The NYU Breast Ultrasound Dataset v1.0

Farah Shamout[a,*], **Yiqiu Shen**[c], **Jan Witowski**[d,b], **Jamie Oliver**[d], **Kawshik Kannan**[f], **Nan Wu**[c], **Jungkyu Park**[d,g], **Beatriu Reig**[d,e], **Linda Moy**[d,b,e,g], **Laura Heacock**[d,e], **and Krzysztof J. Geras**[d,b,c,g,*]

[a]Engineering Division, NYU Abu Dhabi; [b]Center for Advanced Imaging Innovation and Research, NYU Langone Health; [c]Center for Data Science, New York University; [d]Department of Radiology, NYU Langone Health; [e]Perlmutter Cancer Center, NYU Langone Health; [f]Courant Institute of Mathematical Sciences, New York University; [g]Vilcek Institute of Graduate Biomedical Sciences, NYU Grossman School of Medicine

This manuscript was compiled on April 28, 2021

Ultrasound imaging is often used to diagnose breast cancer, especially in young women, women with palpable masses, and women with dense breast tissue. In this report, we present the NYU Breast Ultrasound Dataset consisting of 288,767 breast ultrasound exams with 5,442,907 total images acquired from 143,203 patients examined between 2012 and 2019 at NYU Langone Health. We summarize the statistics of the dataset, image collection process, and image preprocessing procedures. This dataset is intended to be used for the development of deep neural networks for the detection of breast cancer. Although this is a private dataset, we are publishing this report to improve reproducibility of our work and to share practices and insights that may be useful to others.

1. Statistics of the dataset

The NYU Breast Ultrasound Dataset v1.0 consists of 288,767 breast ultrasound exams (containing 5,442,907 images) from 143,203 patients examined between 2012 and 2019 in the NYU Langone Health system (New York, NY, USA). It was obtained with approval from the NYU Langone Institutional Review Board. Consecutive breast ultrasound examinations, performed in a screening setting (asymptomatic patient) or in a diagnostic setting (symptomatic patient, follow-up examination of a previously identified ultrasound finding, patient had an abnormal screening ultrasound, or patient had an abnormality detected on other breast imaging examinations e.g. mammography or magnetic resonance imaging (MRI)) were included. All exams were performed using hand-held ultrasound technique by sonographic technologists who specialize in breast imaging. Each exam contains images of one or both breasts. The images were originally stored using the Digital Imaging and Communications in Medicine (DICOM) Standard. We pre-processed each image and applied the filtering criteria, as described in Section 2. At the time of examination, the patients' ages ranged between 16 and 102 years, with a mean age of 55 (see Figure 1). The number of images per exam ranged between 4 and 70 images, with 18.8 images per exam on average (see Figure 2). The average size of a processed image is 665 pixels in width and 603 pixels in height. In Table 1, we show the average width and height of the images grouped by the year of examination. In the same table, we also show the three most frequently used ultrasonic transducers in each year of the study. A few examples of the final cropped ultrasound images are shown in Figure 3.

To develop and evaluate deep neural networks, the dataset was further split into training, validation, and test sets, as described in the following Section 1.A. The dataset also includes two breast-level binary labels for each breast: one indicating whether the imaged breast has at least one malignant finding and the other one indicating whether the breast has at least one benign finding, which will be further described in Section 1.B. The labels were extracted from pathology reports, as described in Section 3. The labels were assigned by matching each exam to pathology reports recorded either within 120 days after or 30 days before the date of examination. The natural language processing pipeline used for extracting the labels is described in our previous mammography data report, The NYU Breast Cancer Screening Dataset (1). In the test set, the cancer labels were further refined to reduce label noise. Further details on the filtering of the test set will be presented in Section 2.E. Each exam is also associated with a Breast Imaging-Reporting and Data System (BI-RADS) risk assessment and mammographic breast density labels as summarized in Section 1.C. The exams were performed by a variety of ultrasound machines and sonographic transducers from multiple vendors, as detailed in Section 1.D.

A. Training, validation and test sets. To split the dataset, the ultrasound exams were first grouped according to their patient identifier. Next, we randomly divided the patients into disjoint training (60% of the patients), validation (10% of the patients) and test (30% of the patients) sets. Hence, if a patient had multiple ultrasound exams over the study period, then the exams were all included in the same dataset (i.e., training, validation, or test set). The test set was further filtered to refine the cancer labels (see Section 2.E). The overall procedure resulted with 209,162, 34,850 and 44,755 exams in the training, validation and test sets, respectively, corresponding to 3,930,347, 653,924, 858,636 images, respectively. The training set contained exams collected between October 23, 2012, and September 30, 2019, the validation set contained exams collected between January 6, 2012, and September 30, 2019, and the test set contained exams collected between September 10, 2012, and September 30, 2019. The distributions of the patients' age in each subset are shown in Figure 1.

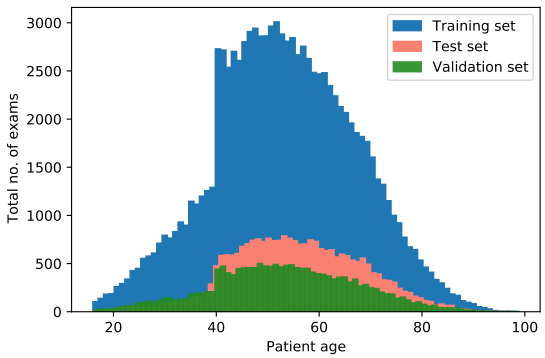


Fig. 1. Distribution of patients' age at the time of examination in the training, validation and test sets.

*To whom correspondence should be addressed. E-mails: k.j.geras@nyu.edu and farah.shamout@nyu.edu.

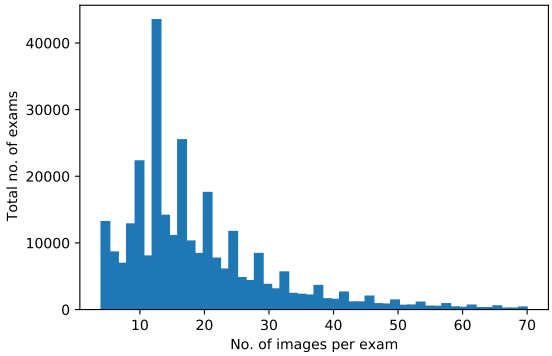
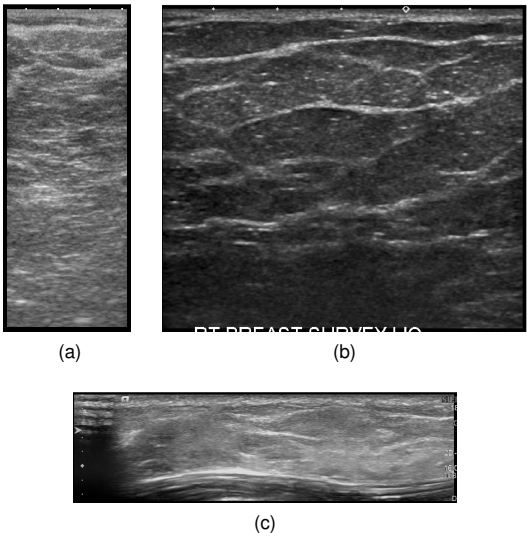


Fig. 2. Distribution of the number of images per exam in the overall dataset.

Table 1. The average image width and height after image cropping, grouped based on the year the ultrasound exam was conducted. The image resolution generally increased over time, reflecting developments in ultrasound technology.

Year	Total	Width	Height	Top 3 Devices
2012	4452	613	492	S2000 (83%), LOGIQ5 (15%), LOGIQ9 (2%)
2013	79149	566	468	S2000 (67%), LOGIQ7 (18%), LOGIQ5 (13%)
2014	222723	516	457	LOGIQ7 (36%), S2000 (27%), Xario (11%)
2015	349139	553	498	LOGIQ7 (22%), S2000 (16%), S3000 (14%)
2016	1060810	633	587	S1000 (42%), S3000 (14%), Affiniti 70G (8%)
2017	1300776	684	628	Affiniti 70G (56%), S1000 (20%), S3000 (13%)
2018	1422517	697	630	Affiniti 70G (57%), S1000 (16%), S3000 (14%)
2019	1003341	710	627	Affiniti 70G (56%), S3000 (15%), S1000 (13%)



(a) (b)

(c)

Fig. 3. Examples of images after applying the cropping procedure. We show three examples with different width to height ratios: (a) example of the minimum ratio present in the dataset, (b) example of the mean ratio present in the dataset, and (c) example of the maximum ratio present in the dataset.

B. Breast-level cancer labels. Here, we describe the distribution of the breast-level cancer labels. In total, the dataset contained images of 510,271 breasts (255,551 left breasts and 254,720 right breasts). There were 28,914 ultrasound exams (10% of total exams) associated with at least one biopsy performed within 120 days after or 30 days before the date of the ultrasound examination. Among these, there were 5,593 breasts with at least one biopsy-confirmed malignant finding and 26,843 breasts with at least one biopsy-confirmed benign finding. Additionally, 2,171 breasts were associated with both biopsy-confirmed malignant and benign findings. The remaining breasts were not matched with any pathology reports. These exams were assigned a 'negative' cancer label. Figure 4 shows the distribution of the time between the study date of the ultrasound exam and the latest pathology results across the entire dataset. Table 2 shows the distribution of malignant and benign findings across left and right breasts in the training, validation, and test sets.

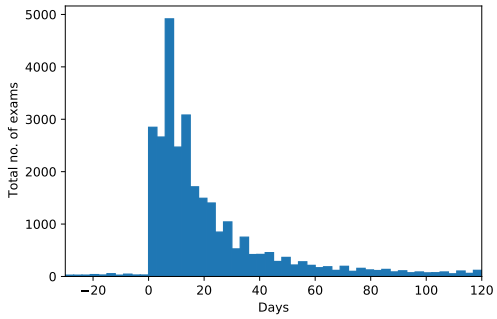


Fig. 4. Distribution of the number of days between ultrasound study date and pathology results. For ultrasound exams associated with multiple pathology reports, we selected the date of the latest report.

Table 2. Number of breasts with malignant and benign findings based on the labels extracted from the pathology reports, across the left and right breasts in the training, validation, and test sets.

	malignant		benign	
	right	left	right	left
training	1,794	1,867	8,116	8,159
validation	310	298	1,276	1,413
test	691	633	4,065	3,814
overall	2,795	2,798	13,457	13,386

C. BI-RADS risk assessment & mammographic breast density labels. We also extracted the associated BI-RADS risk assessment and mammographic breast density labels for each ultrasound exam from their ultrasound report and existing mammography reports, respectively. BI-RADS risk assessment labels are assigned by radiologists to indicate their suspicions of malignancy while reporting breast mammographic, ultrasound and MRI findings (2)*. Mammography reports should also include a visual assessment of the breast density, ranging from entirely fatty to extremely dense breasts. Therefore, we extracted breast density information for patients in our dataset who had undergone a screening or diagnostic mammogram in the past and had a matching radiology report available. Further details on the extraction of the BI-RADS risk assessment labels and the mammographic density labels are described in Section 3.B and Section 3.C, respectively. The distributions of the BI-RADS risk assessment labels and the mammographic breast densities for the exams in our dataset are summarized in Table 3 and Table 4, respectively.

*BI-RADS 0: incomplete exam and needs additional imaging evaluation, BI-RADS 1: negative (normal exam), BI-RADS 2: benign (normal exam with a benign finding) 0% probability of malignancy, BI-RADS 3: probably benign, <2% probability of malignancy, short interval follow-up suggested, BI-RADS 4: suspicious for malignancy, >2-95% probability of malignancy, biopsy should be considered, BI-RADS 5: highly suggestive of malignancy, >95% probability of malignancy, appropriate action should be taken, and BI-RADS 6: known biopsy-proven malignancy.

Table 3. Breakdown of BI-RADS risk assessment labels assigned to each exam. BI-RADS risk assessment labels were extracted from the patient's ultrasound report. 'Unknown' indicates exams with missing or ambiguous information.

BI-RADS risk assessment	Overall	Training set	Validation set	Test set
0	14078 (4.9%)	11151 (5.3%)	1835 (5.3%)	1092 (2.4%)
1	86347 (29.9%)	63499 (30.4%)	10474 (30.1%)	12374 (27.6%)
2	136322 (47.2%)	98044 (46.9%)	16603 (47.6%)	21675 (48.4%)
3	27711 (9.6%)	20722 (9.9%)	3403 (9.8%)	3586 (8.0%)
4	22133 (7.7%)	14266 (6.8%)	2289 (6.6%)	5578 (12.5%)
5	1348 (0.5%)	865 (0.4%)	145 (0.4%)	338 (0.8%)
6	518 (0.2%)	393 (0.2%)	56 (0.2%)	69 (0.2%)
Unknown	310 (0.1%)	222 (0.1%)	45 (0.1%)	43 (0.1%)

Table 4. Breakdown of mammographic breast density labels assigned to each exam. The density labels were extracted from screening and diagnostic mammogram reports associated with the same patient. 'Unknown' indicates exams with missing or ambiguous information.

Mammographic breast density	Overall	Training set	Validation set	Test set
A (breasts are almost entirely fatty)	5384 (1.9%)	4107 (2.0%)	582 (1.7%)	695 (1.6%)
B (scattered areas of fibroglandular density)	69948 (24.2%)	50609 (24.2%)	8291 (23.8%)	11048 (24.7%)
C (breasts are heterogeneously dense)	165855 (57.4%)	119417 (57.1%)	19929 (57.2%)	26509 (59.2%)
D (the breasts are extremely dense)	31829 (11.0%)	22660 (10.8%)	3980 (11.4%)	5189 (11.6%)
Unknown density	15751 (5.5%)	12369 (5.9%)	2068 (5.9%)	1314 (2.9%)

D. Scanner information. We also extracted information on the ultrasound system used to acquire each image using the `ManufacturerModelName` attribute in the DICOM file. The images in the dataset were collected from 8 manufacturers, including Philips, General Electric (GE), Siemens, Toshiba, Medison, Advanced Technology Laboratories, Supersonic Imagine, and Samsung, using 20 different types of ultrasound transducers, as shown in Table 5. The most commonly used ultrasound machines were the Affiniti 70G (Philips), S1000 (Siemens), S3000 (Siemens), and S2000 (Siemens). This highlights the diversity of the dataset in terms of the acquisition devices.

Table 5. Distribution of types of ultrasound machines used to collect the exams in the training, validation and test sets.

Device	Training set	Validation set	Test set
Affiniti 70G	79080	13329	14715
S1000	40097	6684	9148
S3000	29676	4937	5785
S2000	24701	4054	5655
LOGIQ7	6316	1035	1647
Xario	6029	947	1541
iU22	4988	803	1696
LOGIQ9	3659	585	544
TUS-A300	3540	618	954
Accuvix V10	2478	389	788
Antares	2468	395	709
LOGIQ5	2232	471	832
Sequoia	1868	263	28
Accuvix V20	1851	311	680
LOGIQE9	152	19	26
HDI 5000	10	6	1
LOGIQS8	8	1	5
Aixplorer	4	1	0
LOGIQS7	4	2	1
UGEO H60	1	0	0

2. Image collection and preprocessing

In this section, we describe in detail the complete pipeline used for processing the ultrasound images, starting from extracting the images from the DICOM files. This pipeline consists of six phases: image collection and extraction, image cropping, breast laterality extraction using optical character recognition (OCR), filtering of the overall dataset based on the inclusion and exclusion criteria, filtering of the test set, and removal of burnt-in annotations.

A. Image collection and extraction. We extracted certain metadata fields from all DICOM files, where each file represented a single ultrasound image and each exam contained several ultrasound DICOM files. The extracted metadata included (i) exam and patient identifying information (`PatientID`, `AccessionNumber`, & `StudyDate`) (ii) patient demographics (`PatientBirthDate` & `PatientSex`), (iii) characteristics of the image and acquisition process (`Modality`, `InstanceCreationTime`, `SOPInstanceUID` which represents a unique image identiifer, number of `Rows`, number of `Columns`, `ImageType`, `PhotometricInterpretation`, `ManufacturerModelName`, & `Manufacturer`) and (iv) type of procedure (`PerformedProcedureStepDescription`, `RequestedProcedureDescription`, & `StudyDescription`). This metadata was used to filter the exams as described in Section 2.D. Before filtering and pre-processing the data, the data was anonymized by replacing the patients' identifiers and names with randomly generated identifiers.

B. Image cropping. The pixel values of each image were extracted from the `PixelData` attribute in the DICOM file. Each image contained a margin of textual metadata surrounding the ultrasound picture of the breast. Unaltered examples of these ultrasound images are provided in Figure 7(a). The background surrounding the breast ultrasound image was approximately zero-valued, except for the burnt-in metadata. The number of metadata strings and their location varied across exams and different acquisition devices. Additionally, depending on the acquisition probe or transducer that was used by the ultrasound technologists to optimize image quality, some images were in a rectangular, trapezoidal, or convex shape, as shown in Figure 5. Thus, the images required designing a careful cropping procedure to obtain the picture of the breast and discard the surrounding margins. This procedure consisted of two parts: (i) binary erosion and dilation to obtain the largest nonzero connected component and (ii) heuristic assessment of the pixel values at the boundaries of the cropped image to perform any additional cropping. Further details are provided below.

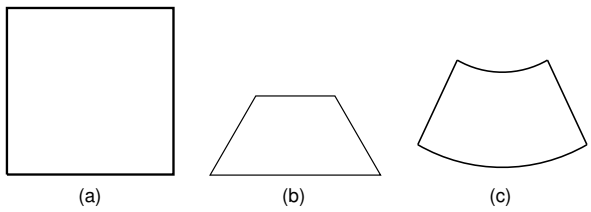


Fig. 5. Examples of image shapes provided by ultrasound transducers: (a) 2D rectangular, (b) 2D trapezoidal, or (c) 2D convex (curved). Our pre-processing procedure obtains a rectangular crop of 2D trapezoidal and 2D convex images.

B.1. Erosion and dilation. To obtain the largest nonzero connected component, all images were first converted to grayscale since some images were not in grayscale. Such images had the attribute `PhotometricInterpretation` set to RGB or YBR. Next, we obtained the nonzero mask of the image, which is

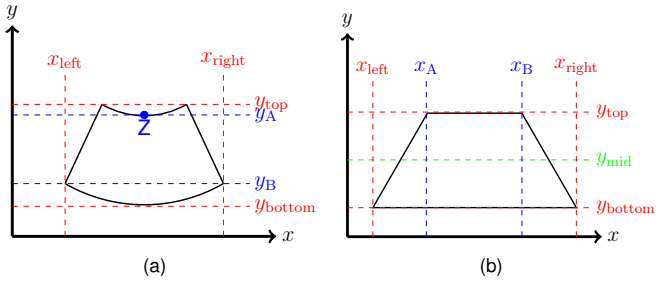


Fig. 6. Further cropping of non-rectangular image shapes. The red lines show the boundaries of the largest mask after erosion and dilation. (a) 2D convex image shape cropped as trapezoid with the new blue boundaries y_{A} and y_{B}. (b) 2D trapezoid image shape further cropped as rectangular with new blue boundaries x_{A} and x_{B}.

simply a binary mask of nonzero pixels. For devices where the background pixel value was not exactly zero, we obtained the binary mask by thresholding based on the most common pixel value in the image (i.e., the mode of all pixel values). Binary erosion and dilation (3, 4) were then applied to the binary mask of the ultrasound image for a number of iterations (`numIter`). In brief, erosion of image A by structural element B is defined as:

$$A \ominus B = \{z \in E | B_z \subseteq A\},$$

where E denotes Euclidean space and B_z denotes the translation of B by the vector z. Dilation of image A by structural element B is defined as

$$A \oplus B = \{z \in E | (B^s)_z \cap A \neq \emptyset\},$$

where $B^s = \{x \in E | -x \in B\}$. The structural element B is defined as:

$$B = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix},$$

where only the pixels directly connected to the center are considered as neighbors. Therefore, one iteration of erosion shrinks nonzero shapes in an image by one pixel, and dilation expands nonzero shapes in an image by one pixel. We applied erosion to the nonzero binary mask, obtained the largest connected component in the image, then discarded all surrounding artifacts that were not connected to the image. Next, we applied dilation to the largest connected component and used that as the initial cropping window location w. This procedure allowed us to eliminate any artifacts surrounding the breast ultrasound image.

B.2. Additional cropping based on image shape. Images in the dataset were acquired using a variety of ultrasound transducers, as shown in Table 5. Each device has its own physical dimensions and scanning properties, providing a range of image resolutions and shapes (5). The shapes that were common in our dataset are shown in Figure 5. The easiest image shape to crop was the 2D rectangular shape (Figure 5(a)) where no further cropping steps were required after the erosion and dilation process. Those images are obtained by the transducer through a linear scanning in the xz plane. On the other hand, images in the 2D trapezoidal shape (Figure 5(b)) are obtained through a shift in the angle of the acoustic beam in an arc in the xz plane, while the images in the 2D convex shape (Figure 5(c)) are obtained along a curved surface rather than the straight xz plane (5). Here, we describe the additional steps taken to (i) detect 2D convex images, (ii) crop 2D convex images as 2D trapezoidal convex images, (iii) detect trapezoidal images, and (iv) crop the trapezoidal images as 2D rectangular images.

The largest mask of each image consisted of four edges after erosion and dilation: bottom (y_{bottom}) and top (y_{top}) edges identified by point coordinates on the y-axis, and right (x_{right}) and left (x_{left}) edges identified by point coordinates on the x-axis, as shown in Figure 6. Assume m_x is the midline between the vertical x_{right} and x_{left} edges. To detect 2D convex images, we retrieved the coordinate of the first occurring non-zero pixel along m_x, which is point Z in Figure 6(a), yielding the horizontal line y_{A}. If y_{A} was more than 20 pixels away from y_{top}, then y_{top} was readjusted to y_{A} since this means that the original top edge (y_{top}) had zero pixels in the middle (i.e., convex). The '20 pixels buffer' was incorporated to ensure that there is a significant gap between y_{top} and y_{A}. After adjusting the top edge of the 2D convex image, the proportion of nonzero pixels were calculated along the y-axis (i.e., per row). We defined y_{B} as the row containing the highest proportion of nonzero pixel values. Then, y_{bottom} was readjusted to y_{B}. Therefore, y_{A} and y_{B} defined the new top and bottom boundaries. This procedure detects and further crops 2D convex images to 2D trapezoids by adjusting the top and bottom edges.

To detect trapezoidal images based on pixel values, we compared the proportion of zero pixels at the top edge of the image (i.e., y_{top}) with the proportion of zero pixels along the horizontal midline of the image (i.e., y_{mid} as shown in Figure 6(b)). We adjusted the right (x_{right}) and left (x_{left}) edges of an image when the proportion of zero pixels at its top edge was at least 3 times the proportion of zero pixels at y_{mid}. In particular, x_{left} was shifted to x_{A} and x_{right} was shifted to x_{B}. Each vertical edge was shifted by half the total number of zero pixels at y_{top}. Finally, we added `bufferSize` pixels as buffer to all edges to ensure that our cropping window was not missing anything. Examples of the images produced with the overall cropping procedure for all image shapes are shown in Figure 7.

B.3. Experimental observations. Prior to dilation and erosion, we removed all non-zero headers (i.e. rectangular banner located at the top of the image; 25-80 pixels in length depending on the device), since some headers were directly connected to the ultrasound image and hence were included as part of the largest connected component after erosion and dilation. In particular, images acquired by Siemens Acuson S1000 (Siemens), S2000 (Siemens), S3000 (Siemens), TUS-A300 (Toshiba), Antares (Siemens), iU22 (Philips), LOGIQ5 (GE), LOGIQ7 (GE), LOGIQ9 (GE), Affiniti 70G (Philips), Xario (Toshiba), LOGIQE9 (GE), and Accuvix V10 (Medison), required the removal of a non-zero header that is 56 pixels in height.

During preliminary experiments, we determined that `numIter` $= 5$ for erosion and dilation was sufficient for most cases. However, due to the variety of transducers and variable image quality, we included sanity checks after cropping to assess whether the new image dimensions were reasonable. For example, some images were largely dark in the bottom or in the center due to acoustic shadow, leading to incorrectly cropped images. Therefore, we determined heuristic procedures to identify and fix such cases. First, some images contained large patches of zero pixels in the upper half of the

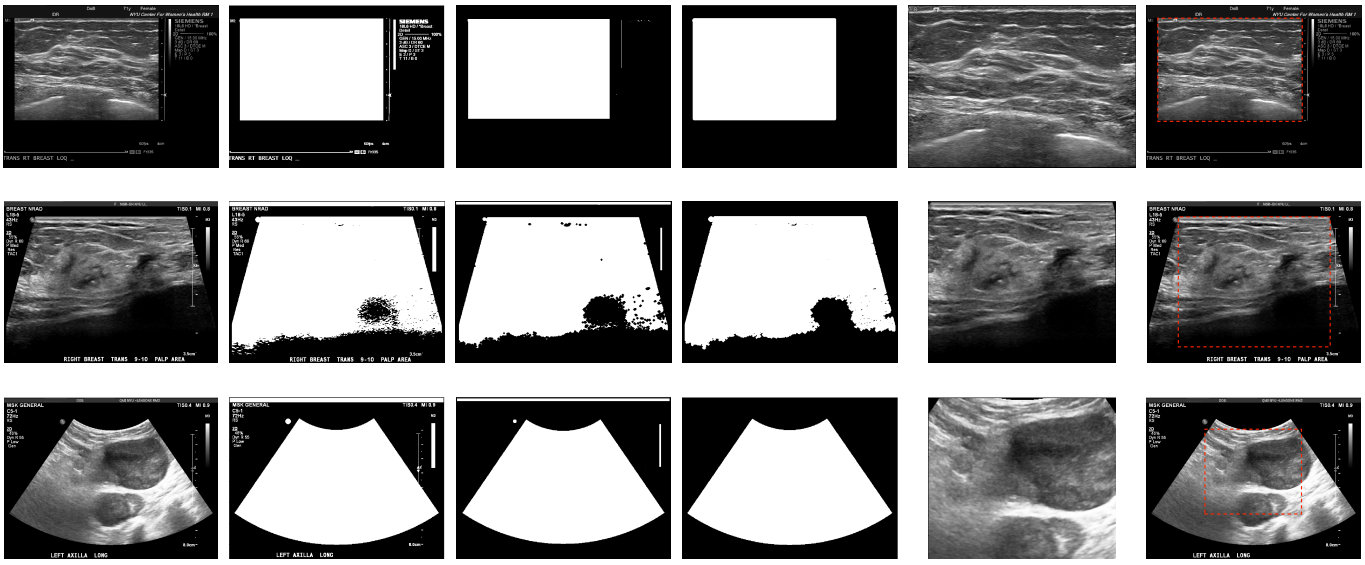


Fig. 7. Examples of the cropping procedure applied to breast ultrasound images with a 2D rectangular shape in the first row, a 2D trapezoidal image shape in the second row, and a 2D convex shape in the third row. (a) An image with surrounding metadata before cropping. (b) Binary image mask obtained by assigning ones to pixels with values greater than 0 and leaving zeroes unchanged. (c) Binary image mask after erosion. (d) Largest nonzero connected component selected after dilation. (e) Final cropped image. (f) The image before cropping with position of the final cropped image indicated by the red rectangle.

image due to acoustic shadowing, resulting in a binary mask similar to the one in Figure 8(h). Applying 5 iterations of erosion to such masks results with the selection of the area under the zero-valued patch as the largest nonzero connected component. This incorrectly discards the zero-valued patch and the area above it, which are supposed to be part of the breast image. To fix those cases, we first checked if the new top edge was more than 200 pixels away from the original top edge (which means the crop was too large). If so, we re-applied the erosion procedure to the original image with `numIter` $= 2$ iterations only.

Second, there were images that had a high proportion of zero pixels in the bottom half of the breast image ($> 50\%$ of all pixels). To address these cases, we set limits for the extent to which images could be cropped vertically. We also assessed the ratio of image width and height. If the width was more than twice the height, then we ignored any adjustments done to y_{top} and y_{bottom} by the convex image detection step. This mainly affected images that had many zero pixels around the top center of the image but were not necessarily convex. Alternatively, if the width was less than half of the height, then we ignored any adjustments done to x_{left} and x_{right} by the trapezoid detection step. Finally, `bufferSize` was set to 5. The full algorithm is summarized in Algorithm 1 to obtain the final cropping window w and the image of the breast cropped by w.

C. Breast laterality extraction. Each ultrasound exam consisted of a series of images that belonged to only one or both breasts. Since the metadata of the DICOM images did not include any information to indicate which breast the image belongs to, we extracted the laterality from the burnt-in annotations in each image. This procedure consisted of a series of three steps: extracting all text within the image using OCR, processing the extracted text, and interpolating laterality for images with missing laterality.

C.1. OCR for text extraction. Burnt-in annotations containing information regarding image laterality were embedded in the bottom portion of each image. We applied the following steps to extract the relevant textual information from each image. First, we selected the bottom half of the image (500 pixels in height), in order to only process the region that includes the burnt-in text pertaining to image laterality and to reduce the amount of text processed by the OCR engine for efficiency purposes. Next, the cropped region was smoothed using median filtering with a kernel size of 3 (6), and thresholded at a value of 150, such that pixel intensities less than 150 were mapped to 0 and pixel intensities greater than or equal to 150 were mapped to 255. Finally, we applied the open-source Python-tesseract (7) OCR engine to detect all text embedded in the processed image.

C.2. Processing of extracted text. The text was first converted to lowercase. Then, we performed exact matching of words `left` and `right`, as well as common abbreviations such as `lt` and `rt`. If phrases for both left and right side were detected in the image, then we selected the one which appeared first. This is because the burnt-in text, on rare occasions, contained additional information regarding the orientation of the transducer and the location on the breast. For example, `lt sag right` designates the sagittal orientation, or longitudinal plane, on the right side of the left breast. Since the OCR procedure would have detected both `lt` and `right`, we chose `lt`, left, as the breast laterality since it appeared first in the string.

After detecting exact string matches, we performed partial string pattern matching, such that any image was matched with left if it contained patterns of `.eft`, `l.ft`, `le.t`, `lef.`, or `ft breast` and with right if it contained patterns of `.ight`, `r.ght`, `ri.ht`, `rig.t`, `righ.`, or `ght breast`, where the dot matches any character except a newline. This was done because, in some images, the laterality information overlapped with other burnt-in annotations in the image and was therefore

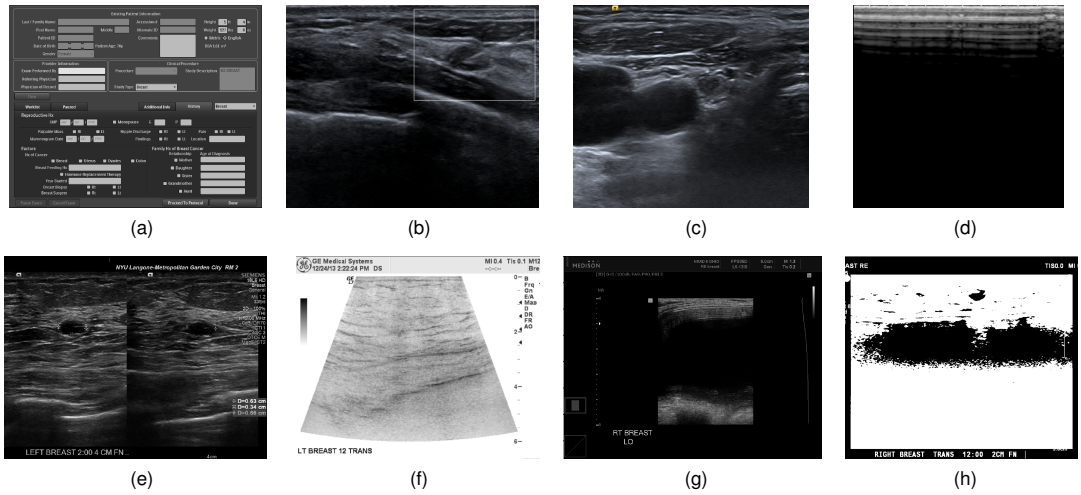


Fig. 8. Examples of images seen during data preprocessing. The first row (images a-d) show examples of images that were automatically rejected since they did not represent valid US images: (a) `ImageType` was 'INVALID', (b) `PerformedProcedureStepDescription` was 'US GUIDED FINE NEEDLE BREAST ASPIRATION', and the needle could be seen in the upper right corner, (c) `PerformedProcedureStepDescription` was 'US HEAD AND NECK', or (d) the fraction of non-zero pixels ratio was smaller than 20%. The second row (images e-h) shows examples of images that we encountered during image pre-processing: (e) Some images contained two adjacent US images and (f) others had to be inverted prior to further pre-processing like the one shown here. For some images, the majority of pixel values were zero or they contained a large zero-pixel path in the center, so either (g) the image was rejected if the ratio of the image width and height was extreme after preprocessing as in the case of the shown image, or (h) the number of erosion steps was decreased to avoid resulting with two disjoint masks.

Algorithm 1 Given image A, obtain the cropping window location w

procedure CROP_ULTRASOUND(`numIter`, `bufferSize`)
  Get mask b where $A > 0$ (or $A > mode(A)$ for certain devices, i.e., pixel value that appeared the most)
  Apply erosion for `numIter` iterations to get eroded mask e
  Get the nonzero largest connected component c of e
  Apply dilation c for `numIter` iterations to get a dilated mask d
  Select a window w_1 from A which contains d and set $w = w_1$
  if image A_1 cropped by w_1 is 2D convex **then**
    Select a new window w_2 from A_1, and set $w = w_2$
  if image A_2 cropped by w_2 is trapezoidal **then**
    Select a new window w_3 from A_2, and set $w = w_3$
  Assess image per sanity checks described in Section 2.B.3
  Include `bufferSize` pixels in all directions, record the final location of w, and save the image cropped by w

not detected accurately by the OCR engine. If there were no exact or partial matching of left and right, then the image laterality was denoted as missing.

In ultrasound examinations, the technologist generally examines the left breast followed by the right breast, or vice versa. Alternatively, the technologist may examine just one of the two breasts during the exam. To detect errors incurred by the OCR engine and the text pre-processing procedure, we sorted images within the same exam based on the `InstanceCreationTime` attribute in the image DICOM file, which indicates the date and time that the image was acquired. For any image with left laterality situated between two images with right laterality, or vice versa, the laterality of the respective image was set as missing because this was highly unlikely to occur in practice and was likely to be an OCR error. One limitation of this approach is that it ignores the rare cases where the OCR extraction fails or the technologist does indeed switch between the left and right breasts several times.

C.3. Interpolation technique for missing laterality. For each exam in which laterality was successfully extracted for at least 10% of its images, we further interpolated the remaining missing lateralities. First, for each image with a missing laterality, situated between two images with available lateralities (i.e., collected before or after the respective image), we interpolated the laterality of the image that is closest in time. Next, we identified the first image during the exam where the laterality switches from one laterality to another from a set of three possible values (`left`, `right`, `missing`). We denote those as 'transition' points. We also define a 'splitpoint' as the transition point when the technologist switches between left and right, or vice versa, and we assume that there is at most one splitpoint per exam. To interpolate the laterality within an exam, we first search for the first occuring transition point, which may or may not be the splitpoint. The transitions may indicate one of several scenarios:

1. **Identification of a `left->right` transition, i.e., the splitpoint**: The technologist switches to examining the right breast after examining the left breast. If such a scenario is detected, then we simply used a backward fill to interpolate the laterality of all images prior to the splitpoint as left, and a forward fill to interpolate the laterality of all images following the splitpoint as right.

The same applies if the technologist switches from right to left (`right->left`).

2. **Identification of a `left->missing` transition and the exam contains images with left and right laterality**: If such a scenario is detected, we first interpolate the laterality of the closest available image laterality for the missing value. Then, we repeat the interpolation procedure and detect a new transition point, until the splitpoint is identified. This same step applies for `right->missing`, `missing->left`, and `missing->right`.

3. **Identification of a `left->missing` transition and the exam only contains images with left laterality**: In this scenario, we simply fill all missing values with the available image laterality. The same applies for `right->missing`, `missing->left`, and `missing->right`, and only a single image laterality is available in the exam.

D. Filtering of overall dataset. The initial dataset originally contained 425,859 exams. Amongst those, we successfully extracted 425,506 ultrasound exams containing 8,448,978 images collected from 212,716 unique patients. The excluded 353 exams represented non-ultrasound imaging modalities (e.g., CT, MRI), which were excluded by examining whether the `Modality` attribute in the DICOM file was 'US', had corrupted metadata, or were ultrasound videos. The extracted dataset was then further filtered to discard images that were not within our clinically-defined inclusion criteria as described below.

1. Overall, there were 96 exams containing images associated with one of 76 non-integer patient IDs, which are invalid. By matching based on the anonymized patient name and birth date, we were able to fix 39 patient IDs. We then excluded the remaining images that had corrupt patient IDs. This resulted with the exclusion of 1,150 images and only 2 exams.
2. Before filtering, the dataset contained exams collected between 2008 and 2019. We excluded all exams collected between 2008 and 2011 (18 exams), and included data from 2012 onwards.
3. We excluded any exams collected from patients younger than 16 years of age. This corresponded to 703 excluded exams (10,697 images).
4. We discarded 307,433 images with duplicate `SOPInstanceUID`, as these indicated duplicates of the same image.
5. We discarded 2,994 exams with `PatientSex` different than 'F' (female).
6. We discarded 74,786 images that had invalid entries in the list `ImageType` attribute. Invalid image types included 'INVALID', 'REPORTDATA', 'DEMOGRAPHICDATA', '0000', '0009', and '0019', as those image types were not associated with breast ultrasound images. Example of an `INVALID` image is shown in Figure 8(a).
7. We excluded 364,296 images that were collected during biopsy procedures based on the values of `PerformedProcedureStepDescription`, `StudyDescription` & `RequestedProcedureDescription`, in that order. Performed procedure step description was prioritized because it contains information about what was actually performed during image acquisition (8). An example of a fine needle aspiration procedure is shown in Figure 8(b). Some of the excluded images were also acquired from other body parts, such as the thyroid gland as shown in Figure 8(c).
8. We further excluded 18,925 images with missing `PerformedProcedureStepDescription`, `RequestedProcedureDescription`, & `StudyDescription` attributes.
9. We excluded 3,568 images which had less than 20% nonzero pixels after cropping, as they usually were mostly empty images as shown in Figure 8(d).
10. We also excluded 1,174 exams that were associated with multiple patient IDs or study dates, as it is not possible for an exam to be performed across multiple dates and each exam must be associated with a single patient.
11. We further excluded 2,636 images whose height or width did not change after the cropping procedure. Images that were not cropped at all in both dimensions mainly consisted of invalid images. They were missed by the filtering procedure presented in step 6 because their `ImageType` attribute did not contain any of the mentioned codes. Other images that were excluded in this step consisted of two adjacent US images, as shown in Figure 8(e), that were only cropped in height but not in width.
12. One drawback of the erosion-dilation procedure is that US images may contain areas that are entirely black, also referred to as anechoic due to the absorption of sound waves. Hence, when the center of the US image is largely black, such as in Figure 8(g), the US image is wrongly split into two large masks, and only one of them is selected leading to a small crop. Although this was generally avoided for the majority of cases, we excluded any images that had an extreme width to height ratio (either greater than the 99th percentile or lower than the 1st percentile of the image ratio distribution.) This led to the exclusion of 81,441 images.
13. We also excluded 12,356 exams that had an extreme number of images. That is, if the total number of images was greater than 70 images (99th percentile of image number distribution) or lower than 3 images. Exams with a very small number of images were either a result of the image exclusion process described above or, if not, were typically highly targeted diagnostic ultrasound exams where only one region of one breast was imaged.
14. After applying the OCR pipeline to extract laterality and interpolate it wherever possible, we excluded 32,042 exams that had at least one image with missing laterality data.

The dataset was then split randomly into training (60%), validation (10%), and test (30%) sets based on patient IDs. To decrease the noise in the cancer labels, we applied additional filtering to the test set as described in the next section. After applying the aforementioned exclusion and inclusion criteria and filtering the test set, the final dataset included 288,767 exams containing 5,442,907 images, which were acquired from 143,203 patients. The training set contained 3,930,347 images collected in 209,162 exams from 101,493 patients. The validation set contained 653,924 images collected in 34,850 exams

from 16,707 patients. The test set contained 858,636 images collected in 44,755 exams from 25,003 patients.

E. Filtering the test set. We performed a rigorous filtering process on our test set to ensure that our benign and malignant labels associated with each ultrasound exam were correct. To confirm benign labels, we evaluated the medical records of patients to see if any follow-up imaging or pathology results contradicted the benign label. For exams associated with malignant labels, we filtered exams based on their BI-RADS risk assessment labels (radiologist's overall diagnostic impression) as well as the the imaging modality that was ultimately used to obtain the breast tissue biopsy (determined using biopsy reports). Since breast ultrasound produces 2D images and does not contain comprehensive images of the entire breast, a proportion of patients diagnosed with breast cancer did not actually have images of the cancer in their ultrasound exams. This analysis of BI-RADS risk assessment labels and biopsy reports was necessary to ensure that all exams associated with a malignant label did in fact have cancers that were visible. Because of these more rigorous restrictions, we excluded 341 test set exams that had missing ultrasound BI-RADS risk assessment labels and were not associated with any biopsy procedures, as these exams could not be processed using this filtering protocol. The extraction procedure of the BI-RADS risk assessment labels is described in Section 3.B.

First, we examined benign exams that did not undergo biopsy within -30 to +120 days of their ultrasound exam. To confirm that these exams were benign, we evaluated the patients' electronic medical record to determine what follow up imaging they received and whether any additional breast pathology was obtained in the 15 months following their ultrasound exam. Non-biopsied patients who had negative (BI-RADS 1) or benign (BI-RADS 2) ultrasound exams were only included in the test set if they did not have any malignant breast pathology found within 0-15 months following their ultrasound exam and had follow up imaging between 6 and 24 months that was also negative or benign (BI-RADS 1-2) as is shown in Figure 9(a). Patients who did not undergo biopsy and had probably benign ultrasound exams (BI-RADS 3) were included in the test set if they did not have any malignant breast pathology found within 0-15 months following their exam and met one of two additional criterion: they had at least one follow up ultrasound exam at 24-36 months (2 year follow-up) after the initial study which were all BI-RADS 1-3 or all of their follow up ultrasound exams in the 4-36 months following their initial ultrasound exam were BI-RADS 1-2. This step of the filtering procedure is outlined in Figure 9(b).

We also examined exams with biopsy-proven benign lesions in the test set. Specifically, benign biopsy reports were evaluated using textual analysis to determine if the pathology results were deemed by the radiologist to be concordant (the imaging findings are accounted for by the pathology results) or discordant with the imaging features of the breast lesion. Discordance usually occurs in the setting where the ultrasound features appear suspicious, but a biopsy yields benign breast pathology. In this setting, there is concern that there was inadequate biopsy sampling of the lesion. Therefore, a repeat biopsy or surgical excision where the entire lesion is removed is typically recommended. Patients with a benign biopsy report that mentioned a discordant finding were only included in the test set if they received a subsequent biopsy (that was not discordant) or breast surgery in the 6 months following the discordant biopsy. Patients with benign discordant biopsies that did not receive subsequent pathological evaluation were excluded (34 exams excluded from this criterion).

Patients with biopsy-proven cancer were also examined in the test set. Since breast ultrasound produces 2D images and ultrasound exams do not contain comprehensive images of the entire breast, a proportion of patients diagnosed with breast cancer did not have images of the cancer in any of their ultrasound exams. Ultrasound exams with a malignant label and a BI-RADS risk assessment label of 1-2 were excluded as these exams typically did not contain images of the cancer. Additionally, patients diagnosed with breast cancer who did not have any breast pathology obtained using US-guided biopsy were also excluded, since the majority of patients diagnosed using MRI and stereotactic-guided biopsies had malignancies that were sonographically occult. Ultrasound exams that received a BI-RADS risk assessment label of 0, 3, and 6, as well as patients who had breast pathology obtained using multimodal image guidance (ultrasound plus stereotactic and/or MRI guided biopsies) had their cases manually reviewed to confirm the breast cancer was visible on the ultrasound exam. Only patients who were given a BI-RADS risk assessment label of 4-5 and had all of their breast pathology obtained using US-guided biopsy were presumed to have visible cancers and were not manually reviewed. This component of the filtering process is outlined in Figure 9(c). In total, of the 1822 US exams associated with a malignant pathology report initially included in the test set, 595 (32.7%) were excluded due to this filtering procedure.

F. Removal of burnt-in annotations. A fraction of the images in this dataset contain burnt-in annotations, which are created by the technologist to highlight findings for the radiologist. These annotations may consist of asterisks and dots, typically placed by a technologist to measure the size of lesions, such as in Figure 10(a,b), as well as bounding boxes around lesions that are used in Doppler ultrasonography to determine the vascularity of breast tissue within a specific region of interest, as shown in Figure 10(c). In order to reduce the likelihood that our deep learning models trained using this dataset would learn to rely on these annotations, we designed a pipeline to remove them. This system removed all asterisks, measurements, and other text that a technologist might write on an image. However, the bounding boxes and color overlay from Doppler ultrasonography were not removed, as this information regarding lesion vascularity can be critical in distinguishing benign and malignant lesions. By nature, these images also always had a region of interest selected as part of the acquisition process. As illustrated in Figure 11, the annotation removal pipeline consisted of three main steps that involve a deep learning classifier. We explain these three steps in detail below.

ResNet-18 Classifier. First, we trained a ResNet-18 (9) classifier to determine if an image contained any burnt-in annotations. To prepare the training data for this ResNet-18, we manually selected 1,000 images that contained annotations along with 2,000 images that did not contain any annotations (training set A). We then trained the ResNet-18 on training set A and applied it to the entire dataset. We denote each image within the entire dataset as a positive image if the trained

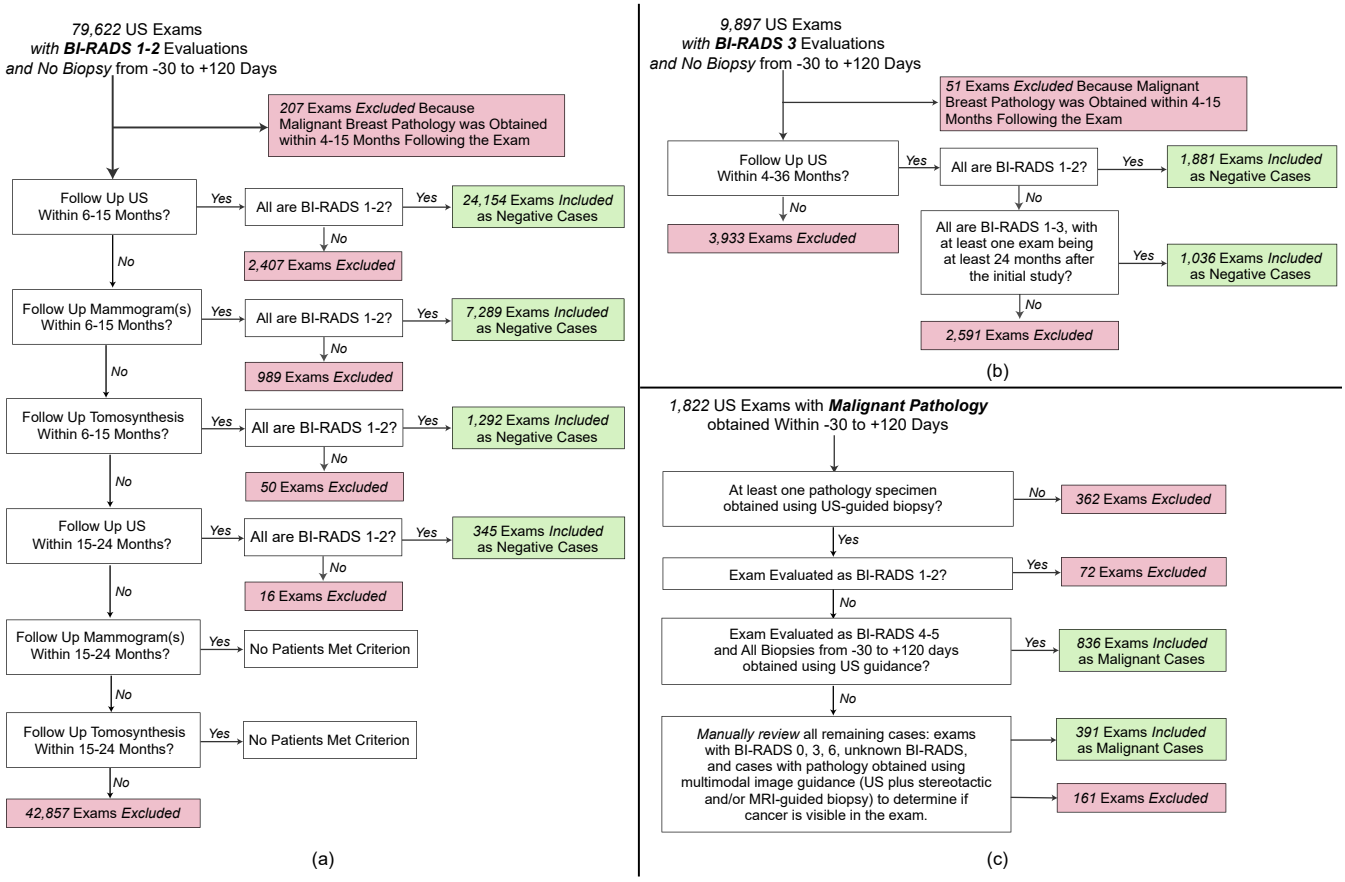


Fig. 9. Filtering protocol for non-biopsied patients whose ultrasound exams had BI-RADS risk assessment labels 1-2 (a), BI-RADS risk assessment label 3 (b), and patients with biopsy-proven cancer (c).

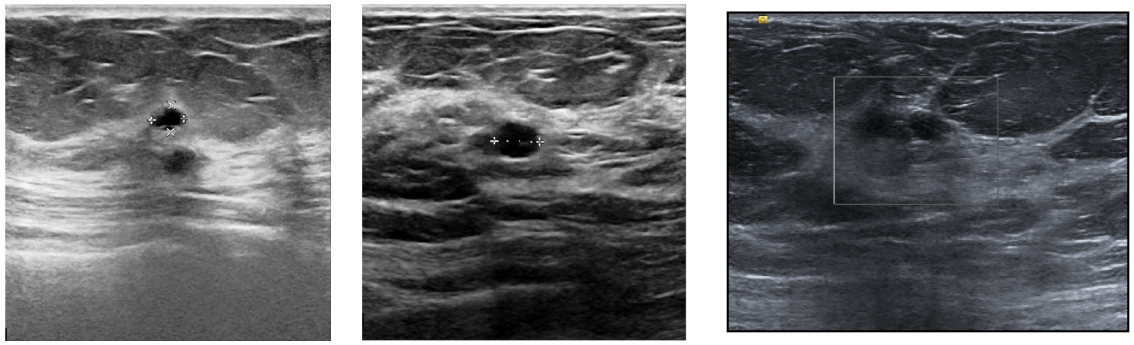


Fig. 10. Example images with burnt-in annotations. (a) Asterisk-like annotations that marks the shape of a finding. (b) Asterisk-like annotations with dots that indicate the size of a finding. (c) Bounding box that indicate the location of a finding.

ResNet-18 computed a positive prediction of the presence of annotations. We then compared each positive image to all other images collected within the same exam. A positive image was matched to another negative image (i.e., the ResNet-18 computed a negative prediction) within the exam if the two images shared more than 95% pixel similarity. This process yielded 380,642 pairs of matched images (training set B). We then trained the ResNet-18 classifier on training set B and applied it on the entire dataset again.

U-Net segmentation network. For those images that were classified as containing annotations by the ResNet-18 classifier, we utilized U-Net (10) to produce a segmentation mask which determined the pixel-level locations of the annotations. To prepare the training data for this U-Net segmentation network, we extracted 462,702 images (training set C) from training set B. Among training set C, 380,642 images were classified as positive by the ResNet-18 and had a matched image within the same exam that shared $> 95\%$ pixel similarity, and 82,060 images were randomly sampled from the negative samples in the dataset (i.e., do not contain annotations). We obtained the segmentation labels for these 380,642 positive images by comparing pixels value of each positive image with its paired

negative image. All pixels whose value in the positive image differed from the value in the paired image were treated as an annotation. A zero matrix was assigned as the segmentation label for the 82,060 negative images. The U-Net was then trained on training set C and applied to all positive images.

Image Inpainting. Finally, we applied image in-painting (11) to the original image to recover the pixels that were selected as burnt-in annotations by the U-Net.

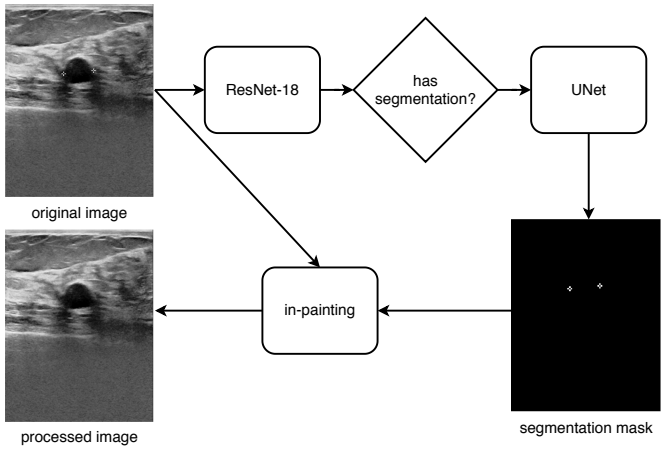


Fig. 11. Annotation removal pipeline. We first utilized a ResNet-18 (9) to predict if an image contained any burnt-in annotations. We then used U-Net (10) to predict the pixel-level locations of annotations. Finally, we applied in-painting to remove the annotations.

3. Label extraction

In this section, we describe how we extracted three types of labels for each breast imaged within an ultrasound exam. The labels include biopsy-proven cancer labels extracted from pathology reports, BI-RADS risk assessment labels extracted from ultrasound reports, and mammographic breast density labels extracted from associated screening and diagnostic mammography reports.

A. Biopsy-proven cancer labels. We extracted benign and malignant labels for each breast within the breast ultrasound dataset from pathology reports. As a first step, we processed pathology reports as described in our screening mammography data report (1). Pathology reports summarize findings by pathologists after examining a small amount of tissue obtained from the breast through a breast biopsy or surgical excision.

Next, for each breast, we assigned a cancer label if the pathology report was dated within 120 days after the ultrasound exam or 30 days before the ultrasound exam. If at least one of the matched pathology reports contained malignant findings, then the breast was assigned a positive malignant label. If at least one of the matched pathology reports contained benign findings, then the breast was assigned a positive benign label, noting that malignant and benign findings are not mutually exclusive. For breasts imaged within ultrasound exams that were not matched with any pathology reports, we assumed that they did not contain any benign or malignant lesions. This is a common assumption, although we acknowledge that patients may have had biopsies at other institutions.

B. BI-RADS risk assessment labels. BI-RADS risk assessment categories were collected from free-text radiology reports, matched with the studies by accession numbers. For extraction of BI-RADS risk assessment labels, we developed a lexicon of phrases used to describe the BI-RADS categories either as a number (e.g. 'BIRADS: 3') or a verbose phrase (e.g. 'bi-rads: probably benign'). We followed categories as defined by the 5th edition of ACR BI-RADS Atlas.

Rarely a single report yielded multiple conflicting BI-RADS risk assessment labels. In cases where it did yield conflicting labels, the reports were manually reviewed to find the correct label.

C. Mammographic breast density labels. The breast tissue density labels were extracted from screening and diagnostic mammography reports. First, we searched for keywords associated with the four breast tissue density classes defined by the BI-RADS lexicon for breast density. In addition to the typical phrases used to describe breast density, we also included phrases that evaluate the percentage of glandular tissue present, which was a criterion used in the 4th edition of the BI-RADS lexicon. All classifying phrases are listed below:

- BI-RADS A (breasts are almost entirely fatty): 'predominantly fatty', 'entirely fatty', 'breasts are comprised of fatty tissue', '10% dense' or '20% dense'.
- BI-RADS B (there are scattered areas of fibroglandular density): 'scattered areas of fibroglandular tissue densities', 'scattered areas of fibroglandular density', 'scattered fibroglandular elements in both breasts', 'scattered fibroglandular densities', 'scattered fibroglandular elements in the left breast', 'scattered fibroglandular elements in the right breast', 'scattered fibroglandular', 'scattered nodular densities', '30% dense', '40% dense' or '50% dense'.
- BI-RADS C (the breasts are heterogeneously dense, which may obscure small masses): 'heterogeneously dense', 'predominantly dense glandular elements', '60% dense' or '70% dense'.
- BI-RADS D (the breasts are extremely dense, which lowers the sensitivity of mammography): 'extremely dense', 'very dense', '80% dense' or '90% dense'.

Next, we matched each breast ultrasound exam with the mammographic breast density label from the closest available mammography exam. If no mammography exam / density label was available then the ultrasound exam was assigned a density label of 'unknown'. Assuming that mammographic breast density does not change much over time and given that our dataset was collected over 7 years, we did not set a time limit on this merging procedure.

1. Wu N, et al. (2019) The nyu breast cancer screening dataset v1, (0. Technical report, 2019. Available at https://cs. nyu. edu/˜ kgeras . . .), Technical report.
2. D'Orsi CJ, Sickles EA, Mendelson EB, Morris EA, , et al. (2013) *ACR BI-RADS Atlas: Breast Imaging Reporting and Data System.* (American College of Radiology).
3. Jamil N, Sembok TMT, Bakar ZA (2008) Noise removal and enhancement of binary images using morphological operations in *International Symposium on Information Technology*.
4. Al-Ghaib H (2016) Morphological procedure for mammogram enhancement and registration in *Applied Imagery Pattern Recognition Workshop*.
5. Szabo TL, Lewin PA (2013) Ultrasound transducer selection in clinical imaging practice. *Journal of Ultrasound in Medicine* 32(4):573–582.
6. Justusson B (1981) Median filtering: Statistical properties in *Two-Dimensional Digital Signal Prcessing II.* (Springer), pp. 161–196.
7. (year?) Python tesseract library (https://pypi.org/project/py-tesseract/).

8. Noumeir R (2005) Benefits of the dicom modality performed procedure step. *Journal of digital imaging* 18(4):260–269.
9. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition in *CVPR*.
10. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation in *International Conference on Medical image computing and computer-assisted intervention*. (Springer), pp. 234–241.
11. Telea A (2004) An image inpainting technique based on the fast marching method. *Journal of graphics tools* 9(1):23–34.