HHS Public Access

Author manuscript

*Mach Learn Appl*. Author manuscript; available in PMC 2022 December 27.

Published in final edited form as:

*Mach Learn Appl*. 2022 December 15; 10: . doi:10.1016/j.mlwa.2022.100430.

# A fully automatic framework for evaluating cosmetic results of breast conserving therapy

**Chenqi Guo**[a,*], **Tamara L. Smith**[b], **Qianli Feng**[a], **Fabian Benitez-Quiroz**[a], **Frank Vicini**[c], **Douglas Arthur**[d], **Julia White**[e], **Aleix Martinez**[a]

[a]Computational Biology and Cognitive Science Laboratory, the Ohio State University, Columbus, OH, USA

[b]Radiation Oncology, Memorial Healthcare System, Hollywood, FL, USA

[c]Radiation Oncology, Genesis Care Pty Ltd, Alexandria, NSW, Australia

[d]Radiation Oncology, Virginia Commonwealth University, Richmond, VA, USA

[e]Radiation Oncology, the Ohio State University, Columbus, OH, USA

## Abstract

The breast cosmetic outcome after breast conserving therapy is essential for evaluating breast treatment and determining patient's remedy selection. This prompts the need of objective and efficient methods for breast cosmesis evaluations. However, current evaluation methods rely on ratings from a small group of physicians or semi-automated pipelines, making the processes time-consuming and their results inconsistent. To solve the problem, in this study, we proposed: 1. a *fully-automatic* Machine Learning Breast Cosmetic evaluation algorithm leveraging the state-of-the-art Deep Learning algorithms for breast detection and contour annotation, 2. a novel set of Breast Cosmesis features, 3. a new Breast Cosmetic dataset consisting 3k+ images from three clinical trials with human annotations on both breast components and their cosmesis scores. We show our fully-automatic framework can achieve comparable performance to state-of-the-art without the need of human inputs, leading to a more objective, low-cost and scalable solution for breast cosmetic evaluation in breast cancer treatment.



*Correspondence to: Computational Biology and Cognitive Science Laboratory, the Ohio State University, Columbus, OH 43210, USA. guo.1648@buckeyemail.osu.edu (C. Guo).

CRediT authorship contribution statement

**Chenqi Guo:** Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization, Supervision, Projection administration. **Tamara L. Smith:** Methodology, Investigation, Resources, Data curation, Writing – review & editing, Visualization. **Qianli Feng:** Methodology, Writing – review & editing, Visualization. **Fabian Benitez-Quiroz:** Methodology, Writing – review & editing. **Frank Vicini:** Conceptualization, Investigation, Resources, Data curation, Funding acquisition. **Douglas Arthur:** Conceptualization, Investigation, Resources, Data curation, Funding acquisition. **Julia White:** Conceptualization, Investigation, Resources, Data curation, Funding acquisition. **Aleix Martinez:** Methodology, Supervision.

Ethical statement

The digital photos collected to assess breast cosmetic outcomes are part of the Quality of Life (QOL) outcome analysis on three national breast cancer clinical trials, all of which were approved by each participating institutions' Institutional Review Board (IRB). The participation in each trial is optional for patients. Each trial has specific instructions for digital photo submission: face and personal or protected information are not included (for example, patient name, medical record number, date of birth, etc.).

**Keywords**

Breast cancer; Breast conserving therapy; Breast Cosmesis scores; Breast detection; Machine learning; Predictive model

## 1. Introduction

Breast conservation for early stage breast cancer is one of the notable achievements of modern cancer care. Breast cancer is the most commonly diagnosed cancer in women. Just in the USA over 280,000 new cases were diagnosed in the year of 2021 (Siegel, Miller, Fuchs, & Jemal, 2021). The widespread adoption of mammography screening has led to earlier detection of smaller, early stage breast cancers that in combination with improved treatments over the past decade has resulted in better cancer outcomes and a growing number of long-term survivors (Desantis et al., 2017). Over 50% of newly diagnosed breast cancers each year are Stage 0 or 1 and most of these patients will elect to undergo Breast Conserving Therapy (BCT) as opposed to mastectomy for treatment of their breast cancer (Sareigo, 2008; Tuttle et al., 2012). BCT includes surgical removal of the cancer from the breast with lumpectomy followed by a course of radiation therapy to the retained breast. Cancer control outcomes from BCT have been extensively studied in clinical trials, meta-analyses and modern registries over the last three decades (Early Breast Cancer Trialists' Collaborative Group (EBCTCG), 2011; van Maaren et al., 2016), demonstrating that breast cancer mortality, overall survival as well as local regional recurrence risks are comparable to that of mastectomy for disease. Given the expectation of comparable cancer outcomes from different local treatment approaches, the appearance of the breast or cosmetic outcome has become an important determinant for treatment selection as a quality of life-related end point (Volders et al., 2017). Fear of a poor cosmetic outcome from BCT may also lead women to select riskier and more costly treatment (e.g., mastectomy with reconstruction for local regional treatment), causing additional psychological pressure for patients (Pataky & Baliski, 2016; Razdan et al., 2016). Therefore, acquiring relatively objective Breast Cosmesis scores is of great importance.

The Breast Cosmesis scores (Harris, Levene, Svensson, & Hellman, 1979) are 4-scale ordinals to evaluate the cosmetic appearance of breasts, concerning skin color, breast size, breast shape, nipple appearance, scar appearance, and a global score (an overall impression of the appearance of the treated breast in comparison to the untreated breast). Although the scores grading criterion is defined in details, physicians may still inherently grade images differently based on their personal views, experience, or grading scales. For example, one physician may grade an image as a "Good" Breast Cosmetic outcome, while another physician rates the same image as "Fair". Besides the inconsistency, evaluating Breast Cosmesis scores is also time consuming, and large clinical trials typically acquire thousands of images. Therefore, a more efficient and unbiased grading method is urgently needed.

In this paper, we propose a machine-learning algorithm which utilizes digital photographs of the breasts as the input and outputs cosmesis score objectively and efficiently. Previous attempts like BCCT.core (Brouwers, Werkhoven, Bartelink, & Fourquet, 2016; Cardoso &

Cardoso, 2007; Cardoso, da Costa, & Cardoso, 2005) (an intelligent medical software for evaluating breast cosmetic outcomes after breast cancer conservative treatment) requires manually annotation on the starting and ending points of a breast. Instead, here we propose a fully-automatic method without any human supervision with a novel set of Breast Cosmesis features. More specifically, given a digital photo of a patient with frontal view of two breasts, our algorithm is able to: 1. automatically detect breasts, areolas and nipples, 2. automatically extract landmarks of each breast components, 3. automatically extract Breast Cosmetic features, 4. automatically predict Breast Cosmesis scores for breast skin color, breast shape, breast size, scar appearance, nipple appearance and global cosmetic outcome, see Fig. 1. Experiments show that our fully-automatic algorithm can achieve comparable performance to the BCCT.core algorithm without the need for human input.[1]

The remainder of this paper is arranged as follows: Section 5.1 introduces how the Breast Cosmesis images, scores and other data are collected and processed. Section 5.2 describes how the breast compontents detection and landmark detection is designed and trained. Section 5.3 illustrates how the Breast Cosmesis features are constructed. Section 5.4 presents the machine learning algorithms we used for feature selection and Breast Cosmesis scores prediction. Section 6 measures the performance of the proposed system per components and as a whole. The conclusion and discussion of using machine learning algorithms to evaluate Breast Cosmesis scores are given in Section 7.

## 2. Related work

For evaluating the breast cosmetic outcome after conservative therapy, Harris scoring (Harris et al., 1979) with 4-scale (excellent, good, fair, and poor) introduced in 1979 is the most widely used subjective approach, both for physicians' grading and patient-reported self-scoring. Other score criteria can be either 3-scale (good, fair, and poor) used for non-expert breast cosmesis graders, or 2-scale (good versus bad) in a binary evaluation system (Fitzal et al., 2007). In this research, we adopt the 4-scale Harris ordinal grading criterion.

Christie, Sharpley, and Curtis (2005) proposed a subjective evaluation system using digital photos on a computer which enables observers quantitatively comparing breast retraction on the treated side with the untreated side. This system requires the photographed patient holding a perspex scaling device as reference for scale calibration in later assessment process. A white nipple markers (WNM) adhesive on the patient is also used at the time of photography to help reducing numbers of observations required to minimize cosmesis evaluation error. This system provides insights on how to design a breast cosmesis subjective grading system. Yet, this subjective evaluation system requires careful scale calibration for photo standardization and burdensome nipple markers for observation aiding, while our proposed framework is fully-automatic without any adjustment requirement during the photographing procedure.

As a time-efficient and result-consistent approach compared with subjective evaluation, several attempts have been made for breast cosmesis objective assessment and the

[1]Codes, models and data are available at https://codeocean.com/capsule/2887058/tree.

corresponding computer imaging techniques have achieved promising advances. Based on the breast symmetry index (BSI) computed by subtracting the size and the shape between two breasts from both frontal and side view photos, a software system called Breast Analyzing Tool (BAT) (Fitzal et al., 2007) was created and able to differentiate between a binary (good versus bad) cosmesis score. However, this method requires additional inputs like manual annotations on breast markers and only takes breast shape and size into account. The objective evaluation is limited to good and bad cosmesis instead of the 4-scale Harris grading criterion, which hinders more detail exploration. In contrast, our method is fully-automatic without any need of human supervision, also considers breast skin color and nipple appearance factors, and predicts on 4-scale Harris scores.

Another well established objective breast cosmesis intelligent evaluation software using digital photo is BCCT.core (Brouwers et al., 2016; Cardoso & Cardoso, 2007; Cardoso et al., 2005). As a semi-automated pipeline which also requires manual annotations on breast, it measures breast asymmetry, skin color difference and scar visibility and predicts on a 4-point Harris scoring system with Machine Learning algorithm. Compared with BAT, BCCT.core achieves superior results especially when analyzing high-quality digital photographs (Cardoso, Cardoso, Wild, Krois, & Fitzal, 2009). In this paper, we mainly focus on and compare our algorithms and results with this BCCT.core framework. Compared with BCCT.core, our fully-automatic framework takes more breast features into consideration and achieves better classification accuracy without the need of any additional manual annotations.

Inspired by the previous work, Soror et al. (2016) proposed Objective Breast Cosmesis Score (OBCS) for breast conservative therapy by measuring certain anatomic distances without scale calibration using breast frontal images. These measurements represent the nipple displacement and the asymmetry in breast dimensions and contour. Experiments showed that the proposed OBCS is eligible for objective cosmesis assessment without scale calibration. Yet, this method does not take breast skin color and scars into account and still needs human annotations like the nipple and breast locations as extra inputs, which are addressed by our framework.

Recently, Deep Learning (DL) based methods have also been studied on biomedical breast detection and classification. To assist radiologists and healthcare professionals in the breast cancer diagnosis, Ragab, Albukhari, Alyami, and Mansour (2022) developed an Ensemble Deep-Learning-Enabled Clinical Decision Support System for the detection and classification of breast cancer using ultrasound images. In this system, a multi-level thresholding-based image segmentation technique was designed to identify the tumor-affected regions, together with an ensemble of three deep learning models for feature extraction and a machine learning classifier for breast cancer detection. Althobaiti et al. (2022) introduced a social engineering optimization with deep transfer learning-based breast cancer detection and classification (SEODTL-BDC) model using photoacoustic multimodal imaging. In this model, a lightweight LEDNet is employed for biomedical images segmentation, a residual network for feature extraction, and an SEO-RNN classifier for biomedical images classification.

The above DL models are focusing on breast cancer diagnosis on tumor detection and classification with medical imaging like ultrasound or photoacoustic, while our study is dealing with breast, areola and nipple detection and cosmetic outcome evaluations using RGB photographs directly taken by a digital camera. To our knowledge, our method is the first *fully-automatic* breast cosmetic evaluation framework achieving state-of-the-art performance without any human supervision.

## 3. Project objectives

Specifically, the objectives in this breast cosmesis project contain the following:

1. Standardize the breast digital photos from 3 clinical trials by color adjusting and image cropping.
2. Grade the standardized images by a group of physicians to subjectively evaluate the cosmesis results using 4-scale Harris scoring criterion.
3. Annotate the landmarks manually for breasts, areolas, nipples, surgical scars, and areas of radiation changes in images by a radiation oncology physician.
4. Create a fully-automatic algorithm that can objectively generate the cosmetic score of a patient's breast image after conservative therapy for early-stage breast cancer. Technically, input breast digital photographs, and output landmarks of each breast component and 4-scale Harris breast cosmesis scores.

## 4. Pipeline algorithm overview

As the flowchart of our fully-automatic Breast Cosmesis evaluation system shown in Fig. 1, here we provide an overall algorithm for this proposed method in Algorithm 1. The time complexity for each step in this algorithm is linear, since it is a one-image-in, one-result-out algorithm without any iterations or recursions.

**Algorithm 1:**

Overall algorithm for our fully-automatic Breast Cosmetic evaluation pipeline

**Input:** A patient's frontal-view breast image in RGB $\mathbf{X}_o \in \mathbb{R}^{3 \times w \times h}$

**Data:** Breast component $c \in$ {lb, rb, la, ra, ln, rn}

**Result:** Objective breast cosmesis scores prediction $\mathbf{y} \in \mathbb{R}^6$

1 Pre-processing $\mathbf{X}_o$ by color adjusting and image cropping to obtain image $\mathbf{X}$

2 Breast, areola and nipple detection: by YOLO detectors $D_c(\mathbf{X})$ to get bounding boxes $\mathbf{b}_c \in \mathbb{R}^4$

3 Breast, areola and nipple landmark detection: by Ensemble Regression Trees detectors $G_c(\text{crop}(\mathbf{X}, \mathbf{b}_c))$ to get landmarks $\mathbf{S}_c \in \mathbb{R}^{2 \times d_c}$

4 Feature extraction: by Python programs to get $p = 298$ dimensional breast cosmesis feature vector $\mathbf{v}_f$

5 Machine Learning prediction: by Lasso Regression, kSVC and OrdinalNet to get $\mathbf{y} = F_{ML}(\mathbf{v}_f)$

## 5. Material and methods

This section is organized as follows:

Section 5.1 first describes the data collection methods used in this study, including the breast image dataset, cosmesis grading scores and landmark annotations. Next, the detection of the breast components and landmarks are introduced in Section 5.2. Section 5.3 then illustrates the breast feature extractions of our framework. Finally, details of our Machine Learning algorithms are described in Section 5.4.

### 5.1. Data collection

we first describe our data collection method, as data itself plays an important role in any data-driven method. The method used for data collection and the quality of the acquired data is directly related to the performance of the algorithm.

**5.1.1. Image data collection, processing and grading—**All images are collected from three clinical trials (NRG NSABP B39/RTOG 0413, NRG RTOG 1005 and NRG RTOG 1014 with https://clinicaltrials.gov/ registration number NCT00103181, NCT01349322 and NCT01082211). All patients signed informed consent forms to participate and to permit research study on their breast images. The protocol for taking the digital photos consisted of plain background, high image quality and patients naked in akimbo position. For image pre-processing, Low-quality images (i.e., too noisy or low resolution) were excluded. Color standardization and cropping are then also applied as pre-processing (Sekhon et al., 2017). Table 1 describes the number of images from each clinical trial before and after selection. Only the frontal view of the breasts is used in this study.

After pre-processing, all image were graded by a group of physicians to evaluate the cosmesis result from the breast cancer treatment. There are 6 reviewers in the group, all of whom are Breast Cancer specific Radiation Oncologists with a range of experience from around 6 years to 25 years. All of them are clinical investigators that have run clinical trials including an assessment of breast cosmetic outcomes using a validated tool called Global Cosmetic Score (GCS). The scores contain a global cosmetic score and five different criteria (sub-scores), including breast size, breast shape, nipple appearance, skin color and scar appearance. Each of these five criteria is graded with 4 points, 0 = no or minimal difference between the treated and untreated breast, 1 = slight or small difference between the treated and untreated breast, 2 = more marked difference between breasts, 3 = disturbing difference between breasts. The global cosmesis score takes into account the scores from the 5 criteria and the overall appearance of the patient's breasts. The global cosmesis score is also graded in 4 points, 1 = Excellent, 2 = Good, 3 = Fair, and 4 = Poor. As the score distributions shown in Fig. 2, this dataset is unbalanced, suggesting difficulties in down-stream Machine Learning predictions.

**5.1.2. Annotation protocol—**To achieve fully automated estimation of the global and categorical scores, the algorithm needs to first detect the anatomical structures of the breast in the image to avoid the information in background and other areas of the patient body

being extracted. Landmarks estimation is then performed for each part of the breast to extract the precise area and shape. We use machine learning algorithms for both tasks and it is thus essential to acquire high-quality manual annotation on the landmarks. In our study, we are interested in anatomical annotation on 6 breast components: left breast, right breast, left areola, right areola, left nipple, and right nipple.

More specifically, a radiation oncology physician manually annotated the landmarks for breasts, areolas, nipples, surgical scars, and areas of radiation changes. The starting landmark for a single breast is at the armpit, and the breast contour is outlined until the medial part of the breast or where the contour of bottom breast disappears. Due to a lack of anatomical definition of landmarks other than the start and end points, we created pseudo-landmarks by resampling the physician's contour annotation with equal-arc sampling and 30 pseudo-landmarks per each breast. For the areolas and nipples, since both contours are closed, we choose the top-most landmark with smallest $y$ coordinates as the first (i.e., starting) landmark, and then arranged the remaining landmarks counter-clockwise along the curve. Equalarc length sampling is also used to generate the pseudo-landmarks for areola (12 points) and nipple (6 points). An inspector is asked to check the annotation to assure the quality. Example of the annotation is shown in Fig. 3.

## 5.2. Detection and landmarks estimation

To evaluate breast cosmesis score fully automatically, for a given image, the algorithm needs to locate the area of each breast component. Section 5.2.1 describes the algorithm for breast, areola and nipple detection. Once the bounding boxes are estimated for each breast component, a landmark detection algorithm is used for estimating the shape of each component, which is described in Section 5.2.2. The image is then further processed for skin analysis and feature extraction in Section 5.2.3.

**5.2.1. Breast, areola and nipple detection—**For a given pre-processed patient image in RGB $\mathbf{X} \in \mathbb{R}^{3 \times w \times h}$, the bounding boxes for each component $\mathbf{b}_c \in \mathbb{R}^4$ is estimated by,

$$\mathbf{b}_c = D_c(\mathbf{X}) \tag{1}$$

for $c \in \{\text{lb, rb, la, ra, ln, rn}\}$. The components includes left breast (lb), right breast (rb), left areola (la), right areola (ra), left nipple (ln) and right nipple (rn). $D_c$ denote the detector for component $c$ and $\mathbf{b}_c$ denotes the 4-d bounding box vector of the corresponding component.

We use YOLO-v3 (Redmon, Divvala, Girshick, & Farhadi, 2016) trained on our dataset as our breast components detector due to its robust and accurate performance on generic object detection. Instead of training one detector for all the components we train individual detector for right breast, left breast, areola and nipple. A single YOLO-v3 detector is not preferred here since it tends to miss small objects (e.g., nipple or areola) when detects all parts at once. Modeling left and right breast separately also improves the performance of detection. To improve the robustness of the detectors, we augment the data with affine transformations, such as rotation, translation, scaling and sheering.

**5.2.2. Breast, areola and nipple landmark detection—**we want a detector that can automatically detecting the contours for the left breast, right breast, areola, and nipple given the detected bounding boxes, which is also necessary for the breast cosmetic analysis.

For a given image $\mathbf{X}$ and the detected bounding box $\mathbf{b}_c$ for the component $c$, the landmarks $\mathbf{S}_c \in \mathbb{R}^{2 \times d_c}$ can be estimated by,

$$\mathbf{S}_c = G_c(\text{crop}(\mathbf{X}, \mathbf{b}_c)) \tag{2}$$

where crop(·, ·) is the image cropping function. $G_c$ is the landmark detection algorithm for the component $c$ and $d_c$ is the number of landmarks define for the component which is 30 for breast, 12 for areola and 6 for nipple (see Section 5.1.2 for more details).

Specifically, we use the algorithm proposed in Kazemi and Sullivan (2014) for breast landmark detection. The model itself is object agnostic thus suitable for modeling breast landmarks. Similar to the design in breast components detection, we also train separate landmarks detection models for left breast, right breast, nipple and areola. The input to the algorithm is the cropped image and bounding boxes detected by the breast, areola and nipple detectors.

**5.2.3. Image preprocessing for breast skin analysis—**To construct features from images, the images need to first be preprocessed to remove any man-made markings such as tattoos or markings on the breast skin.

The RGB image is first converted to Lab color space since the latter is closer human color perception.

After each image is converted into LAB color space, regions containing only two breasts (the areolas and nipples are excluded) are segmented from the entire image using landmark annotations. An exampled colorful image of breast segmentation result is shown in Fig. 4(a). Since the A and B channels represent color information, the A and B channels of the segmented breasts are chosen to mask out tattoos or markings on the breast skin. Specifically, for each segmented breast in LAB color space, a 2D histogram of its nonzero A and B channels' pixel values are estimated with Kernel Density Estimation to generate a probability map of nonzero pixel locations on the segmented breast. This probability map is then used to filter out small probability values with a threshold to generate a probability mask on the segmented breast. Applying this probability mask on the segmented breast, we get breast images with holes, as shown in Fig. 4(b), where tattoos or markings are removed from breast skin. Finally, the median values of nonzero skin pixels are calculated for each breast on A, B channel separately, to fill in those holes, as shown in Fig. 4(c). Resulting RGB image after this whole preprocessing procedure is shown in Fig. 4(d).

Note that in the later analysis, for each segmented breast, the same probability mask is also applied with the L channel and R, G, B channel, and holes are filled with median values of nonzero skin pixels in each corresponding channel.

### 5.3. Feature extraction

Once the landmarks and image crop for each breast component is extracted, we then need a set of useful features for the subsequent cosmesis score estimation. There are typically two categories for feature extraction: data-driven feature learning and hand-crafted feature design. The former, represented by Deep Convolutional Neural Networks (DCNN), is to learn the appropriate feature representation according to the training data, model architecture and objective function. However, this method typically requires a very large amount of training data, which is not the case in our study. The second method manually design each feature according to the domain specific knowledge, which is more suitable to our study due to the explicit clinical definition of all the comesis scores and the lack of very large annotated datasets.[2]

As described in Section 5.1.1, breast cosmesis score composes the global score and categorical scores focusing on breast size, breast shape, nipple appearance, scar appearance and skin color. In the following sections, we will describe our feature definitions based on our knowledge about individual scores.

**5.3.1. Breast shape and size feature—**To measure the shape and size[3] differences between the two breasts, we propose to use three features: Relative Breast Area Difference (dBA), Breast Bottom Location Difference (dBBL) and Procrustes Distance (dP).

1. Relative Breast Area Difference (dBA) is proposed to measure the size difference between left and right breasts. More specifically, two segmentation masks are generated for each breast according to their breast contour landmarks as in Fig. 4(a). The dBA's are defined as the normalized differences in the number of pixels between the two masks with three normalizing factors. Mathematically,

$$\mathrm{dBA}_i = \frac{|A_{\mathrm{lb}} - A_{\mathrm{rb}}|}{z_i}, \text{ for } i \in \{1, 2, 3\} \tag{3}$$

where $A_{\mathrm{lb}}$ and $A_{\mathrm{rb}}$ are the area of left and right breasts respectively. The normalizing factor $z_i$ for each of $i \in \{1, 2, 3\}$ are defined as,

$$z_1 = \left(d_{\mathrm{ip}}/2\right)^2, z_2 = \min(A_{\mathrm{lb}}, A_{\mathrm{rb}}), z_3 = \frac{A_{\mathrm{lb}} + A_{\mathrm{rb}}}{2} \tag{4}$$

where $d_{\mathrm{ip}}/2$ denotes half inter-pit distance as shown in Fig. 8(a), $d_{\mathrm{ip}} = \|\mathbf{S}_{\mathrm{lb}}[1,] - \mathbf{S}_{\mathrm{rb}}[1,]\|_2$.

2. Breast Bottom Location Difference (dBBL) is proposed to mimic one of the commonly used heuristic on comparing breast sizes and shapes. The dBBL is defined as the normalized height difference between the two breast, i.e.,

[2]Although our dataset is the largest of this kind, it is still small comparing to the datasets typically used in training deep neural networks which is generally in 100K–10M.

[3]The word "shape" and "size" used here are more aligned with the intuitive definition by physicians rather than the more technical definition in computer vision.

$$\mathrm{dBBL} = \frac{|h_{\mathrm{lb}} - h_{\mathrm{rb}}|}{0.5\, d_{\mathrm{ip}}} \tag{5}$$

where the $h_{\mathrm{lb}}$ and $h_{\mathrm{rb}}$ are the heights of the left and right breasts respectively. $d_{\mathrm{ip}}$ is the inter-pit distance of the patients, which is used to normalize the patient body size.

3. Procrustes Distance (dP) is a metric to measure the distance between statistical shapes (Dryden & Mardia, 2016). In this application, we want to measure the distance between the shapes of left and right breasts. Given the landmarks $\mathbf{S}_{\mathrm{lb}}$, $\mathbf{S}_{\mathrm{rb}}$ detected for left and right breasts, the Procrustes distance can be calculated by,

$$\mathrm{dP}(\mathbf{S}_{\mathrm{lb}}, \mathbf{S}_{\mathrm{rb}}) = \|\mathbf{S}_{\mathrm{lb}} - \hat{\beta}\mathbf{S}_{\mathrm{rb}}\widehat{\mathbf{R}} - \mathbf{1}_{d_c}\hat{\mathbf{t}}\|^2 \tag{6}$$

where $\mathbf{1}_{d_c}$ is a $d_c$-by-2 dimensional matrix with ones. The estimated scaling factor $\hat{\beta} \in \mathbb{R}$, rotation (with reflection) transformation $\widehat{\mathbf{R}} \in \mathbb{R}^{2\times 2}$ and translation vector $\hat{\mathbf{t}} \in \mathbb{R}^{2\times 1}$ are the least-square solution of the following orthogonal Procrustes problem,

$$\hat{\beta}, \widehat{\mathbf{R}}, \hat{\mathbf{t}} = \underset{\beta, \mathbf{R}, \mathbf{t}}{\arg\min} \|\mathbf{S}_{\mathrm{lb}} - \hat{\beta}\mathbf{S}_{\mathrm{rb}}\widehat{\mathbf{R}} - \mathbf{1}_{d_c}\hat{\mathbf{t}}\|^2 \tag{7}$$

**5.3.2. Nipple/areola shape and size features—**This set of features aim to model the intuition and concepts used by physicians when grading the nipple appearance[4] score. Similar to the last section, we will focus only on the shape and size as the color differences is graded in the skin color score per definition (Harris et al., 1979).

1. Areola Area Difference (dAA) measures the difference in areas between two areolas. Let us denote the area of left and right areola as $A_{\mathrm{la}}$ and $A_{\mathrm{ra}}$. The relative Areola Area Difference (dAA) feature can be defined as

$$\mathrm{dAA} = \frac{|A_{\mathrm{la}} - A_{\mathrm{ra}}|}{(A_{\mathrm{la}} + A_{\mathrm{ra}} + \epsilon)/2} \tag{8}$$

Comparing to the relative Breast Area Difference (dBA) features, we only use one normalizing factor here since the areola has a clear definition of enclosed contour comparing to the breast.

2. Areola Perimeter Difference (dAP) measures the perimeter difference between two areolas. Let us denote the perimeter of left and right areola as $P_{\mathrm{la}}$ and $P_{\mathrm{ra}}$ respectively. The relative Areola Perimeter Difference (dAP) can then be defined as

[4]This "appearance" is the common sense definition used by physicians rather than the technical definition used in Computer Vision.

$$\mathrm{dAP} = \frac{|P_{\mathrm{la}} - P_{\mathrm{ra}}|}{(P_{\mathrm{la}} + P_{\mathrm{ra}} + \epsilon)/2} \tag{9}$$

The perimeter of an areola can be calculated by

$$P_c = \sum_{i=1}^{11} \|\mathbf{S}_{c,i} - \mathbf{S}_{c,\mathrm{mod}(i+1,12)}\|_2, \text{ for } c \in \{\mathrm{la}, \mathrm{ra}\} \tag{10}$$

where mod($i$ + 1, 12) denotes the remainder of $i$ + 1 divided by 12 (recall that areola has 12 landmarks, see Section 5.2.2). $\mathbf{S}_{c,i}$ denotes the $i$th row of matrix $\mathbf{S}_c$ (landmark matrix for object $c$).

3. Areola Height Difference (dAH) and Areola Width Difference (dAW) measures the relative differences between the heights of areolas and the widths of areolas. They can be calculated by

$$\mathrm{dAH} = \frac{|h_{\mathrm{la}} - h_{\mathrm{ra}}|}{(h_{\mathrm{la}} + h_{\mathrm{ra}} + \epsilon)/2} \tag{11}$$

and

$$\mathrm{dAW} = \frac{|w_{\mathrm{la}} - w_{\mathrm{ra}}|}{(w_{\mathrm{la}} + w_{\mathrm{ra}} + \epsilon)/2} \tag{12}$$

where $h_{\mathrm{la}}$ ($h_{\mathrm{ra}}$) and $w_{\mathrm{la}}$ ($w_{\mathrm{la}}$) denotes the width and heights of the left (right) areola on the 2D image plane, see Fig. 5(a).

4. Areola Shape Difference (dAS) measures the difference between the fatness of the left and right areolas, which can be calculated by

$$\mathrm{dAS} = \frac{\|h_{1\mathrm{a}} - w_{1\mathrm{a}}| - |h_{\mathrm{ra}} - w_{\mathrm{ra}}\|}{\left(\sqrt{h_{1\mathrm{a}}^2 + w_{1\mathrm{a}}^2} + \sqrt{h_{\mathrm{ra}}^2 + w_{\mathrm{ra}}^2} + \epsilon\right)/4} \tag{13}$$

5. Local Nipple Location Difference (dLNL) is defined to measure the difference in terms of the location of nipple relative to areola for left and right breasts, which contains the difference in $x$ and $y$ axis on the 2D image plane, see Fig. 5(b).

$$\mathrm{dLNL} = (\mathrm{dLNL}_x, \mathrm{dLNL}_y) \tag{14}$$

$$\mathrm{dLNL}_x = \left|\frac{D_{hll}}{D_{hrl} + \epsilon} - \frac{D_{hlr}}{D_{hrr} + \epsilon}\right| \tag{15}$$

$$\mathrm{dLNL}_y = \left| \frac{D_{vtl}}{D_{vbl} + \epsilon} - \frac{D_{vtr}}{D_{vbr} + \epsilon} \right| \tag{16}$$

The features described above assume all areola and nipples are present in the image, which might not be the case (for example, missing due to operation). We described our heuristics for handling missing areola and nipple in Section 5.3.3.

**5.3.3. Heuristics for missing nipple and areola—**Since nipples and areolas from either side of the breast could be potentially missing, it is necessary to develop rules to handle these cases. Based on the detection result defined in Section 5.2.1, there are three possibilities for missing components: areola detected but nipple missing, areola missing but nipple detected, or both areola and nipple missing. The heuristics is described in Algorithm 2.

**Algorithm 2:** Heuristics for missing nipple and areola

**Input:** Detected breast, areola and nipple landmarks $\mathbf{S}_b$, $\mathbf{S}_a$, $\mathbf{S}_n$
**Result:** Pseudo nipple and areola center coordinate $\mathbf{Cp}_a$, $\mathbf{Cp}_n$

```
if S_n is None then
    if S_a is None then
        Cp_a = Cp_n = center of S_b;
    else
        Cp_a = Cp_n = center of S_a;
    end
else
    if S_a is None then
        Cp_a = Cp_n = center of S_n;
    else
        Cp_a = center of S_a;
        Cp_n = center of S_n;
    end
end
```

**5.3.4. Skin color features—**To measure the skin color difference, an intuitive method is to measure the divergence of the color distributions between the left and right breasts. We use Jensen–Shannon divergence (JSD) and Earth Mover distance (EMD) for the divergence measurement, which is more suitable in our application than Kullback–Leibler divergence (KLD) as the side of treated/untreated breast is unknown. Other ones are also tried but no good results are provided.

1. JS-divergence between left and right breast sampling pixels (JSD-sample): A mere divergence itself between left and right breasts might not be sufficient due to lack of reference of its scale. A large between-breast color difference could be accompanied by a large within-breast difference. Inspired by discriminant analysis methods (e.g., linear discriminant analysis, LDA), we proposed a number of features comparing color distributions considering both within-breast differences and between-breast differences.

   More specifically, given an image of a patient in LAB color space which decouple the illumination from colors (L for perceptual lightness, and A and B for human vision colors) and its corresponding breast segmentation mask (pre-processed according to Section 5.1.1), we first calculate per channel histogram

$p_{q,c}$ for each breast, with channel $q \in \{L, A, B\}$ and component $c \in \{lb, rb\}$, from $m$ pixels sampled uniformly randomly from non-zero pixels. To account for sampling variance, we repeat the aforementioned procedure $N$ times independently, resulting in $p_{q,c,i}$ with $i = 1, \ldots, N$, yielding a total of $3 \times 2 \times N$ histograms. The within-breast histogram distance ($d_{W,q,c,i,j}$) for the channel $q$, and breast $c$, between $i$th and $j$th sample is calculated by,

$$d_{w,q,c,i,j} = \mathrm{JSD}\left(p_{q,c,i} \| p_{q,c,j}\right) \tag{17}$$

with $i = 1, \ldots, N-1$ and $j = i+1, \ldots, N$. The $\mathrm{JSD}(\cdot\|\cdot)$ is the Jensen-Shannon divergence between two distributions. This yields a total of $3 \times 2 \times \binom{N}{2}$ distances (3 channels, 2 breasts, $\binom{N}{2}$ pairs within breasts). We represent these distances in six $\binom{N}{2}$-dimensional $\mathbf{d}_{w,q,c}$ vectors for each breast and LAB channel, with each element of the vector calculated from Eq. (17). Similarly, we can define between-breast histogram distance $d_{b,q,\mathrm{lb},\mathrm{rb},i,j}$ as

$$d_{b,q,\mathrm{lb},\mathrm{rb},i,j} = \mathrm{JSD}\left(p_{q,\mathrm{lb},i} \| p_{q,\mathrm{rb},j}\right) \tag{18}$$

with $i = 1, \ldots, N$ and $j = 1, \ldots, N$. This yields a total of $3 \times N^2$ distances (3 channels, $N^2$ pairs between breasts). We represent these distances in three $N^2$-dimensional $\mathbf{d}_{b,q}$ vectors for each breast and LAB channel, with each element inside the vector calculated from Eq. (18).

Once both $\mathbf{d}_{w,q,c}$ and $\mathbf{d}_{b,q}$ are calculated, we estimate the distributions of the distances across the elements of each vector, denoting as $p_{w,q,c}$ and $p_{b,q}$. The final JSD-sample feature set that will be use to characterize the color difference between the two breasts is calculated by,

$$\mathrm{JSD}_{q,c} = \mathrm{JSD}(p_{w,q,c} \| p_{b,q}) \text{ for } q \in \{L, A, B\}, c \in \{lb, rb\} \tag{19}$$

We repeat the entire process $T$ times and use the average $\mathrm{JSD}_{q,c}$ as our final feature to reduce variance of the estimate against sampling variation. This process yields a total of 6 features (2 breasts and 3 channels). We use $m = 100$, $N = 50$, $T = 5$.

**2.** JS-divergence between left and right breast sampling patches (JSD-patch): The feature described in the last section uses pixel samples across the entire breast, which characterize the overall color heterogeneity of the breast, but it might also overlook local color changes. In this feature, instead of sampling across the entire breast, we extract samples from small patches in the breast to better preserve the local color pattern.

More specifically, we uniformly randomly sample $N$ patch centers from the segmented breast image independently for each LAB channel. Given the $q$-

channel of the segmented image of component $c$, $\mathbf{X}_{q,c} \in \mathbf{R}^{w \times h}$, the area of each square is equal to $(2\alpha \min(w, h) + 1)^2$. $\alpha$ is a pre-determined parameter controlling the size of the square. This setup enables a variable patch size accounting for different image resolution while preserving the visual angle of each patch (similar to the physicians' grading process). If the number of non-zero pixels within the patch is smaller than a pre-defined percentage ($\theta$), the algorithm resamples the patch center until there is sufficient number of non-zero pixels within the patch. We show the pseudo-code of the patch sampling algorithm in Algorithm 3.

Once a patch is extracted according to Algorithm 3, a 255-bin histogram is used to approximate its distribution across all pixels, denoting as $p_{q,c,i}$ for the $i$th patch $\mathbf{X}_{q,c,i}$. Note that this notation of histogram is the same as in the last feature (List 1), which is an intentional choice since they serve the same purpose. Once $p_{q,c,i}$ is extracted, the exact same process as described in the last section is used to extract the 6 between breasts features and 3 within breasts features to characterize local color heterogeneity.

**Algorithm 3:** Algorithm for sampling patches

**Data:** $N = 200$, $\alpha = 0.08$, $\mathbf{X}_{q,c}$, $\theta = 0.35$
**Result:** $N$ patches with sufficient non-zero pixels within each patch $\mathbf{P}_{q,c,i}$

set half-width of the patch $r = \alpha \min(w, h)$;
**for** $i = 0$ **to** $N$ **do**
  set $n_p$ (number of non-zero, i.e., positive, pixels) = 0;
  **while** $n_p < \theta(2r + 1)^2$ **do**
    get patch center $c_i = (c_{i,x}, c_{i,y})$ uniformly and randomly sample from $\mathbf{X}_{q,c}$;
    get square patch $\mathbf{X}_{q,c,i} = \mathbf{X}_{q,c}[c_{i,x} - r : c_{i,x} + r, c_{i,y} - r : c_{i,y} + r]$;
    get number of non-zero pixels $n_p = \#\{P_{q,c,i} > 0\}$
  **end**
**end**

3. EMD between left and right breast sampling pixels (EMD-sample): Although JSD can measure the distance between two distributions, it remains constant between two non-overlapping distributions regardless of their distances (Arjovsky, Chintala, & Bottou, 2017). The Earth Mover's Distance (EMD) based feature introduced in this section and the next two are proposed to address this issue.

   The EMD-sample feature is exactly the same as the sample pixel based JSD feature (JSD-sample) introduced in List 1, except the use of Earth Mover's Distance (EMD). We replace JSD($\cdot \| \cdot$) with EMD($\cdot$, $\cdot$) for Eqs. (17)–(19). The EMD between two distribution represents the minimum amount of "work" needed to move the probability mass from one distribution to the other.

   We use the same set of features as in JSD-sample features (Section 1) with the additional maximum probabilities for within-breast histogram distance (6 features for two breasts and 3 channels) and between-breast histogram distance (3 features for 3 channels). This process yields a total of 15 (6+9) features.

4. EMD of pixel location & color difference between left and right breast (EMD-XYLAB): The four sets skin color features introduced earlier only considers the color distribution itself, without encoding the spatial information of the color pattern between breasts, which leads to identical features even if one randomly shuffle the pixels within the breast (and thus destroy all the spatial color patterns).

   To encode the spatial-color information, it is necessary for it to be shape-invariant, since we already modeled the shape information in Sections 5.3.1 and 5.3.2. Otherwise, the spatial-color pattern differences between two breasts might be due to the shape differences instead of skin color.

   To construct this feature, we first align the shape of each breast to the breast mean shape using non-linear warping according to the detected landmarks. For every aligned breast image, we extract spatial-color information (*x, y*, L, A, B) for all nonzero pixels in the left and right breast. The (*x, y*) coordinate for each pixel is defined with respect to their own coordinate system whose the origin located at the first landmark of the corresponding breast (Fig. 6(c)).

   The EMD-XYLAB feature set is then computed as the EMD of the spatial-color joint distributions between two breasts in both original and standardized (*x, y*, L), (*x, y*, A), (*x, y*, B), (*x, y*, A, B) and (*x, y*, L, A, B) spaces. This process yields a total of 10 features.

5. PCA in AB space (PCA-AB): For some patients, noticeable skin rashes can appear after the treatment. The gradual change in color from normal skin to skin rashes is reflected as high correlation between the A and B channels (of LAB space). One example is provided in Fig. 7, with left breast treated and right breast untreated. As shown in Fig. 7(c), the A and B channel of pixels within the left breast are significantly more linearly correlated than the untreated right breast. If we inspect the pixels along the axis of treated breast, it can be found that increasing values in both A and B channel will move the region from normal skin to skin rash (regions 1 to 2 to 3).

   To mathematically extract this information, we employ Principle Component Analysis (PCA) in the AB space of the two breasts independently, which will analyze the extent of linear correlation between A, B channels within each breasts.

   Let us denote the (variance-)covariance matrices for left and right breasts in A–B space as $\Sigma_{\mathrm{lb}}$ and $\Sigma_{\mathrm{rb}}$. Then the PCA solves the following eigenvalue decomposition problem,

$$\mathbf{V}_c \boldsymbol{\Sigma}_c = \boldsymbol{\Lambda}_c \mathbf{V}_c \text{ for } c = \{\mathrm{lb}, \mathrm{rb}\} \tag{20}$$

   where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_{c,1}, \lambda_{c,2})$ is the diagonal matrix of eigenvalues (in decreasing order) and the orthonormal matrix $\mathbf{V}$ contains columns of corresponding eigenvectors (or in this context Principle Components, PCs).

If the left breast's pixels are more correlated in A–B space than the right breast, then we will have $\lambda_{lb,2} < \lambda_{rb,2}$. We use $\lambda_{c,1}$, $\lambda_{c,2}$ and vectorized **V** of both left and right breast as our final feature. This yields a total of 12 (2 × 6) features.

## 5.4. Machine prediction

Once all features and ground-truth scores are obtained according to Sections 5.1.1 and 5.3, we can train a machine learning algorithm to automatically predict the Breast Cosmesis scores.

As described in Section 5.1.1, the Breast Cosmesis score has four levels for both global score and category criteria (breast size, breast shape, nipple appearance, skin color and scar appearance). The four levels are discrete quantization of a continuous perception of breast cosmesis quality. Thus the problem of Breast Cosmesis scores prediction can be viewed from three perspectives, as a classification problem, a regression problem or a ranking problem. We use Lasso regression, Kernel Support Vector Classification (kSVC) and OrdinalNet for the regression, classification and ranking formulation respectively.

### 5.4.1. Machine learning algorithms

1. Lasso Regression: Lasso (Tibshirani, 1996) (Least Absolute Shrinkage and Selection Operator) is a linear regression method with least-square loss and $l_1$ penalization. Once the Lasso model is fitted to the training set, variables corresponding to zero coefficients are discarded.

2. Kernel Support Vector Classification: Kernel Support Vector Machine is a classical supervised learning method developed primarily for classification purposes (Cortes & Vapnik, 1995) (which we will refer as kernel Support Vector Classification, kSVC, to discern from Support Vector Regression). One-vs-all (i.e., fitting one classifier per class) method is used to convert 4-class classification problem to four binary classification problems. kSVC has been used for breast cosmesis score classification previously in Cardoso and Cardoso (2007) and Cardoso et al. (2005). In this study, we use Radius Basis Function (RBF) kernel kSVC.

3. OrdinalNet: As Section 5.1.1 describes, the grading score provided by the physicians are naturally ordered as the 4-point scale represents Excellent, Good, Fair, Poor (which is also observed by Cardoso et al., 2005). It is thus necessary to model the relative relationship between the ratings, rather than treating their integer coding as merely continuous variable or discrete categories. To achieve this, we model the data with regularized ordinal regression, ordinalNet (Wurm, Rathouz, & Hanlon, 2017). Here we use General Linear Model (GLM), non-parallel formulation Wurm et al., 2017) and elastic net penalty (Zou & Hastie, 2003) to model relationship between covariate (breast cosmesis feature) and response (cumulative probability) with logit link function (McCullagh, 1980).

# 6. Experiments and results

A series of experiments are conducted to evaluate the performance of the proposed fully automatic algorithm, as well as the effectiveness of each of its components. In this section, we will describe the experimental details and results. Section 6.1 first describes the general testing framework which includes 10-fold and leave-one-trial-out (LOTO) cross-validation. The evaluation metric and results for breast components detection, landmarks detection and machine prediction are provided in Sections 6.2–6.4 respectively.

## 6.1. Testing framework

We test our algorithm using 10-fold cross-validation and leave-one-trial-out (LOTO) cross-validation. The former provides a well balanced (in terms of bias and variance) estimation of the generalization error and the latter provides an estimate on performance when the applying our algorithm on a completely new trial.

**6.1.1. 10-Fold cross-validation—**The entire dataset is randomly split into 10 non-overlapping partitions. Nine partitions will be used for training and the remaining partition will be used for testing. This process repeats 10 times until all the partitions have been served as training and testing set. The aggregated evaluation is then reported as the final performance of the algorithm.

**6.1.2. Leave-one-trial-out (LOTO) cross-validation—**Instead of using random data split for training and testing as in 10-fold, samples from one trial will be served as testing set and the rest as training set. The process will also repeat until all trials have been served as training and testing set. We also combine the trial NSABP B39 and trial RTOG 041 due to the significantly smaller number of samples from RTOG 1014 trial (89 samples after selection).

To our knowledge, it is the first time such algorithm being tested in an across-trial setup, which poses significant challenge to the algorithm. One should also notice that because our dataset only has two trials with sufficient number of samples, the LOTO cross-validation results provided here should be interpreted as an anecdotal report on the across-trial generalization of the algorithm. One should be cautious on generalizing reported performance to the cases where high statistical power is required.

## 6.2. Breast, areola and nipple detection

**6.2.1. Evaluation metric—**We use mean Average Precision (mAP) metric and F1-score as evaluation metrics for our breast components detectors, which are commonly used in object detection literature (Everingham, Van Gool, Williams, Winn, & Zisserman, 2010; Lin et al., 2014). More specifically, a true positive detection is when intersection-over-union (IOU) between predicted and ground-truth bounding boxes is above a pre-defined threshold (we use 0.5 as in Everingham et al., 2010; Lin et al., 2014).

The mean Average Precision (mAP) for each detector is calculated by the area under the precision–recall curve. We use the same 11-point interpolation method introduced in Everingham et al. (2010) to calculate the mAP score. Since in our case, each detector only

focuses on one class of object, we do not distinguish mAP and AP, which is also commonly the case as in Lin et al. (2014). The final mAP and F1-score is aggregated across folds and trials in the 10-fold and LOTO cross-validation.

**6.2.2. Results—**Table 2 shows mAP and F1 score results of each detector tested in 10-fold and LOTO cross-validation. The first block (first and second rows) shows the average mAP and F1 score (s.e. in parenthesis) in 10-fold cross-validation. The second and third block shows results in LOTO cross-validation. Since there are two trials in total, we report the performance in each LOTO fold instead of aggregating the results (using mean and s.e.).

As shown in the Table 2, all of the breast components detectors can perform with > 0.85 mAP and F1 score even if we train and test them using two independent trials, with very different image and score distributions. Among all the detectors, the left/right breast detector can accurately detect almost all the breasts at IOU threshold = 0.5, regardless of testing setups. The areolas and nipples detector perform slightly worse as their size much smaller. For reference, a state-of-the-art generic object detection algorithm like EfficientDet has mAP at around 0.734 tested on Microsoft COCO dataset (Tan, Pang, & Le, 2020), trained on millions of images. Permutation tests are also conducted on all the experiments in this table, which demonstrate a $p$-value < 0.002 and thus suggest statistical significance.

### 6.3. Breast, areola and nipple landmarks detection

**6.3.1. Evaluation metric—**Given a set of detected landmarks, We first use the interpolation method described in Section 5.1.2 to create pseudo-landmarks, which will provide identical landmarks as long as the detected curve is the same as the ground truth.

Once the pseudo-landmarks are properly resampled, we measure the detection error by the Euclidean distance between detected landmarks and the ground truth. To mitigate differences in image resolution and breast sizes, the error is normalized by half-inter-pit distance (see Fig. 8(a)).

Two methods are used to aggregate the error in landmarks detection, **within image** and **within landmarks**. The former calculates mean error across landmarks within an image then the mean and standard error of the mean across images, which indicates the robustness of the algorithm for different images with different skin tone, breast shape, illumination, etc. The latter calculates mean error and its standard error per landmark across images, which aims to show the performances differences across landmarks (which landmarks are more susceptible to detection error). These two evaluation will be reported in the next section as quantitative results.

**6.3.2. Results—**Table 3 shows the quantitative evaluation results for our landmarks detectors. More specifically, the table shows the *within-image* evaluation under 10-fold cross-validation. The table shows that both left and right breast landmarks detection has an average error below 0.02 (2% of half inter-pit distances) across subjects. Performance is slightly worse for areola due to potential ambiguous boundaries. Accuracy computed with normalized $l_2$ between the ground truth and detected landmarks smaller than a threshold

is also provided. Results show that LOTO breast landmark detection deteriorates and thus suggest biases of the model. Overall, the results show that the algorithm can perform well across subjects and conditions even if the training and testing set are very different (from LOTO cross-validation). Permutation tests are also conducted on all the landmark detectors in this table, which demonstrate a $p$-value < 0.002 and thus suggest statistical significance.

Fig. 8 shows quantitative evaluation for within-landmark metric. Fig. 8(a) shows an intuitive visualization of the within-landmark error across images. The radius around each landmark indicates the error mean where the thickness of the circle indicates one s.e. This figure shows that although the detection error increases moving from arm pit to the chest center, the overall magnitude is still consistently small w.r.t breast size regardless of image condition. Fig. 8(b)–(e) also provide bar plots for landmark detection error, which demonstrate a high detection accuracy.

### 6.4. Full pipeline

In this section, we report the performance of the full pipeline of the proposed automatic breast cosmesis score estimation algorithm, including the breast components detection, landmarks detection (which both are tested independently in previous sections) and the data-driven predictors defined in Section 5.4.

**6.4.1. Evaluation metric—**We use 4-way classification accuracy as the evaluation metric used for all the Lasso, kSVC, OrdinalNet algorithms described in Section 5.4, due to its intuitive interpretation. For Lasso regression, the predicted response is first rounded to the closest integer. Once predicted score is acquired for each machine learning methods across all the cross-validation folds, the mean accuracy and its standard error (s.e.) will be reported for performance evaluation.

**6.4.2. Additional setup—**Beside the 10-fold and LOTO cross-validation setup that used in both evaluating components and landmarks detectors, we also tested our fully automatic pipeline with different feature sets. As mentioned in Section 1, although our study is the first (to our knowledge) to develop a fully automatic algorithm for breast cosmesis score estimation, there is nevertheless previous attempts on using computer vision technique to partially automatize the grading procedure such as BCCT.core algorithms (Brouwers et al., 2016; Cardoso & Cardoso, 2007; Cardoso et al., 2005). By adapting the BCCT.core features to our fully automatic framework, we can compare the BCCT.core feature with our full feature set, gaining more insight on the effectiveness of our machine learning predictors and additional features.

The feature sets used for testing are, BCCT features (adapted to fully automated landmarks section), full features (BCCT features + our additional features as described in Section 5.3) and Lasso selected features described in Section 1. The Lasso selected feature is used only with kSVC since the Lasso Regression and OrdinalNet both already equipped with $l_1$ penalty.

As discussed in 1, Lasso regression selects a subset of the original $p$ dimensional breast cosmesis features defined in Section 5.3 ($p$ = 298 in our study). To further understand

the performance differences between Lasso regression and kSVC is due to the feature set differences or the algorithmic differences, we apply kSVC on the feature set selected by Lasso. The number of features selected by Lasso from the full set for each global and categorical score are provided in Table 4. As shown in this table, the number of selected features vary according to the breast cosmesis aspects. For skin color score and scar appearance score, the selected features (i.e., those with non-zero Lasso regression coefficients) mainly include the skin color features described in Section 5.3.4, and, the Relative Breast Area Difference (dBA) feature which measures the size difference between left and right breast in Section 5.3.1. While for breast size, shape and nipple appearance score, they are mainly about the breast, nipple and areola shape and size features as in Sections 5.3.1 5.3.2. For global score, they are mainly about the breast shape and size features in Section 5.3.1. From these feature selection results, it is obvious that breast shape and size features play an essential role in determining all the breast cosmetic scores. This correlates with previous studies in Fitzal et al. (2007) and Soror et al. (2016) demonstrating that the breast symmetry index is important for breast cosmesis evaluation.

**6.4.3. Results—**Fig. 9 shows the bar plot for average cross-validation accuracy for our algorithm under 3 machine learning (ML) frameworks estimating 6 scores (1 global score + 5 categorical scores). The height of the each bar denotes average accuracy using the corresponding ML framework and feature set and the error bar denotes standard error of the mean. *A priori* chance level is 25% for each score.

As shown in Fig. 9, when Lasso regression or OrdinalNet is used, the proposed full feature set outperforms or as good as BCCT features. When kSVC is used the proposed full feature set outperforms or equal to BCCT features in all but nipple score. This shows that our fully-automatic Breast Cosmetic evaluation pipeline with proposed feature set achieves comparable performance to BCCT.core without human supervision.

In practice, physicians' grading on size and shape are typically affected by other cosmesis like breast scars and nipple appearances, and grading on skin color and scare are often interfered with each other. These scores require inferences from all cosmetic aspects. Therefore, for global, size and shape scores, one may use OrdinalNet on all features, and for skin color and scare score, one need to use kSVC on all features. For nipple score, one may use kSVC on Lasso selected features since physicians' focuses are only limited to the nipple and areola regions during grading.

To test the generalizability of each of the methods across different trials, we also run experiments with LOTO cross-validation setup. The results is provided in Table 5.

## 6.5. Ablation study

In Sections 6.2 and 6.3, we directly measure the performance of the breast components and landmarks detection algorithm using commonly used performance metric. However, a high mAP or F1-score does not necessarily means effective for the downstream cosmesis score prediction. To further test the effectiveness of the breast components detection and landmarks detection in the full pipeline, we ablate both detection modules with ground truth physician annotation. If our fully automatic framework can produce similar result with the

human landmarks annotation, then it is reasonable to conclude that both components and landmarks detection are effective.

Additionally, we also tests the proposed algorithm with additional information on the side of treated/untreated breast. This information is added by reordering the left/right breast feature concatenation in PCAAB, EMD-XYLAB, KMEANS-XYLAB, BOVW-XYLAB and GLCM features. In a practical application, when a physician inputs a patient's image to the system, it is usually trivial to provide additional information on which side of the breasts is treated/untreated. However, if a system can perform equally well without such information, the capability of such system to process large batches of images without human input can be drastically improved.

Table 6 shows results for the ablation study when using OrdinalNet methods. The other two ML methods behaves similarly in this experiment. The table shows that when replacing the detected landmarks with ground-truth landmarks, the performance change of the final prediction is negligible. Same applies when adding treated/untreated information to the algorithm. This shows that our detection framework can performs effectively in the fully automatic framework and the additional treated/untreated breast input is not necessary.

## 7. Conclusion and discussion

In this study, we propose a fully-automatic Breast Cosmetic appearance grading system, which contains automatic breast components detection, landmarks detection and cosmesis score prediction. We evaluated both components detection and landmarks detection individually with quantitative methods, showing that both detectors are accurate and robust. When combined with machine learning predictors, both detector can provide sufficient information to the level similar to human annotation. We also propose a new set of features characterizing breast cosmesis score estimation, which is shown to be more effective and accurate than the previous attempts.

Hindrances to clinical adoption of our proposed method lie in concerns about the ability to integrate AI framework into the existing clinical workflow, the issues on clinical breast data privacy and integrity, the biases of models when trained on specific clinical trials but applied on a different one, the lack of ethics regulation on AI system applications, the acceptance and explainability of AI diagnosing breast cosmesis by patients, and costs and return on investment (ROI). More data from different clinical trials are also required for training a robust system to reduce the inherent biases. Ideally, if we can make this whole framework a website and available publicly, radiation oncologists will be able to send images of their patients and evaluate the cosmetic outcomes objectively compared to their own grading. Therefore, the breast database can grow by collecting more data, and the radiation oncologists will have the opportunity to improve the radiation delivery and get better breast cosmesis results for the patients.

However, there are still limitations of the proposed system. Areolas or nipples are occasionally difficult to be detected, particularly when they are inverted or occluded. This issue could be addressed with a finer-grained annotation. It is also hard for our algorithm to

predict on the scar appearance score, partially because of inconsistent physicians' grading and vague definition of scars. Besides, the prediction results skew to class excellent due to the unbalanced data distribution with excellent as the head class, suggesting difficulties in other down-stream Machine Learning tasks. Finally, our model purely operated in 2D image plane, where physician's grading on breast shape and size are typically based on their 3D understanding of the breast. Thus a 3D reconstructed breast model might be more suitable to describe breast shape or size (volume).

In conclusion, a fully-automatic Breast Cosmesis analysis system was proposed as a more objective and time-efficient method assisting in physicians' evaluation on breast cancer recovery and patients' decision on breast reconstruction surgery. We believe this grading system, together with its design methods shown in this paper, provides great potential for improving Breast Cancer patients' life quality, improving productivity of physicians and might be potentially beneficial for other breast cancer medical imaging research.

## Acknowledgments

This work was supported by the FUJIFILM Medical Systems/RSNA Research Resident, USA Grant, NCI U10CA180868, U10CA180822.

### Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Chenqi Guo reports financial support was provided by The Ohio State University. Chenqi Guo reports a relationship with The Ohio State University that includes: employment and funding grants.

## Data availability

All the codes, trained models and data are released on Code Ocean.

## Appendix

## Appendix

## Appendix A. Implementation details, running time, and memory usage

Here we provide the implementation details as follow:

1. MATLAB is used for the implementation of image color standardization before subjective grading, as in Section 5.1.1.
2. For the breast, areola and nipple detection described in Section 5.2.1, we use a PyTorch-YOLOv3 implementation https://github.com/eriklindernoren/PyTorch-YOLOv3/.
3. For the breast, areola and nipple landmark detection provided in Section 5.2.2, we use the Ensemble Regression Trees landmark detector implemented by Python dlib.shape_predictor package.
4. Python is used to implement the image preprocessing in Section 5.2.3 and feature extractions in Section 5.3.

5. For the Machine Prediction described in Section 5.4, Lasso regression is implemented with Python sklearn.linear_model.Lasso package, Kernel Support Vector Classification (kSVC) with Python sklearn.svm.SVC package, and OrdinalNet with R language ordinalNet package.

Running time and memory usage are also provided when running on a GPU machine with Quadro RTX 6000 GPU, Intel i9-7900X CPU and 128 GB memory, as in Table 7.

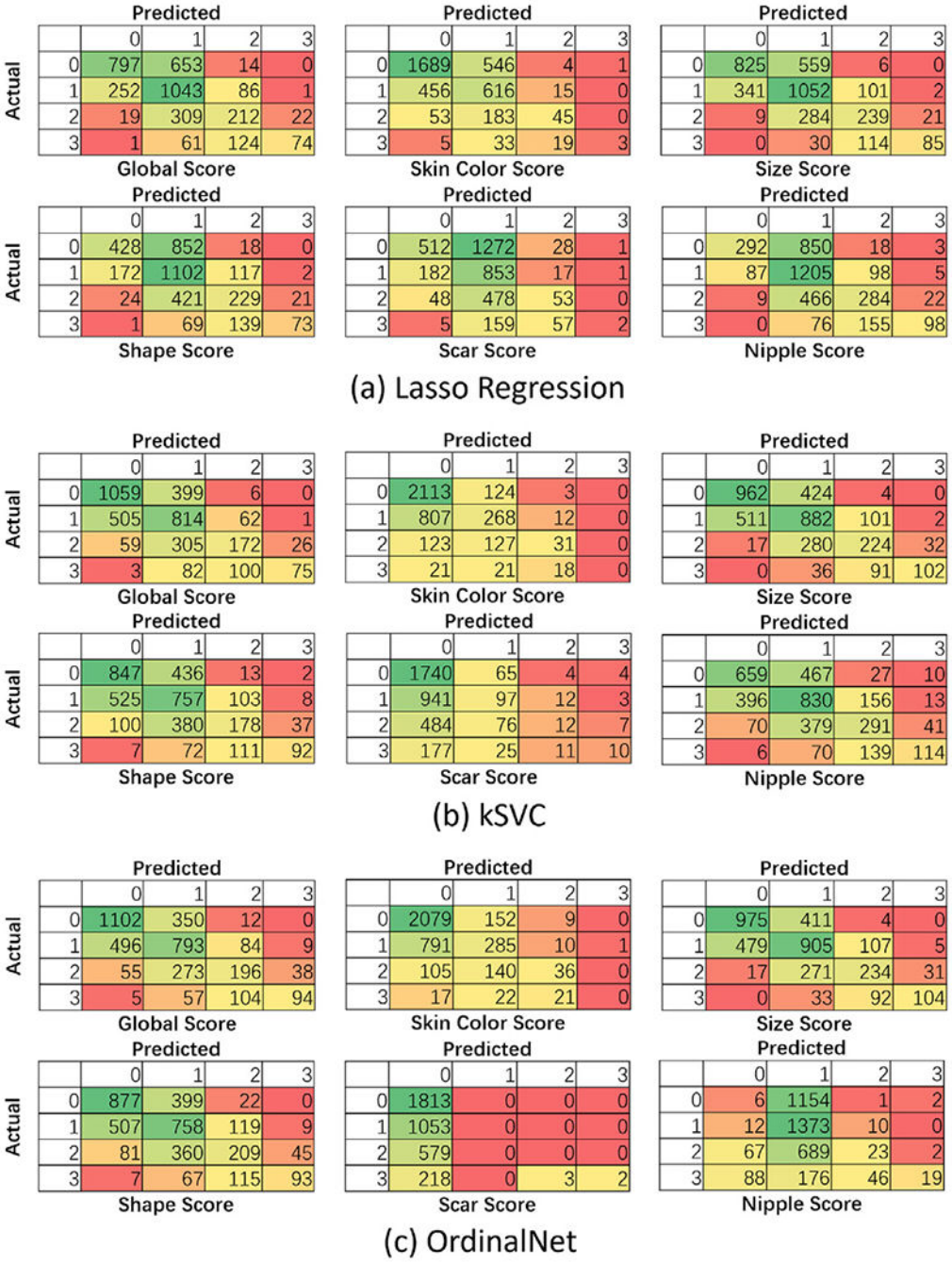


**Fig. 10.**

Confusion matrices of our framework with 3 machine learning algorithms estimating 6 scores (1 global score and 5 categorical scores).

## Appendix B. Extracted breast cosmesis feature summary

Feature extraction is one of the key factors of our fully-automatic breast cosmesis evaluation framework. Here we provide a detailed table of all the 298 features for illustration to help the readers better understand, as in Table 8.

Previous studies Fitzal et al. (2007) and Soror et al. (2016) have demonstrated that the breast symmetry index is important for breast cosmesis evaluation. In our proposed full breast cosmesis feature set shown in Table 8, the features dBA, dBBL, dP, BRA, LBC, BCD, BAD, BOD are related to the symmetry index.

## Appendix C. Experimental results of patients with different levels of cosmesis score

To show our prediction performance on the patients with different levels of the breast cosmesis score, Fig. 10 provides the confusion matrices on each score using 3 machine learning algorithms (Lasso regression, kSVC and OrdinalNet), full pipeline 10-fold Cross Validation and the feature set with highest accuracy for each algorithm (as in Fig. 9).

**Table 8**

List of full breast cosmesis feature set (including the BCCT.core features adapted to our fully-automatic framework).

| Feature name | Feature numbers | Mathematical meaning | Clinical significance |
|---|---|---|---|
| dBA[a] | 3 | Relative breast area difference | Breast size and shape asymmetry |
| dBBL[a] | 1 | Breast bottom location difference | Breast size and shape asymmetry |
| dP[a] | 1 | Procrustes distance | Breast size and shape asymmetry |
| dAA | 1 | Areola area difference | Nipple appearance asymmetry |
| dAP | 1 | Areola perimeter difference | Nipple appearance asymmetry |
| dAH | 1 | Areola height difference | Nipple appearance asymmetry |
| dAW | 1 | Areola width difference | Nipple appearance asymmetry |
| dAS | 1 | Areola shape difference | Nipple appearance asymmetry |
| dLNL | 3 | Local nipple location difference | Nipple appearance asymmetry |
| JSD-sample | 6 | JS-divergence between left and right breast sampling pixels | Skin color and scar appearance |
| JSD-patch | 9 | JS-divergence between left and right breast sampling patches | Skin color and scar appearance |
| EMD-sample | 15 | EMD between left and right breast sampling pixels | Skin color and scar appearance |
| EMD-global | 2 | EMD between left and right breast globally | Skin color and scar appearance |
| EMD-XYLAB | 10 | EMD of pixel location & color difference between left and right breast | Skin color and scar appearance |

| Feature name | Feature numbers | Mathematical meaning | Clinical significance |
|---|---|---|---|
| PCA-AB | 12 | PCA eigenvalues & eigenvectors in breast AB space | Skin color and scar appearance |
| KMEANS-XYLAB | 90 | K-means clustering of pixel location-color information for left and right breast | Skin color and scar appearance |
| BoVW-XYLAB | 40 | Bag-of-Visual-Word on pixel location-color information for left and right breast | Skin color and scar appearance |
| GLCM-texture | 24 | Grey Level Co-occurrence Matrix texture measurement on left and right breast | Skin color and scar appearance |
| BRA[a,b] | 1 | Breast retraction assessment | Breast size and shape asymmetry |
| LBC[a,b] | 1 | Level of lower breast contour | Breast size and shape asymmetry |
| BCD[a,b] | 1 | Breast contour difference | Breast size and shape asymmetry |
| BAD[a,b] | 1 | Breast area difference | Breast size and shape asymmetry |
| BOD[a,b] | 1 | Breast overlap difference | Breast size and shape asymmetry |
| UNR[b] | 1 | Upward nipple retraction | Nipple appearance asymmetry |
| BCE[b] | 1 | Breast compliance evaluation | Nipple appearance asymmetry |
| ang-sec EMD[b] | 65 | EMD between angular sections of left and right breast | Skin color and scar appearance |
| ar&np EMD[b] | 5 | EMD between left and right areola+nipple region | Skin color and scar appearance |

[a]Features related to the breast symmetry index.

[b]Features adapted from the BCCT.core algorithm.

As demonstrated in these confusion matrices, score fair and poor, as well as score excellent and good, are difficult to be distinguished from each other, partly due to inconsistent physicians' grading and vague definition of fair versus poor and excellent versus good during the subjective grading sessions. However, our algorithm hardly predicts score excellent as poor and vice versa, suggesting a decent performance on binary classification task. In addition, the prediction results skew to class excellent (1 for global score and 0 for other cosmesis scores) because of the unbalanced data distribution with excellent as the head class. Scar scores which is hard to detect from digital images, show unsatisfactory performance with OrdinalNet algorithm by predicting most of the sample as excellent. Global scores achieve better performance compared with other breast cosmesis scores in terms of the True Positive (TP) and thus have the highest precision values.

## References

Althobaiti MM, Ashour AA, Alhindi NA, Althobaiti A, Mansour RF, Gupta D, et al. (2022). Deep transfer learning-based breast cancer detection and classification model using photoacoustic multimodal images. BioMed Research International.

Arjovsky M, Chintala S, & Bottou L (2017). Wasserstein GAN. arXiv:1701.07875.

Brouwers P, Werkhoven E, Bartelink H, & Fourquet A (2016). Factors associated with patient-reported cosmetic outcome in the Young boost breast trial. Radiotherapy and Oncology.

Cardoso JS, & Cardoso MJ (2007). Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment. Artificial Intelligence in Medicine.

Cardoso MJ, Cardoso JS, Wild T, Krois W, & Fitzal F (2009). Comparing two objective methods for the aesthetic evaluation of breast cancer conservative treatment. Breast Cancer Research and Treatment.

Cardoso JS, da Costa JFP, & Cardoso MJ (2005). Modelling ordinal relations with SVMs: An application to objective aesthetic evaluation of breast cancer conservative treatment. Neural Networks.

Christie D, Sharpley C, & Curtis T (2005). Improving the accuracy of a photographic assessment system for breast cosmesis. Clinical Oncology.

Cortes C, & Vapnik V (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

Desantis CE, Ma J, Sauer AG, et al. (2017). Breast cancer statistics, 2017, racial disparity in mortality by state. CA: A Cancer Journal for Clinicians, 67(6), 439–448, [PubMed: 28972651]

Dryden IL, & Mardia KV (2016). vol. 995, Statistical shape analysis: with applications in r. John Wiley & Sons.

Early Breast Cancer Trialists' Collaborative Group (EBCTCG) (2011). Effect of radiotherapy after breast-conserving surgery on 10 year recurrence and 15 year breast cancer death: Meta-analysis of individual patient data for 10 801 women in 17 randomised trials. Lancet, 378(9804), 1707–1716. [PubMed: 22019144]

Everingham M, Van Gool L, Williams CK, Winn J, & Zisserman A (2010). The pascal visual object classes (voc) challenge. International Journal of Computer Vision, 88(2), 303–338.

Fitzal F, Krois W, Trischler H, Wutzel L, Riedl O, Kuhbelbock U, et al. (2007). The use of a breast symmetry index for objective evaluation of breast cosmesis. The Breast.

Harris JR, Levene MB, Svensson G, & Hellman S (1979). Analysis of cosmetic results following primary radiation therapy for stages I and II carcinoma of the breast. International Journal of Radiation Oncology Biology Physics.

Kazemi V, & Sullivan J (2014). One millisecond face alignment with an ensemble of regression trees. CVPR.

Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. (2014). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740–755). Springer.

van Maaren MC, de Munck L, de Bock GH, et al. (2016). 10 year survival after breast-conserving surgery plus radiotherapy compared with mastectomy in early breast cancer in the netherlands: A population-based study. The Lancet Oncology, 17(8), 1158–1170. [PubMed: 27344114]

McCullagh P (1980). Regression models for ordinal data. Journal of the Royal Statistical Society. Series B. Statistical Methodology, 42(2), 109–127.

Pataky RE, & Baliski CR (2016). Reoperation costs in attempted breast-conserving surgery: A decision analysis. Current Oncology, 23(5), 314–321. [PubMed: 27803595]

Ragab M, Albukhari A, Alyami J, & Mansour RF (2022). Ensemble deep-learning-enabled clinical decision support system for breast cancer diagnosis and classification on ultrasound images. Biology, [ISSN: 2079-7737] 11(3), http://dx.doi.org/10.3390/biology11030439.

Razdan SN, Cordeiro PG, Albornoz CR, Ro T, Cohen WA, Mehrara BJ, et al. (2016). Cost-effectiveness analysis of breast reconstruction options in the setting of postmastectomy radiotherapy using the BREAST-q. Plastic and Reconstructive Surgery.

Redmon J, Divvala SK, Girshick RB, & Farhadi A (2016). You only look once: Unified, real-time object detection. CVPR.

Sareigo J (2008). Regional variation in breast cancer treatment throughout the united states. The American Journal of Surgery, 196(4), 572–574, [PubMed: 18809065]

Sekhon A, Zhao R, Wang Y, Grant D, Winter KA, Moughan J, et al. (2017). Creating a review process of a digital photo database collected on NRG NSABP B39/RTOG 0413 phase III clinical trial for evaluation of cosmetic results from breast conserving therapy (BCT) [abstract]. Proceedings of the American Association for Cancer Research Annual Meeting.

Siegel RL, Miller KD, Fuchs HE, & Jemal A (2021). Cancer statistic 2021. CA: A Cancer Journal for Clinicians, 71(1), 7–33. [PubMed: 33433946]

Soror T, Kovacs G, Kovacs A, Seibold N, Melchert C, Baumann K, et al. (2016). New objective method in reporting the breast cosmesis after breast-conservative treatment based on nonstandardized photographs: The objective breast cosmesis scale. Brachytherapy.

Tan M, Pang R, & Le QV (2020). Efficientdet: Scalable and efficient object detection. In Proceedings of the ieee/cvf conference on computer vision and pattern recognition (pp. 10781–10790).

Tibshirani R (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B. Statistical Methodology, 58(1), 267–288.

Tuttle TM, Jarosek S, Habermann EB, et al. (2012). Omission of radiation therapy after breast-conserving surgery in the United States. Cancer, 118(8), 2004–2013. [PubMed: 21952948]

Volders JH, Negenborn VL, Haloua MH, et al. (2017). Cosmetic outcome and quality of life are inextricably linked in breast-conserving therapy. Journal of Surgical Oncology, 115(8), 941–948. [PubMed: 28334419]

Wurm MJ, Rathouz PJ, & Hanlon BM (2017). Regularized ordinal regression and the ordinalnet r package. arXiv preprint arXiv:1706.05003.

Zou H, & Hastie T (2003). Regression shrinkage and selection via the elastic net, with applications to microarrays. JR Stat Soc Ser B, 67, 301–320.

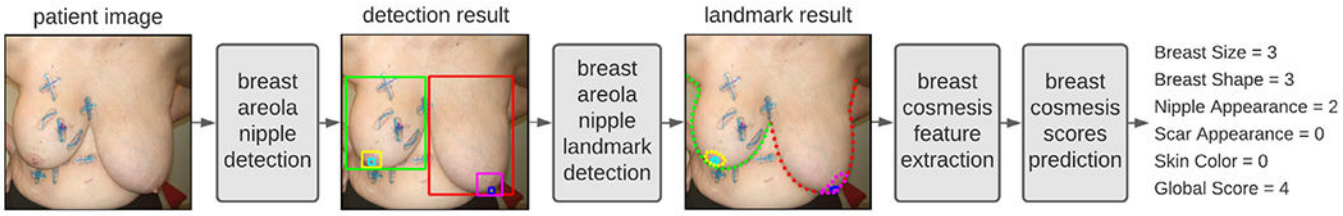


**Fig. 1.**
Overview of the proposed Breast Cosmesis evaluation system.

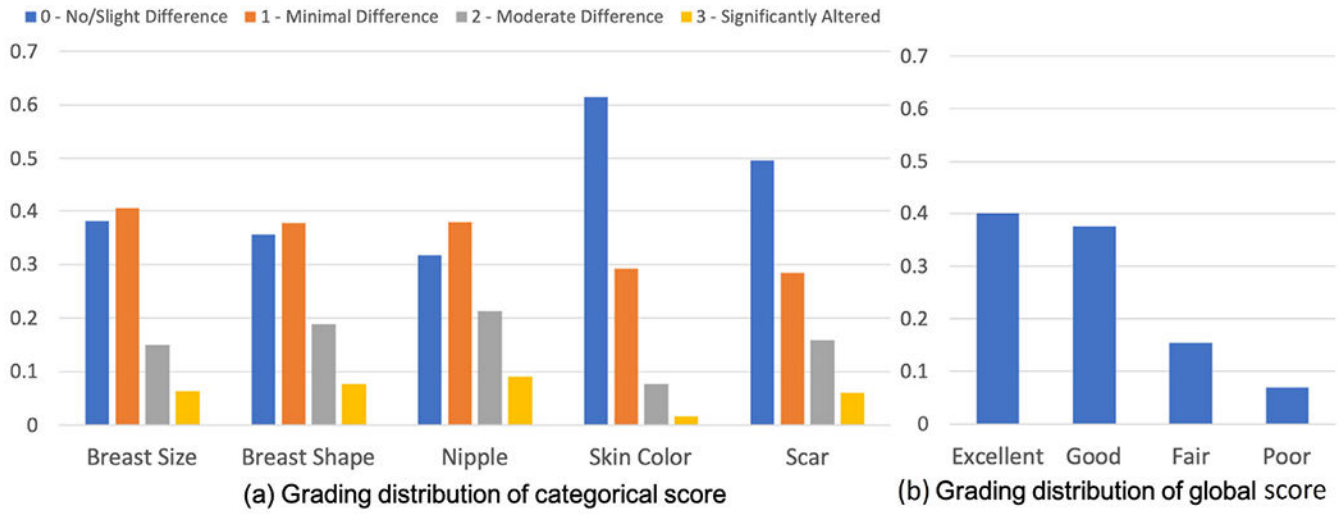


**Fig. 2.**
Score distributions for (a) individual categories and (b) global score.

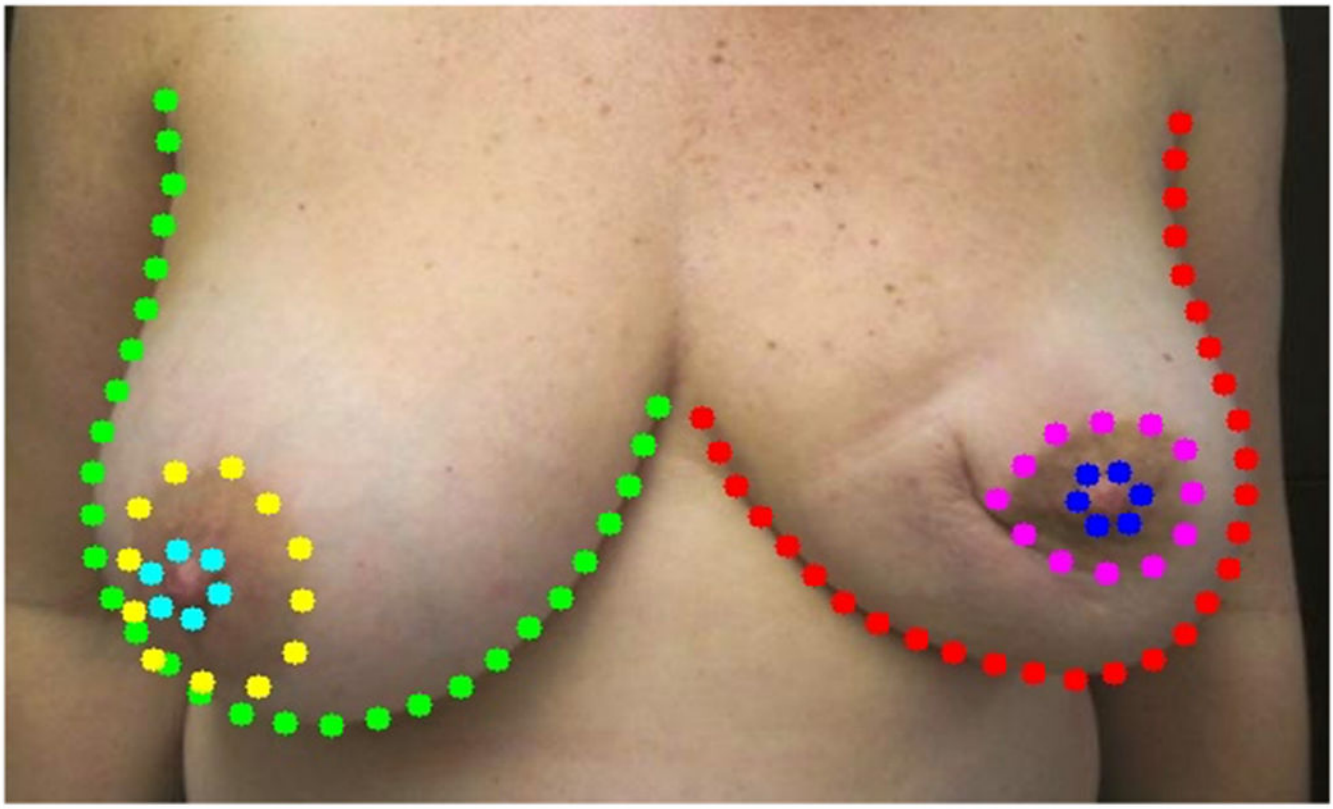

**Fig. 3.**
Landmark annotations used in our study. There are 30 landmarks on each breast, 12 landmarks on each areola, and 6 landmarks on each nipple.

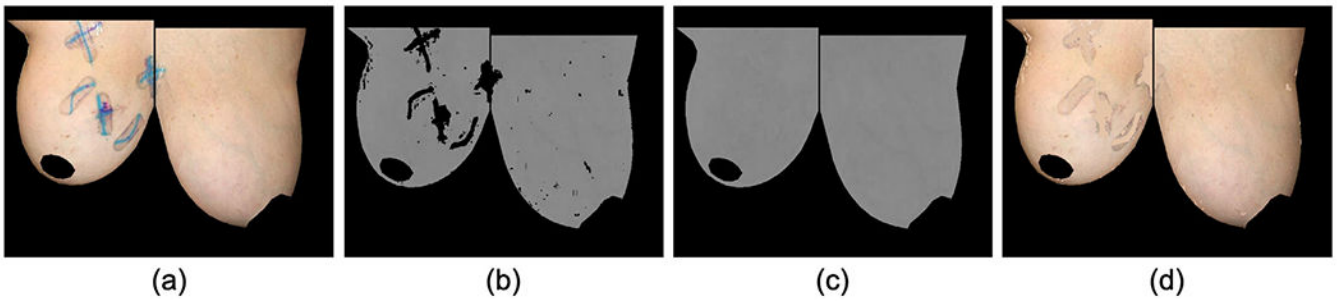


**Fig. 4.**

Image preprocessing to mask out tattoos or markers on breast skin. (a) Segmentation from landmarks. (b) A channel after thresholding. (c) A channel after correction. (d) Resulting RGB image after this preprocessing.

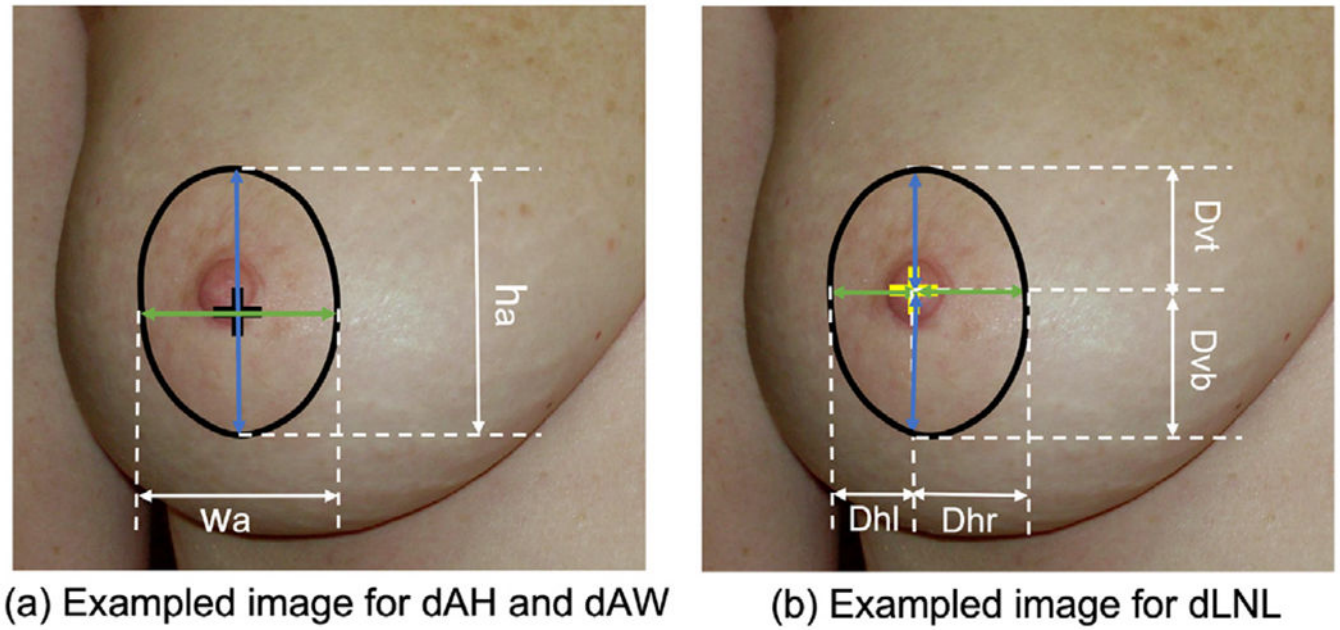


**Fig. 5.**

Here the areola is highlighted using the black contour. (a) This diagram illustrates the definitions of the width and height of areola, denoted as $\omega_a$ and $h_a$, respectively, when the areola exists. The black cross denotes the center of areola. If the areola does not exist but the nipple exists, compute these values from the nipple instead. (b) For cases where both the areola and nipple exist, this diagram illustrates the definitions of the nipple location relative to the areola, denoted as $D_{ut}$, $D_{ub}$, $D_{hl}$, and $D_{hr}$. The yellow cross denotes the center of nipple. If only the areola exists, then these relative locations are set to be semi-vertical and semi-horizontal axis length of the existing areola. If only the nipple exists, then these relative locations are set to the semi-vertical and semi-horizontal axis length of the existing nipple. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

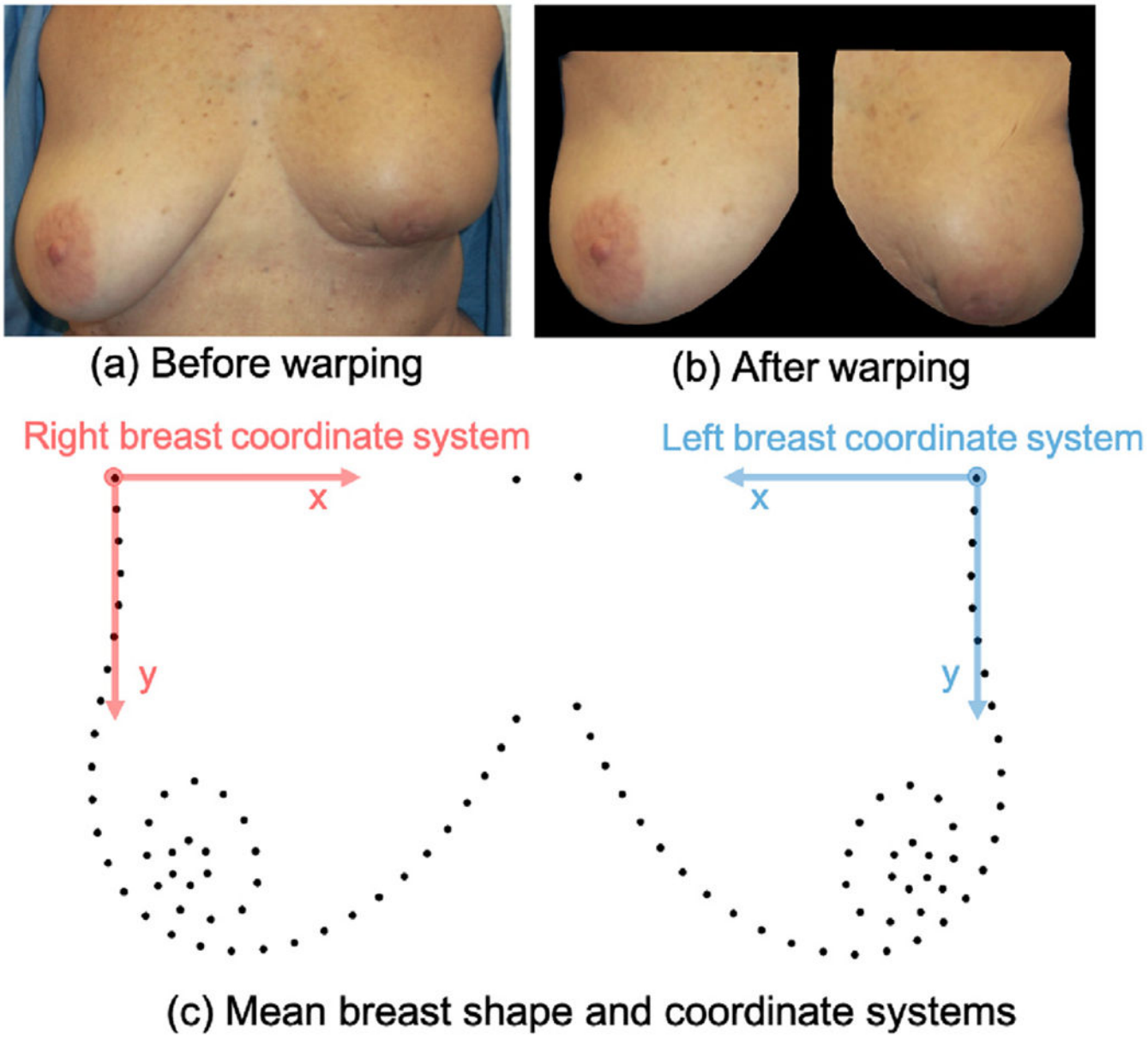


**Fig. 6.**

Example of breast mean shape warping and coordinate systems. To calculate EMD-XYLAB features, each breast of the patient needs to be warped to the mean breast shape of the same side. (a) A patient's image before warping. (b) The same image after warping and masking. (c) mean breast shape used for warping and left (right) coordinate system for pixel location ($x$, $y$) used in the EMD-XYLAB features.

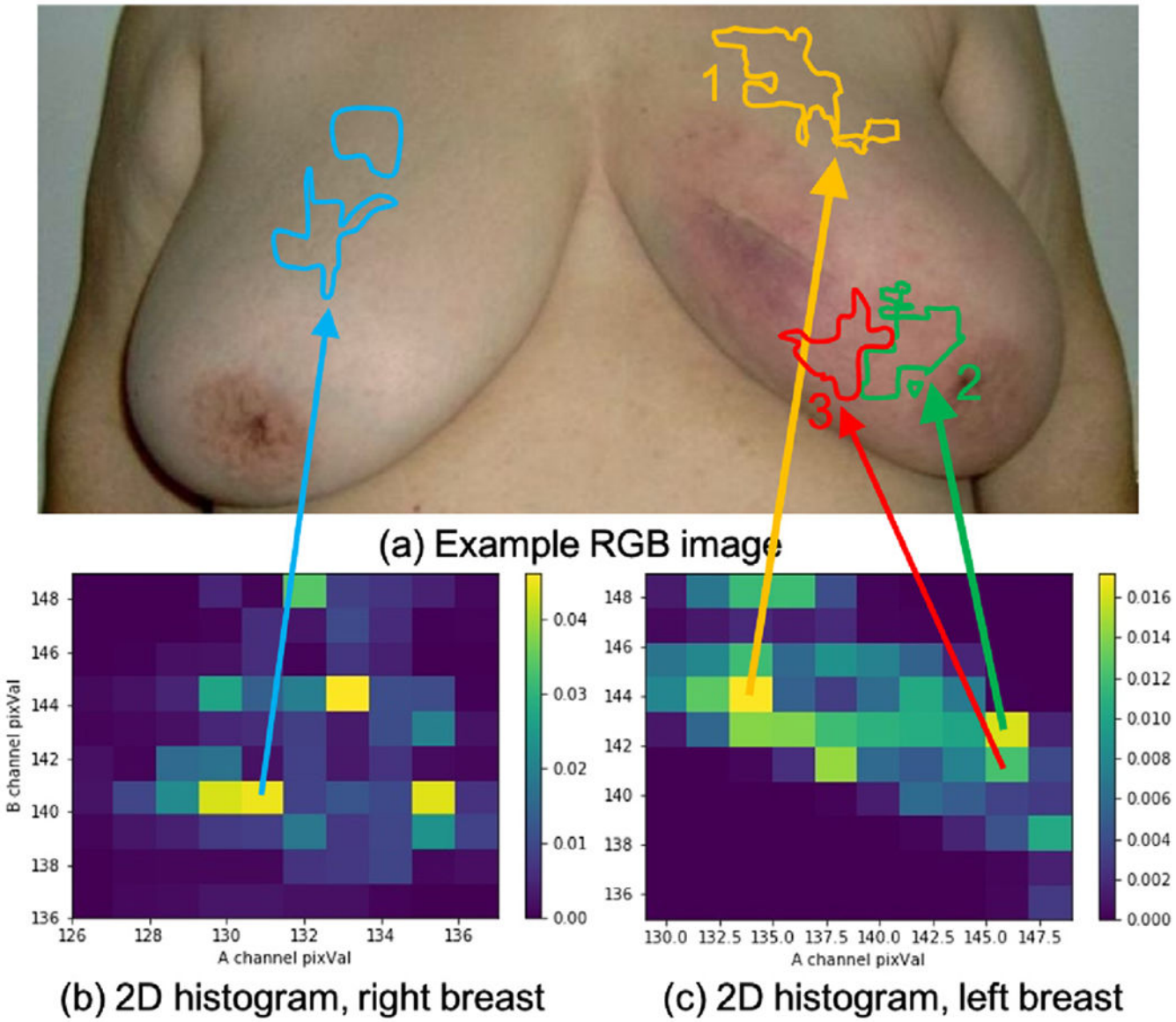


**Fig. 7.**

Color distribution differences between treated and untreated breast of a patient. The left breast of this patient is treated and the right is untreated. (a) Patient's image in RGB color space. (b) 2D histogram in AB space for right breast. (c) 2D histogram in AB space for left breast. Comparing to the untreated breast, the pixels of treated breast are more linearly correlated between A and B channels. The color contour shows the breast pixels with corresponding color range in the histogram. Regions smaller than contour width are discarded for visual clarity. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

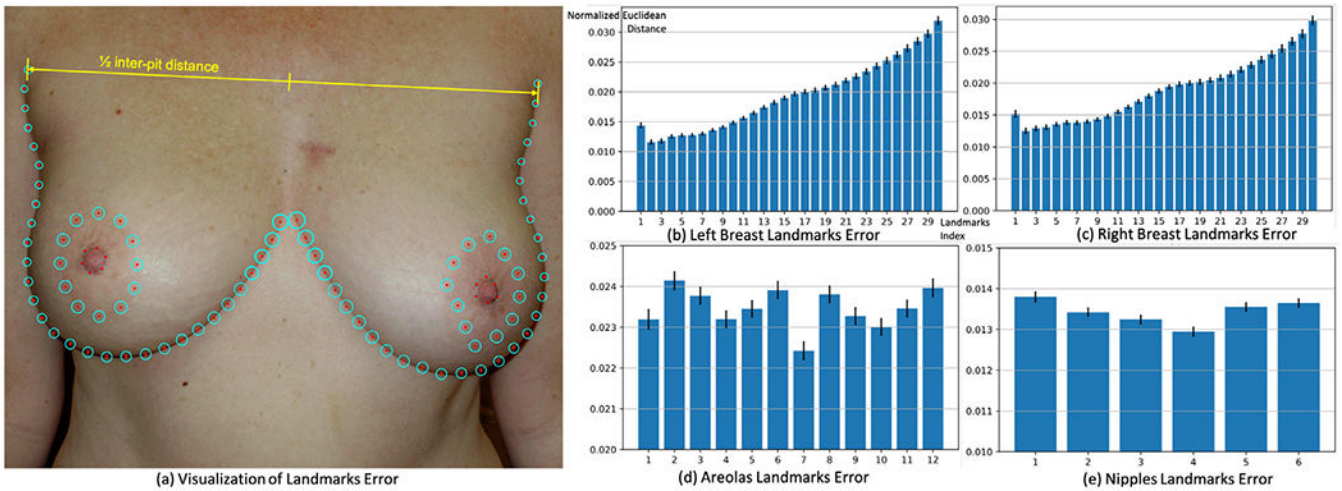


**Fig. 8.**

Quantitative results for landmarks detection using within landmark evaluation. (a) Visualization of the mean and standard error of the normalized Euclidean distance for each landmark across images. The radius of each circle represents the average Euclidean distance (across images) between the ground truth and the detected breast landmark after equal arc sampling. The thickness of each circle equals 2 times the standard error. The half-inter-pit distance is defined as half of the distance between the starting landmarks of left breast and right breast, and both are defined at the armpit of each side. (b) within landmarks result of detection error for left breast, (c) right breast, (d) areolas and (e) nipples.

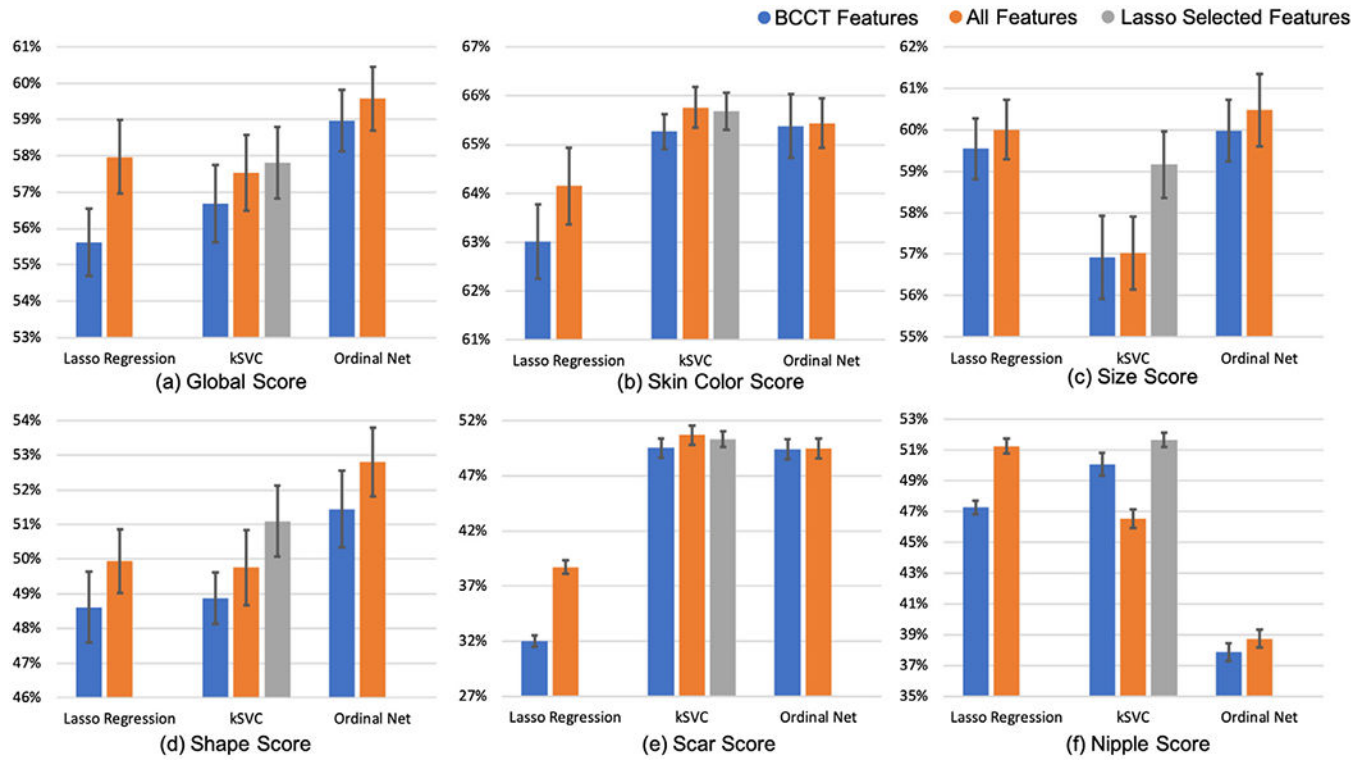


**Fig. 9.**

Bar plot of full pipeline performance with 10-fold Cross Validation, for each individual categorical score and global score. Bar heights represent average accuracy with error bar denoting standard error.

**Table 1**

The number of images from each clinical trial before and after selection, for both frontal and lateral view.

| Clinical trial | Before selection: frontal (lateral) | After selection: frontal (lateral) |
|---|---|---|
| NSABP B39/RTOG 0413 | 2172 (2343) | 1920 (2341) |
| RTOG 1014 | 94 (94) | 89 (94) |
| RTOG 1005 | 1760 (1884) | 1753 (1881) |
| Total number | 4026 (4321) | 3762 (4316) |

In this study, only the frontal view images after selection were graded during the grading sessions and used to train and test our algorithm.

**Table 2**

Quantitative results for breast components detectors with 10-fold and LOTO Cross Validation.

| Metric | Cross-validation | Left breast | Right breast | Areolas | Nipples |
|---|---|---|---|---|---|
| mAP | 10-fold | 1.00000 (0.00000) | 1.00000 (0.00000) | 0.90226 (0.01781) | 0.96201 (0.00610) |
| $F_1$ | 10-fold | 0.99987 (0.00013) | 1.00000 (0.00000) | 0.94331 (0.00958) | 0.97248 (0.00413) |
| mAP | LOTO (1005→0413)[a] | 0.99998 | 1.00000 | 0.86051 | 0.85105 |
| $F_1$ | LOTO (1005→0413) | 0.99922 | 0.99922 | 0.89133 | 0.86954 |
| mAP | LOTO (0413→1005) | 0.99943 | 0.99886 | 0.91321 | 0.90756 |
| $F_1$ | LOTO (0413→1005) | 0.99914 | 0.99914 | 0.93671 | 0.92934 |

[a]LOTO (1005→0413): Leave-one-trial-out, trial 1005 for training and trial 0413 for testing.

$p < 0.002$.

**Table 3**

Within-image results for landmarks detectors with 10-fold and LOTO Cross Validation.

| Metric | Cross-validation | Left breast | Right breast | Areolas | Nipples |
|---|---|---|---|---|---|
| Normalized $l_2$ [a] | 10-fold [b] | 0.0194 (9.697e–4) | 0.0190 (1.003e–3) | 0.0235 (6.224e–4) | 0.0134 (3.183e–4) |
| Normalized $l_2$ | LOTO (1005→0413) [c] | 0.0474 (7.497e–4) | 0.0481 (7.691e–4) | 0.0385 (5.889e–4) | 0.0216 (5.341e–4) |
| Normalized $l_2$ | LOTO (0413→1005) | 0.0517 (9.316e–4) | 0.0518 (9.691e–4) | 0.0350 (4.413e–4) | 0.0223 (4.394e–4) |
| Accuracy [d] | 10-fold | 92.98% | 93.25% | 92.10% | 99.80% |
| Accuracy | LOTO (1005→0413) | 65.02% | 64.76% | 76.74% | 97.02% |
| Accuracy | LOTO (0413→1005) | 62.68% | 62.46% | 79.80% | 95.88% |

[a] Normalized $l_2$ is the normalized Euclidean Distance defined in Section 6.3.1.

[b] In 10-fold cross-validation, the mean and s.e reported are averaged across validation folds.

[c] LOTO (1005→0413): Leave-one-trial-out, trial 1005 for training and trial 0413 for testing. The mean and s.e here are computed from trial 0413.

[d] Accuracy: computed with normalized $l_2$ threshold of $\eta < 0.05$.

$p < 0.002$.

**Table 4**

The number of selected features for each score.

| Total | Skin color | Size | Shape | Scar | Nipple | Global |
|---|---|---|---|---|---|---|
| 298 | 67 | 40 | 50 | 79 | 53 | 61 |

**Table 5**

Quantitative results using full pipeline with LOTO cross-validation.

| training→testing | ML methods | Features | Global | Skin color | Size | Shape | Scar | Nipple |
|---|---|---|---|---|---|---|---|---|
| 1005→0413 | Lasso | BCCT | 0.569 | 0.615 | 0.560 | 0.473 | 0.280 | 0.475 |
| 1005→0413 | Lasso | Full[a] | 0.573 | 0.616 | 0.562 | 0.484 | 0.342 | 0.489 |
| 1005→0413 | kSVC | BCCT | 0.544 | 0.654 | 0.551 | 0.479 | 0.518 | 0.462 |
| 1005→0413 | kSVC | Full | 0.543 | 0.638 | 0.531 | 0.467 | 0.519 | 0.455 |
| 1005→0413 | kSVC | Selected[b] | 0.557 | 0.643 | 0.557 | 0.491 | 0.512 | 0.470 |
| 1005→0413 | OrdinalNet | BCCT | 0.569 | 0.659 | 0.578 | 0.501 | 0.515 | 0.473 |
| 1005→0413 | OrdinalNet | Full | 0.577 | 0.648 | 0.571 | 0.507 | 0.518 | 0.389 |
| 0413→1005 | Lasso | BCCT | 0.551 | 0.625 | 0.605 | 0.489 | 0.355 | 0.459 |
| 0413→1005 | Lasso | Full | 0.564 | 0.633 | 0.615 | 0.503 | 0.421 | 0.505 |
| 0413→1005 | kSVC | BCCT | 0.547 | 0.593 | 0.565 | 0.495 | 0.484 | 0.478 |
| 0413→1005 | kSVC | Full | 0.526 | 0.599 | 0.551 | 0.477 | 0.486 | 0.467 |
| 0413→1005 | kSVC | Selected | 0.536 | 0.609 | 0.580 | 0.495 | 0.485 | 0.478 |
| 0413→1005 | OrdinalNet | BCCT | 0.566 | 0.609 | 0.605 | 0.529 | 0.482 | 0.508 |
| 0413→1005 | OrdinalNet | Full | 0.566 | 0.615 | 0.619 | 0.535 | 0.484 | 0.410 |

All results are reported in accuracy on testing trial.

[a]Full feature set combining BCCT features and proposed additional features.

[b]Lasso selected features from the full features.

**Table 6**

Ablation results with OrdinalNet and 10-fold CV.

| Score | Full Auto[a] | | GT landmarks[b] | | GT treated[c] | |
|---|---|---|---|---|---|---|
| | BCCT | Full | BCCT | Full | BCCT | Full |
| Global | 0.589 (0.008) | 0.595 (0.008) | 0.583 (0.008) | 0.599 (0.006) | 0.589 (0.008) | 0.601 (0.009) |
| Color[d] | 0.653 (0.006) | 0.654 (0.005) | 0.652 (0.009) | 0.655 (0.006) | 0.653 (0.006) | 0.654 (0.007) |
| Size | 0.599 (0.007) | 0.604 (0.008) | 0.602 (0.006) | 0.605 (0.007) | 0.599 (0.007) | 0.605 (0.009) |
| Shape | 0.514 (0.011) | 0.528 (0.010) | 0.517 (0.007) | 0.516 (0.005) | 0.514 (0.011) | 0.525 (0.010) |
| Scar | 0.494 (0.008) | 0.494 (0.009) | 0.493 (0.005) | 0.493 (0.005) | 0.494 (0.008) | 0.494 (0.008) |
| Nipple | 0.378 (0.005) | 0.387 (0.005) | 0.379 (0.007) | 0.387 (0.006) | 0.378 (0.005) | 0.387 (0.005) |

All results are reported in mean accuracy (and its s.e.).

Using BCCT feature, Full Auto and GT treated results are the same since BCCT feature does not involve left/right breast feature concatenation.

[a] Fully automatic pipeline.

[b] Replacing detected landmarks with ground-truth.

[c] Adding additional ground-truth treated/untreated information.

[d] Shorted for skin color.

**Table 7**

Running time and memory usage for each inference task.

| Task | GPU memory usage | CPU memory usage | Running time |
|---|---|---|---|
| Breast, areola and nipple detection[a] | 1386 MB | 18.83 GB | 3 m18.72 s |
| Breast, areola and nipple landmark detection[a] | 0 | 2.51 GB | 22.59 s |
| Feature extraction[b] | 0 | 75.30 GB | 5 h2 m14.28 s |
| Machine prediction: Lasso regression[c] | 0 | 2.51 GB | 3 m38.20 s |
| Machine prediction: kSVC[c] | 0 | 2.51 GB | 6 m24.31s |
| Machine prediction: OrdinalNet[c] | 0 | 7.53 GB | 1 h10 min24.03 |

[a]Experiment on a testing set with 376 images.

[b]Experiment on all 3 clinical trials of 3762 images.

[c]Experiment with 10-fold cross-validation for all breast cosmesis scores on all 3 clinical trials of 3762 images.