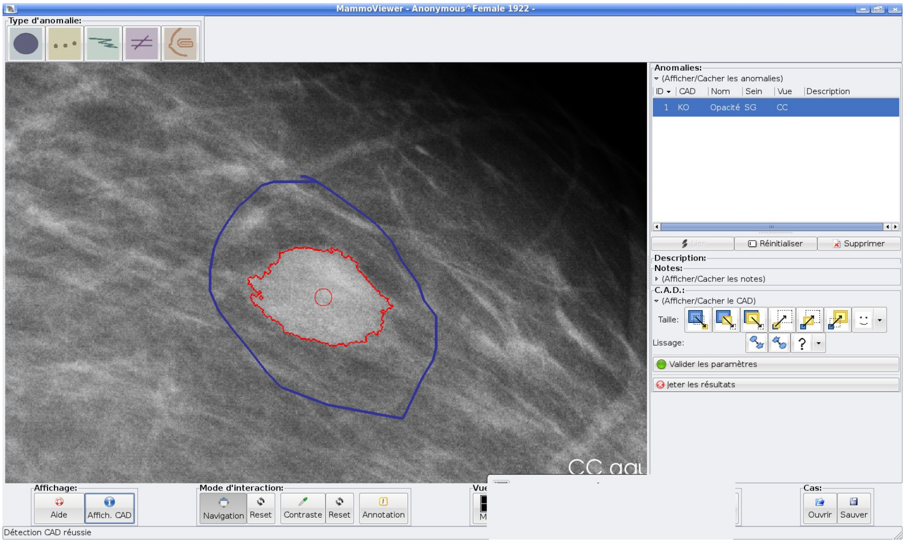


Fig. 11: Main screen: selection of a ROI (blue) and automatic mass segmentation (red). Right: CAD button-panel intended for relevance feedback

clicks (M=3.81), counter-feedback (M=0.61). Participants returned more feedback to enhance than to reduce the size of microlobulated and speculated masses. Moreover, they moderately improved and slightly reduced most of the mass area. Nevertheless, the optimal segmentation threshold could not be automatically reduced to increase these performances, as this would have increased the probability of an erroneous initial segmentation at the same time.

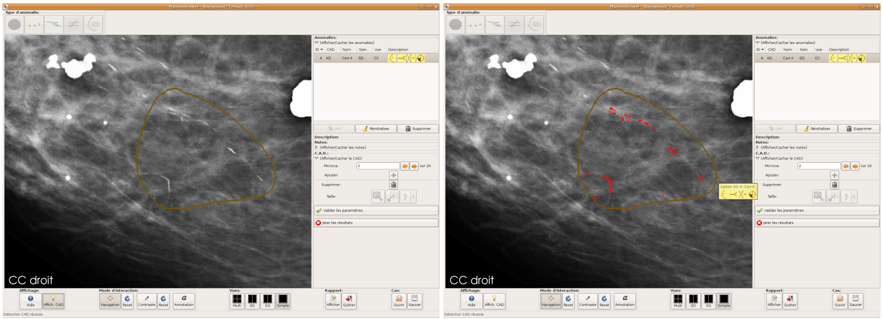
Regarding the smoothness of the contour, the mean number of relevance feedback per image is "Sharpen" (M=.52), "Smooth" (M=1.32), number of clicks (M=1.42), counter-feedback (M=0.1). Participants needed to regularise the contours at least once. To limit the number of clicks (and satisfy radiologists' requirement), contour regularization must be included in the initial segmentation.

5.1.2 Subjective Assessment

First, participants provided a large set of auto-declarations through think-aloud protocol. They reported that relevance feedback improved the accuracy of segmentation and that interactive segmentation provides a well-localised solution, even if the initial segmentation failed. According to one participant, offering the possibility of selecting and moving points in the contour would improve relevance feedback. In that case, the algorithm should calculate an optimal segmentation from the current segmentation and the points moved during interaction. Participants asked sharp contours to surround irregular findings and smooth contours to surround round or oval ones. Second, using the 7-point Likert scale questionnaire, participants subjectively assessed the

Measure	Value
Number of initially detected microcalcifications	367
Number of microcalcifications after radiologist validation (TP: true positives)	323
Number of added microcalcifications (FN: false negatives)	12
Number of deleted microcalcifications (FP: false positives)	41

Table 7: Evaluation of the detection of the microcalcifications.



(a) Before detection. (b) After detection.

Fig. 12: Selection of a region of interest for a microcalcification detection.

accuracy of the segmented area (M=5.88) and the accuracy of the contour regularity (M=5.68). Both scores are consistent with user satisfaction with the graphical design of the user interface, the relevance feedback components, and the reactions of the system to user commands.

Overall, the system was judged to be easy-to-use and participants felt they had the ability to improve the segmentation results with a simple request to a button.

6 Third Round of User Testing: Microcalcification Segmentation

We carried out this third round similarly to Round 2 to evaluate the usability of microcalcification detection and segmentation within the mammography viewer. We recorded in a log file the features extracted from the segmentation and relevance feedback according to the same protocol defined in Round 2. To evaluate the performance of the detection algorithm, we recruited three radiologists (3 male, who were different from the initial 9 participants involved in the first round to avoid any carryover effect) who annotated 20 clusters, *i.e.* 323 microcalcifications, and three false clusters (Fig. 12).

Table 7 reports the results regarding the detection of microcalcification: the false-negative rate is low (FN/TP=13/323=2.7%) for a reasonable false-positive rate (FP/TP=41/323=12.7%). The false positive markers correspond to visual artifacts that the system detects as findings. The missed findings are due to breast image density or mass densities. Therefore, the overall sensitivity is equal to TP/TP+FN=96.4%. We asked participants to draw a ROI with no findings inside and to annotate it as a cluster of microcalcifications even no microcalcifications exist. Regardless of the erroneous annotation, the detection found the true result with zero microcalcification inside the ROI (true negative). The participants returned an average number of 2.94 user feedback per cluster (Table 7), which corresponds to the average number of "Add" feedback (0.66) plus the average number of "Delete" feedback (2.27). This average is fair and reasonable, depending on the detection rate and image quality. Some clusters that include many visual artifacts require more user feedback. Then, the microcalcification suppression mode is not appropriate for this kind of image. Consequently, participants asked for the opportunity to remove several microcalcifications at once. During the thinkin-aloud protocol, all participants underlined the importance of computing the number of microcalcifications in a cluster and the importance of not missing any cluster. Furthermore, one participant stated that to miss a few microcalcifications is less significant than to miss a cluster. Fig. 13 reports the results of the IBM CSUQ questionnaire and their related metrics: SysUse (M=5.00, SD=1.29) and InterQual (M=5.22, SD=1.09) ae above the threshold of 5, while InfoQual (M=4.71, SD=1.38) and Satisfaction (M=4.67, SD=1.15) are below. Similarly to the first round of tests, we computed a Wilcoxon signed rank test for a single sample with one tail. Three metrics are significantly above the median value of 4 in a positive manner: SysUse (z-score=3.05, p=0.0011***) with a large effect size (Pearson's r=0.62), InfoQual (z-score=2.13, p<0.016*) with a medium effect size (Pearson's r=0.46), InterQual (z-score=2.26, p<0.0012***) with a large effect size (Pearson's r=0.75). Satisfaction (p=0.5, *n.s.*) is the only one not significant.

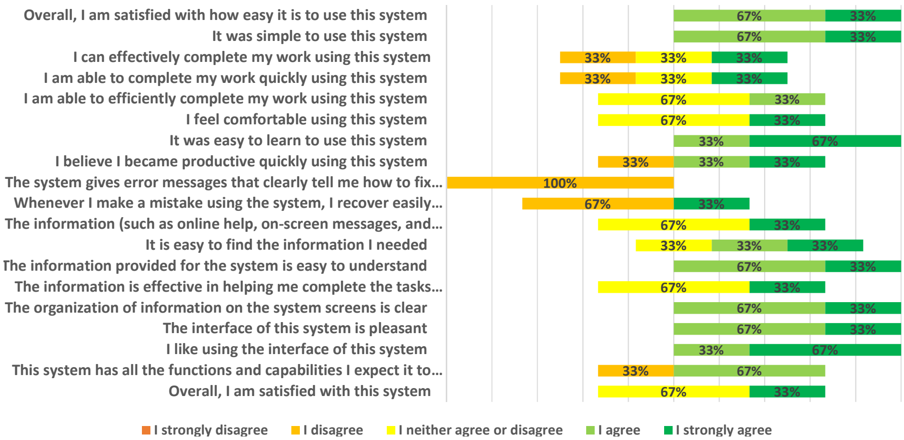


Fig. 13: Distribution of answers to IBM CSUQ for the third round.

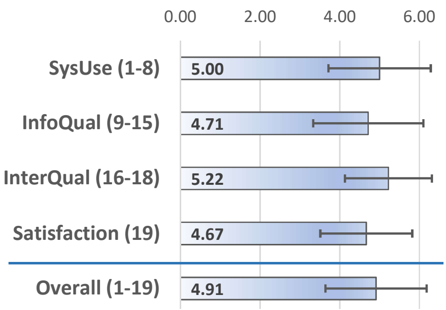


Fig. 14: IBM CSUQ metrics for the third round. Error bars show a confidence interval of 95% (α=0.05).

To compare Round 3 and Round 1 IBM CSUQ answers, we calculated a series of Mann-Whitney tests for two independent samples: SysUse is significantly lower in the third round than in the first round (z score=1.67, p=0.046*) with a small effect size (Pearson's r=0.17), but not InfoQual (p=0.063, *n.s.*) and Satisfaction (p=0.10, *n.s.*).

7 Conclusion and Future Work

This paper presents how we combined human-computer interaction and image processing methods to develop an integrated CDSS for BC annotation and reporting. HCI methods included elicitation of user requirements, design of multi-fidelity prototypes, and usability testing. Image processing methods include detection and segmentation algorithms. We report how we selected, customized and combined the methods borrowed from both fields to develop a CDSS for BC that satisfies radiologists' needs and expectations for multimodality, usability and flexibility.

We achieved multimodality by selecting stroke gesture interaction as input modality, which allows direct manipulation with the system and is natural for annotating regions of interest in medical images. We achieved usability by adopting a user-centered approach throughout the development life cycle, to identify radiologists' needs and expectations and test and refine the features of the proposed system. Specifically, we used domain knowledge elicitation interviews and scenario-based questionnaires with paper mock-ups to specify user requirements, and we carried out three rounds of user testing to evaluate the usability of the proposed system. Moreover, the user-centred approach allowed us to develop a star menu specially designed for the standardised and gesture-based annotation of BC findings that participants judged relevant, efficient and usable. We achieved flexibility by providing users with two modes for annotation: a manual annotation relying on the star menu and a semi-automatic annotation relying on both detection and segmentation algorithms and relevance feedback to improve segmentation results and better support human labor. Overall, participants expressed very positive reactions to the relevance feedback feature and perceived the system as easy-to-use and adapted to the human labour.

The significance of this work can be highlighted with regard to two perspectives: the users of the CDSS and the method adopted for its design and evaluation. First, we pay attention to users and usability throughout the software development life cycle. The main achievement is the improvement of users' experience with BC screening and reporting, as the resulting CDSS is not only useful but also usable. In particular, the pie and star menus led to an increased user performance compared to the array of icons. This is remarkable, as users had no prior experience with these unusual interaction styles. The major challenge to overcome was the lack of availability of domain-expert users at the beginning of the process, when we needed in-depth and extensive face-to-face meetings so as to collect enough data about their needs and expectations with the system.

Second, we combined prototyping tools and techniques such as user-centred design and model-based approaches [17]. In their work, Beaudouin-Lafon and Mackay [3] present these tools and techniques as essential to increase creativity, to explore a design space, and to help designers select the best solution. What is relevant and innovative in the method adopted here is, on the one hand, how the design process, the prototypes, and the usability evaluations were unified with the central goal to put the user at the center of the development process; and, on the other hand, how they provided specific and complementary information. What is promising is how this combination of selected HCI techniques led to the acceptability of the system by the target final users.

We evaluated neither detection accuracy nor prediction accuracy, as these are addressed elsewhere in the literature [49]. Instead, we focused on multimodality, usability, and flexibility as the first software quality criteria, which, to the best of our knowledge, were not explicitly addressed in related work.

Further development will combine detection and segmentation tools for interpretation and reporting tasks to have complete assistance with mammography screening. New image modalities, such as ultrasound or magnetic resonance imaging (MRI), are emerging, and the development can be extended to these new inputs. Another research avenue is the integration of vocal synthesis to the semi-automatic annotation, as vocal message can help support visual exploration in 2D displays [7, 8].

Acknowledgments. The authors would like to thank anonymous reviewers for their constructive comments on previous versions of this manuscript. We acknowledge the support of the VIGILE project (Visualisation Intégrée de Grandes Images médicaLEs) project by "Service Public de Wallonie" under Grant No. 616446 and of the DIAMANT project by "Innoviris".

Declarations. On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

[1] Amendoeira, I., Perry, N., Broeders, M., de Wolf, C., Törnberg, S., Holland, R., & von Karsa, L. European guidelines for quality assurance in breast cancer screening and diagnosis (pp. 1-160). European Commission. (2013), https://op.europa.eu/en/publication-detail/-/publication/4e74ee9b-df80-4c91-a5fb-85efb0fdda2b

[2] American College of Radiology. Breast imaging reporting and data system (BI-RADS®) atlas, 5th ed., Reston, VA, USA, (2007), https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads, https://www.acr.org/-/media/ACR/Files/RADS/BI-RADS/BIRADS-Reference-Card.pdf

[3] Beaudouin-Lafon, M. & Mackay, W. Prototyping Tools and Techniques. In *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies And Emerging Applications*. pp. 1006-1031 (2002), https://www.kth.se/social/upload/52ef5ee4f2765445a466a28a/mackay-lafon-prototypes-52-HCI.pdf

[4] Berg, W., D'Orsi, C., Jackson, V., Bassett, L., Beam, C., Lewis, R. & Crewson, P. Does Training in the Breast Imaging Reporting and Data System (BI-RADS) Improve Biopsy Recommendations or Feature Analysis Agreement with Experienced Breast Imagers at Mammography?. *Radiology*. **224**, 871 (2002), https://doi.org/10.1148/radiol.2243011626

[5] Bourguet, M.-L. Designing and Prototyping Multimodal Commands. *Proc. of 9th IFIP TC 13 International Conference, INTERACT '03*. IOS Press, Amsterdam, pp. 717-720 (2003), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.98.1958&rep=rep1&type=pdf

[6] Callahan, J., Hopkins, D., Weiser, M. & Shneiderman, B. An Empirical Comparison of Pie vs. Linear Menus. *Proc. of The ACM Conf. on Human Factors in Computing Systems CHI '88*. pp. 95-100 (1988), https://doi.org/10.1145/57167.57182

[7] Carbonnell, N. & Kieffer, S. Do oral messages help visual exploration. *Proc. of The International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems*. pp. 27-36 (2002), https://hal.inria.fr/inria-00107591

[8] Carbonnell, N. & Kieffer, S. Do oral messages help visual search? *Advances in Natural Multimodal Dialogue Systems. Text, Speech and Language Technology*, van Kuppevelt, J.C.J., Dybkjær, L., Bernsen, N.O. (eds), vol. 30, pp. 131-157 (2005). Springer, https://doi.org/10.1007/1-4020-3933-6_7

[9] Cheng, H., Cai, X., Chen, X., Hu, L. & Lou, X. Computer-aided detection and classification of microcalcifications in mammograms: a survey. *Pattern Recognition.* **36**, 2967-2991 (2003), http://www.sciencedirect.com/science/article/pii/S0031320303001924

[10] Costabile, M., Fogli, D., Fresta, G., Mussio, P. & Piccinno, A. Computer Environments for Improving End-User Accessibility. *Universal Access Theoretical Perspectives, Practice, and Experience.* 129-140 (2003), https://doi.org/10.1007/3-540-36572-9_10

[11] Coyette, A., Kieffer, S. & Vanderdonckt, J. Multi-fidelity Prototyping of User Interfaces. *Proc. of 11th IFIP TC 13 International Conference, INTERACT '07.* LNCS, volume 4662, Springer, pp. 150-164 (2007), https://doi.org/10.1007/978-3-540-74796-3_16

[12] Coyette, A., Schimke, S., Vanderdonckt, J., & Vielhauer, C. Trainable Sketch Recognizer for Graphical User Interface Design. *Proc. of 11th IFIP TC 13 International Conference, INTERACT '07.* LNCS, volume 4662, Springer, pp. 124-135 (2007), https://doi.org/10.1007/978-3-540-74796-3_14

[13] Dengler, J., Behrens, S. & Desaga, J. Segmentation of Microcalcifications in Mammograms. *IEEE Trans. Med. Imaging*, **12**, *4* pp. 634-642. https://pubmed.ncbi.nlm.nih.gov/18218457/

[14] DeSantis, C.E., Bray, F., Ferlay, J., Lortet-Tieulent, J., Anderson, B.O. & Jemal, A. Variation in Female Breast Cancer Incidence and Mortality Rates. *Cancer Epidemiol Biomarkers Prev.*, **24**, *10* pp. 1495-1506. https://pubmed.ncbi.nlm.nih.gov/26359465/

[15] Esmaeili, M., Ayyoubzadeh, S.M., Ahmadinejad, N., Ghazisaeedi, M., Nahvijou, A., & Maghooli, K. A decision support system for mammography reports interpretation. *Health Inf Sci Syst.*, **8**, *1* pp. 1495-1506. https://pubmed.ncbi.nlm.nih.gov/32257128/

[16] Fadil, R., Jackson, A., Majd, B., Ghazi, H. & Kaabouch, N. Segmentation of Microcalcifications in Mammograms: A comparative Study. *Proc. of IEEE Int. Conf. on Electro Information Technology, EIT '19.* pp. 18-23 (2019), https://doi.org/10.1109/EIT.2019.8833735

[17] Furtado, E., Furtado, V., Silva, W.B., Rodrigues, D.W.T., da Silva Taddeo, L., Limbourg, Q., Vanderdonckt, J. An ontology-based method for designing multiple user interfaces. *Proceedings of International Workshop on Multiple User Interfaces, MUI' 01* (2001). https://www.researchgate.net/publication/2567741_An_Ontology-Based_Method_for_Universal_Design_

[18] Gemo, M., Gouze, A., Debande, B., M.D., A., Mazy, G. & Macq, B. A versatile knowledge-based clinical imaging annotation system for breast cancer screening. *Proc. of Medical Imaging 2007.* **6514** pp. 580-588 (2007), https://doi.org/10.1117/12.709522

[19] Gouze, A., Kieffer, S., Brussel, C., Moncarey, R., Grivegnée, A. & Macq, B. Interactive breast cancer segmentation based on relevance feedback: from user-centered design to evaluation. *Proc. of Medical Imaging 2009: Computer-Aided Diagnosis.* **7260** pp. 593-602 (2009), https://doi.org/10.1117/12.813538

[20] Hulley, S.B., Cummings, S.R., Browner, W.S., Grady, D., & Newman, T.B. Designing clinical research: an epidemiologic approach, 4th ed. Wolters Kluwer-Lippincott Williams & Wilkins, Philadelphia, PA (2013). ISBN 978-1-60831-804-9, https://www.amazon.com/Designing-Clinical-Research-Stephen-Hulley/dp/1608318044

[21] ISO 9241-11:2018 – Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts. ISO/TC 159/SC 4 Ergonomics of human-system interaction, International Standard Organization (2018), https://www.iso.org/standard/63500.html

[22] ISO/IEC 25010:2019 – Software Quality Product Standard, International Standard Organization (2019), https://iso25000.com/index.php/en/iso-25000-standards/iso-25010?limit=3&limitstart=0

[23] Khoulqi, I. & Idrissi, N. Breast Cancer Image Segmentation and Classification. *Proceedings of The 4th International Conference on Smart City Applications, SCA '19.* (2019), https://doi.org/10.1145/3368756.3369039

[24] Kieffer, S., Gouze, A., Moncarey, R., van Brussel, Ch., Wispelaere, J.-F., Kayser, F., & Macq, B. Towards Standardized Pen-Based Annotation of Breast Cancer Findings. *Proc. of 13th International Conference on Human-Computer Interaction, HCI International '09.* LNCS, Vol. 5613, Springer, pp. 524-533 (2009). 10.1007/978-3-642-02583-9_57

[25] Kieffer, S., Coyette, A. & Vanderdonckt, J. User Interface Design by Sketching: A Complexity Analysis of Widget Representations. *Proc. of ACM Int. Conf. on Engineering Interactive Computing Systems, EICS '10.* ACM Press, pp. 57-66 (2010), https://doi.org/10.1145/1822018.1822029

[26] Kieffer, S. ECOVAL: Ecological Validity of Cues and Representative Design in User Experience Evaluations. *AIS Transactions On Human-Computer Interaction.* **9**, 149-172 (2017), https://aisel.aisnet.org/thci/vol9/iss2/4

[27] Kieffer, S., Gouze, A. & Vanderdonckt, J. User-centred Development of a Clinical Decision-support System for Breast Cancer Diagnosis and Reporting based on Stroke Gestures. *Proc. of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - VISIGRAPP (2: HUCAPP). 2021.* https://doi.org/10.5220/0010258900600071

[28] Kupinski, M. & Giger, M. Automated seeded lesion segmentation on digital mammograms. *IEEE Trans. on Medical Imaging.* **17**, 510-517 (1998), https://doi.org/10.1109/42.730396

[29] Lewis J.R. IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction.* **7**, 57-78 (1995), https://doi.org/10.1080/10447319509526110

[30] Lewis, J.R. Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies. *International Journal of Human-Computer Interaction.* **14**, 463-488 (2002), https://doi.org/10.1080/10447318.2002.9669130

[31] Lewis, J.R. Sample Sizes for Usability Tests: Mostly Math, Not Magic. *Interactions.* **13**, 29-33 (2006), http://doi.acm.org/10.1145/1167948.1167973

[32] Likert, R. A technique for the measurement of attitudes. *Archives of Psychology.* **22**, 55- (1932), http://psycnet.apa.org/record/1933-01885-001

[33] Mazo, C., Kearns, C., Mooney, C., & Gallagher, W.M. Clinical Decision Support Systems in Breast Cancer: A Systematic Review. *Cancers* **12**, *2*, 369 (2020, Feb. 6). http://doi:10.3390/cancers12020369.

[34] Magrofuoco, N., Roselli, P. & Vanderdonckt, J. Two-Dimensional Stroke Gesture Recognition: A Survey. *ACM Computing Surveys.* **54** (2021), https://doi.org/10.1145/3465400

[35] Maguire, M. Methods to support human-centred design. *Int. J. Hum. Comput. Stud..* **55**, *4*, 587-634 (2001), https://doi.org/10.1006/ijhc.2001.0503

[36] Massafra, R., Latorre, A., Fanizzi, A., Bellotti, R., Didonna, V., Giotta, F., La Forgia, D., Nardone, A., Pastena, M., Ressa Cosmo, M., Rinaldi, L., Russo A., Tamborra, P., Tangaro, S., Zito, A., & Lorusso, V.. A Clinical Decision Support System for Predicting Invasive Breast Cancer Recurrence: Preliminary Results. *Frontiers in Oncology.* **11** (2021), https://doi.org/10.3389/fonc.2021.576007

[37] Mayhew, D. Usability Engineering Lifecycle: A Practitioner's Handbook for User Interface Design. Elsevier Ltd, Oxford (March 1999), https://www.elsevier.com/books/the-usability-engineering-lifecycle/mayhew/978-0-08-052028-5

[38] Montero, F., López-Jaquero, V., Vanderdonckt, J., González, P., Lozano, M. & Limbourg, Q. Solving the Mapping Problem in User Interface Design by Seamless Integration in IdealXML. *Proc. of Int. Workshop on Design, Specification, and Verification of Interactive Systems, DSV-IS'2006*. Springer, Berlin, pp. 161-172 (2006), https://doi.org/10.1007/11752707_14

[39] Moyle, M. & Cockburn, A. Analysing Mouse and Pen Flick Gestures. *Proc. Of The Symp. On Comp.-Human Int.*. pp. 19-24 (2002), https://doi.org/10.1145/2181216.2181220

[40] Nielsen, J. & Molich, R. Heuristic Evaluation of User Interfaces. *Proceedings of The ACM SIGCHI Conference on Human Factors In Computing Systems*. pp. 249-256 (1990), https://doi.org/10.1145/97243.97281

[41] Pawloski, P., Brooks, G., Nielsen, M. & Olson-Bullis, B. A Systematic Review of Clinical Decision Support Systems for Clinical Oncology Practice. *Journal Of The National Comprehensive Cancer Network*. **17**, 331-338 (2019), https://doi.org/10.6004/jnccn.2018.7104

[42] Rosson, M. & Carroll, J. Scenario-Based Design. *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies And Emerging Applications*. pp. 1032-1050 (2002), https://www.taylorfrancis.com/chapters/edit/10.1201/b11963-ch-48/scenario-based-design-mary-beth-rosson-john-carroll

[43] Ruthven, I. & Lalmas, M. A Survey on the Use of Relevance Feedback for Information Access Systems. *Knowl. Eng. Rev.* **18**, *6*, 95-145 (2003), https://doi.org/10.1017/S0269888903000638

[44] Sainfort & Booske, B. Human-Computer Interaction in Healthcare. (2017,5), https://www.crcpress.com/Human-Computer-Interaction-Designing-for-Diverse-Users-and-Domains/Sears-Jacko/p/book/9781138115057

[45] Saslow, D., Hannan, J., Osuch, J., Alciati, M., Baines, C., Barton, M., Bobo, J., Coleman, C., Dolan, M., Gaumer, G., Kopans, D., Kutner, S., Lane, D., Lawson, H., Meissner, H., Moorman, C., Pennypacker, H., Pierce, P., Sciandra, E., Smith, R. & Coates, R. Clinical Breast Examination: Practical Recommendations for Optimizing Performance and Reporting. *CA: A Cancer Journal For Clinicians*. **54**, 327-344 (2009), https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.

3322/canjclin.54.6.327

[46] Sefelin, R., Tscheligi, M. & Giller, V. Paper Prototyping - What is It Good for? A Comparison of Paper- and Computer-Based Low-Fidelity Prototyping. *Proc. of ACM Conf. on Human Factors In Computing Systems, CHI '03, Extended Abstracts*. pp. 778-779 (2003), https://doi.org/10.1145/765891.765986

[47] Simonin, J., Kieffer, S. & Carbonell, N. Effects of Display Layout on Gaze Activity During Visual Search. *Proc. of IFIP TC13 Int. Conf. on Human-Computer Interaction, Interact '05*. LNCS, Vol. 3585, Springer, pp. 1054-1057 (2005) https://doi.org/10.1007/11555261_103

[48] Skaane, P. & Skjennald, A. Screen-Film Mammography versus Full-Field Digital Mammography with Soft-Copy Reading: Randomized Trial in a Population-based Screening Program—The Oslo II Study. *Radiology*. **232**, 197-204 (2004), https://doi.org/10.1148/radiol.2321031624

[49] Sluÿters, A., Lambot, S., & Vanderdonckt, J. Hand Gesture Recognition for an Off-the-Shelf Radar by Electromagnetic Modeling and Inversion. *Proc. of 27th ACM Int. Conf. on Intelligent User Interfaces, IUI '22*. ACM, New York, NY, USA, pp. 506-522 (2022), https://doi.org/10.1145/3490099.3511107

[50] Smith AP, M. Emerging technologies in breast cancer detection. *Radiology Management*. **26**, 16-24 (2004), https://www.ncbi.nlm.nih.gov/pubmed/30726647

[51] Sung, H. Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians **77** *3* (2021), https://doi.org/10.3322/caac.21660

[52] Vanderdonckt, J. Accessing Guidelines Information with Sierra. *Proc. of IFIP TC13 Int. Conf. on Human—Computer Interaction, Interact '95*. pp. 311-316 (1995), https://doi.org/10.1007/978-1-5041-2896-4_52

[53] Vanderdonckt, J., Roselli, P. & Pérez-Medina, J. !FTL, an Articulation-Invariant Stroke Gesture Recognizer with Controllable Position, Scale, and Rotation Invariances. *Proc. of The 20th ACM International Conference on Multimodal Interaction, ICMI '18*. pp. 125-134 (2018), https://doi.org/10.1145/3242969.3243032

[54] Vinnicombe, S., Pinto Pereira, S., McCormack, V., Shiel, S., Perry, N. & Santos Silva, I. Full-Field Digital versus Screen-Film Mammography: Comparison within the UK Breast Screening Program and Systematic

Review of Published Data. *Radiology*. **251**, 347-358 (2009), https://doi.org/10.1148/radiol.25120812350A202020200A

[55] Wang, D. Scenario-focused development method for a pen-based user interface: model and applications. *The Journal Of Supercomputing*. **46** 139-149 (2008), https://doi.org/10.1007/s11227-007-0161-1

[56] Wittenberg, T., Elter, M. & Schulz-Wendtland, R. Complete Digital Iconic and Textual Annotation for Mammography. *Bildverarbeitung Für Die Medizin 2007*. 91-95 (2007), https://doi.org/10.1007/978-3-540-71091-2_19

[57] Zheng, Y., Wu, M., Cole, E. & Pisano, E. Online annotation tool for digital mammography. *Academic Radiology*. **11**, 566-572 (2004), http://www.sciencedirect.com/science/article/pii/S1076633203007268

References

Appendix: Catalogue of Icons.

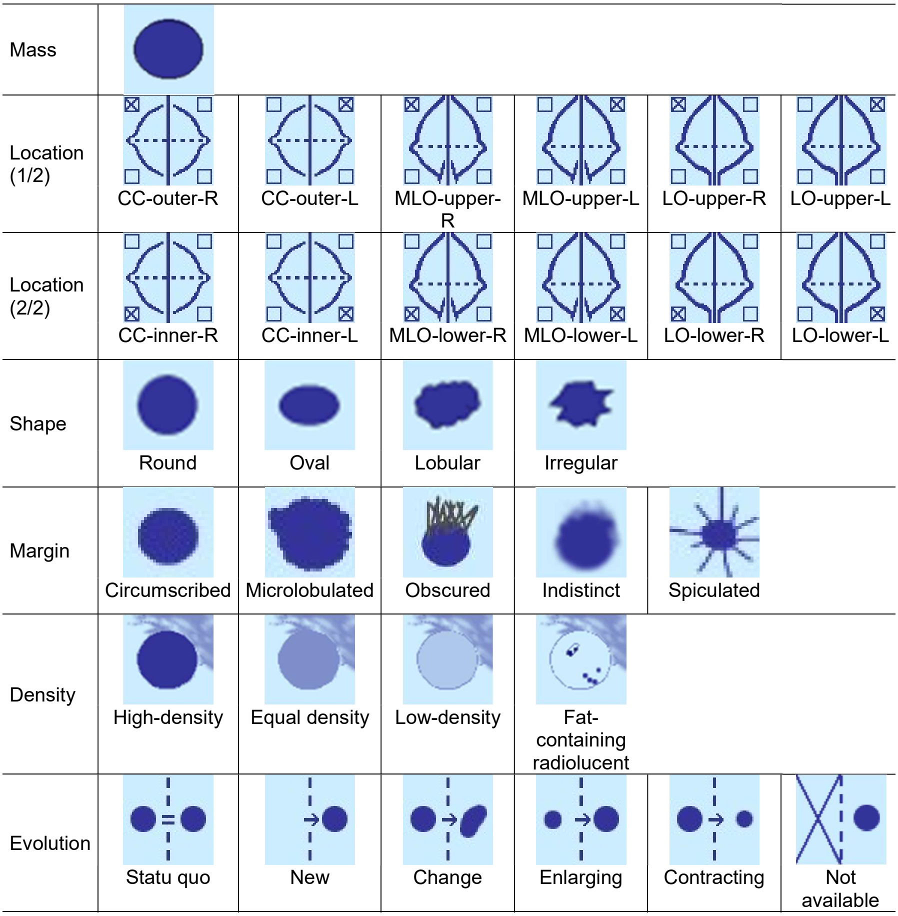


Fig. 15: Mass icons: mediolateral oblique (MLO); craniocaudal (CC); right (R); left (L).

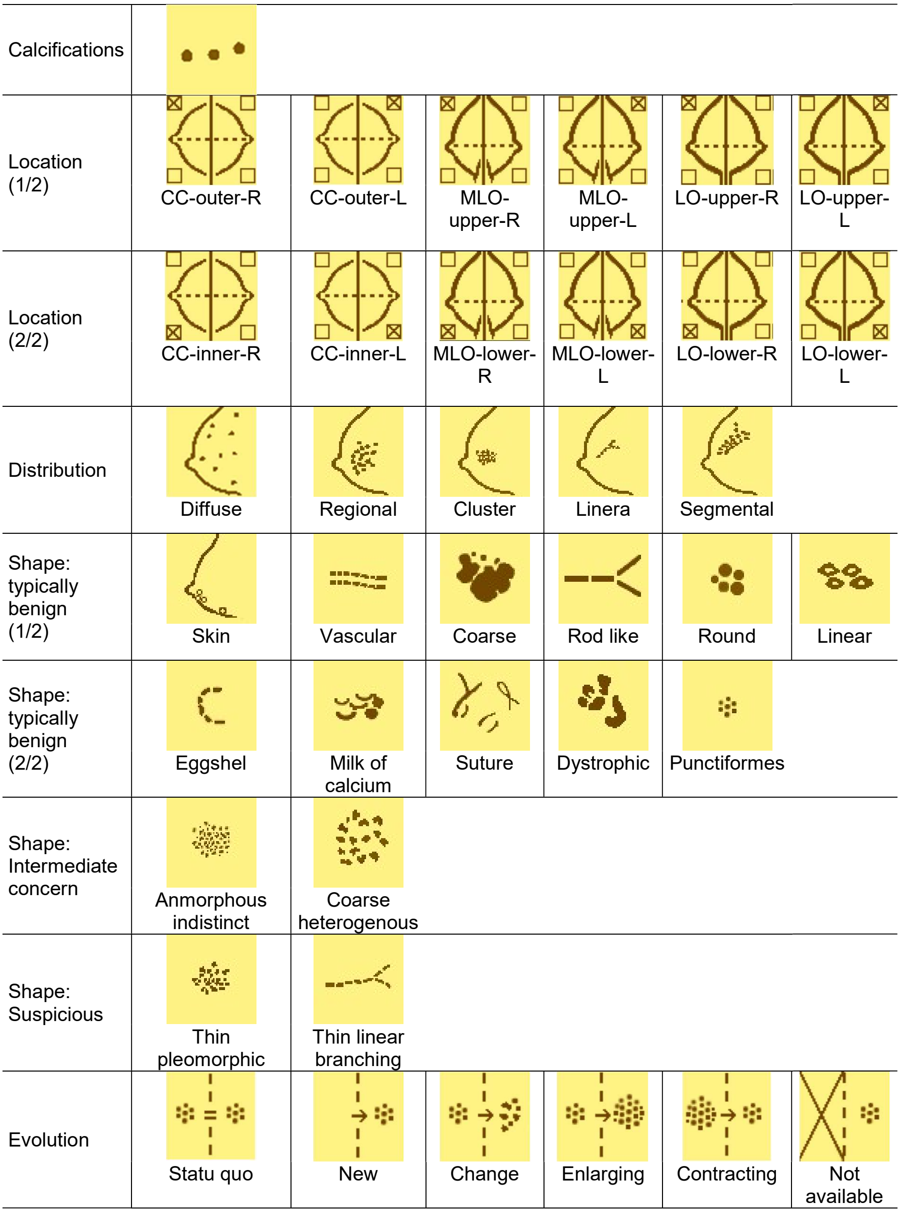


Fig. 16: Calcifications (ca++) icons: mediolateral oblique (MLO); craniocaudal (CC); right (R); left (L)

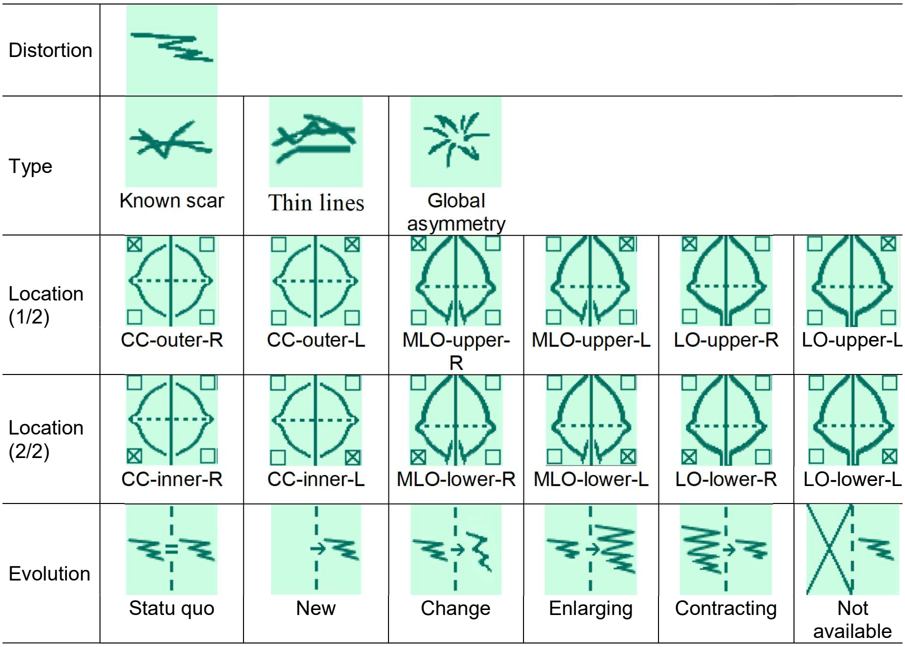


Fig. 17: Architectural distortion icons: mediolateral oblique (MLO); cranio-caudal (CC); right (R); left (L).

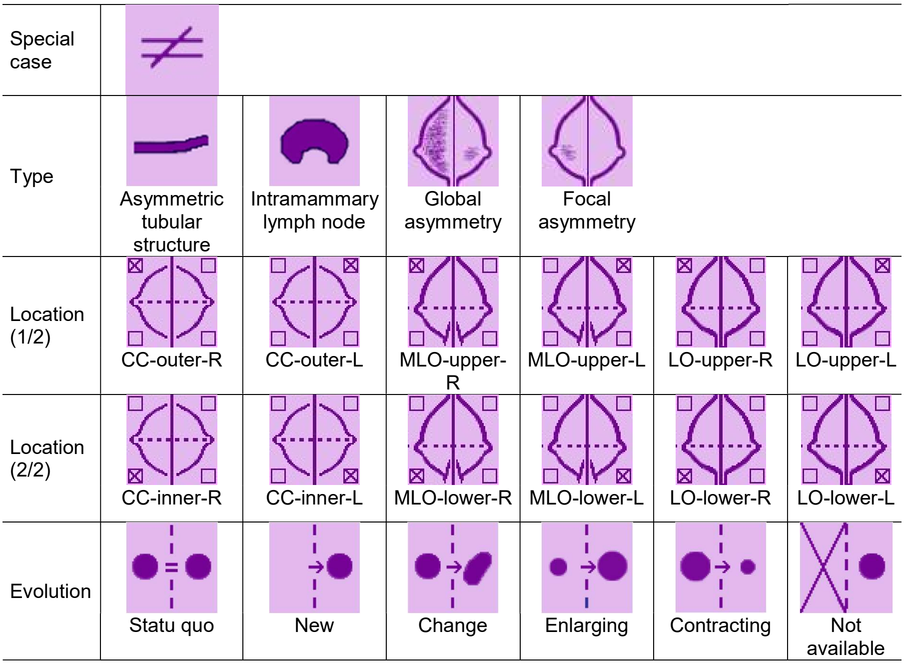


Fig. 18: Special case icons: mediolateral oblique (MLO); craniocaudal (CC); right (R); left (L).

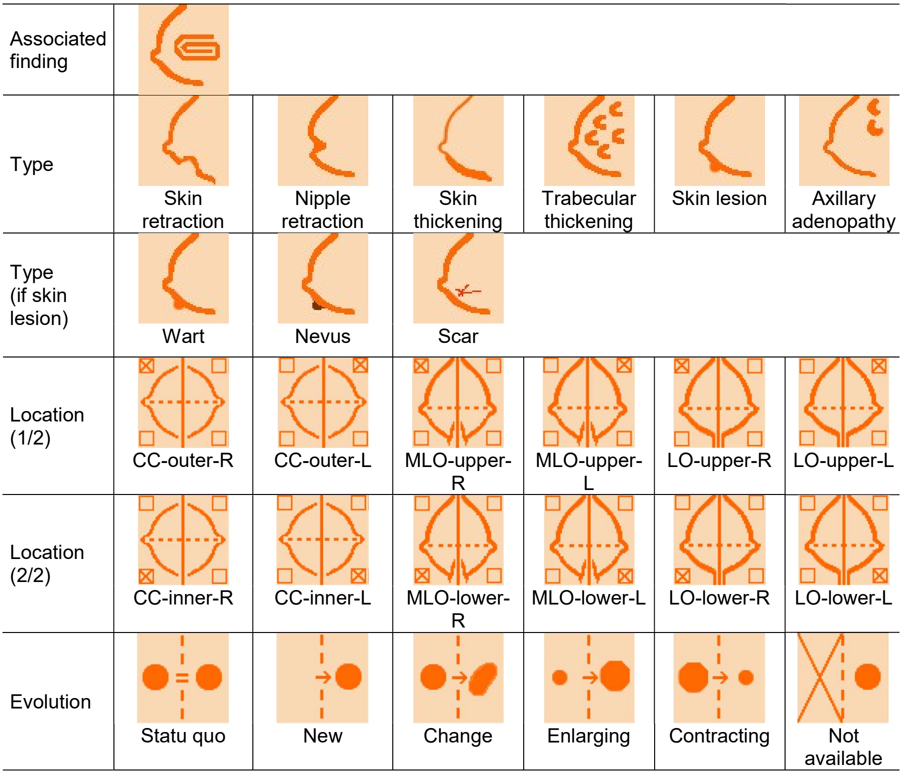


Fig. 19: Associated finding icons: mediolateral oblique (MLO); craniocaudal (CC); right (R); left (L).