The NYU Breast Cancer Screening Dataset v1.0

Nan Wu[b], Jason Phang[b], Jungkyu Park[b], Yiqiu Shen[b], S. Gene Kim[a,d,e], Laura Heacock[a], Linda Moy[a,d,e], Kyunghyun Cho[b,c,f], and Krzysztof J. Geras[a,b,1]

[a]Department of Radiology, New York University School of Medicine, 660 First Ave, New York, NY 10016, USA; [b]Center for Data Science, New York University, 60 5th Ave, New York, NY 10011, USA; [c]Courant Institute of Mathematical Sciences, New York University, 251 Mercer St, New York, NY 10012, USA; [d]Perlmutter Cancer Center, NYU Langone Health, 160 E 34th St, New York, NY 10016, USA; [e]Center for Advanced Imaging Innovation and Research, NYU Langone Health, 660 First Ave, New York, NY 10016, USA; [f]CIFAR Global Scholar

This manuscript was compiled on September 16, 2019

Deep learning models' generalization ability relies on the size and the quality of the data available for training. Therefore, a large, well curated and representative dataset with high-quality labels is needed to develop a highly predictive model for breast cancer screening, and to further provide insights for other problems in medical imaging. This technical report introduces the NYU Breast Cancer Screening Dataset, which consists of 229,426 digital screening mammography exams (1,001,093 images) from 141,473 patients screened between 2010 and 2017 at NYU Langone Health. Below we outline its statistics, details of image collection and preprocessing, as well as details of the extraction of corresponding labels from screening and pathology reports. Although we do not make this dataset public, we hope to make our research more transparent and reproducible through this technical report. We also hope that other research groups entering the area of medical image analysis with machine learning will find this report a valuable resource.

1. Statistics of the dataset

This dataset is a larger and more carefully curated version of a dataset used in our earlier work (1, 2). It consists of 229,426 digital screening mammography exams (1,001,093 images) from 141,473 patients screened between 2010 and 2017 at NYU Langone Health. Each exam contains at least four[*] images, corresponding to four standard views used in screening mammography: R-CC (right craniocaudal), L-CC (left craniocaudal), R-MLO (right mediolateral oblique) and L-MLO (left mediolateral oblique). A few examples of exams are shown in Figure 1.[†] Patients' ages at the time of the exam range from 16 to 99 years with a majority between 40 to 80 years (see Figure 2).

In addition to the images, the dataset contains four types of associated labels:

- **Breast-level cancer labels**: (i) a binary label for each breast indicating whether a biopsy showed that there is at least one malignant finding in the breast and (ii) a binary label for each breast indicating whether a biopsy showed that there is at least one benign finding in the breast.
- **Pixel-level cancer labels**: indicating exact position of malignant and benign findings in the images in which biopsied findings are visible.
- **Exam-level BI-RADS labels**: one label for each exam, indicating what the initial diagnosis of the radiologist was after screening mammography.
- **Exam-level breast density labels**: one label for each exam, indicating radiologists assessment of breast density.

For breast-level cancer labels, our reference standard was the biopsy report where a pathologist, a doctor who specializes in the evaluation of cells and samples of tissue to diagnose disease, made a final pathologic diagnosis. For pixel-level cancer labels, the annotation of the mammograms were performed by radiologists who have specific training in breast imaging.

A. Splitting the data into training, validation and test sets. For each patient, we obtained outcomes through linkage to our electronic health records within our healthcare system and reviewed each mammogram and pathology reports for patients who underwent a biopsy and/or breast surgery. We sorted the patients according to the date of their latest exam and divided them into disjoint training (first 80%), validation (next 10%) and test (last 10%) sets. The validation set contains the exams of patients whose latest exams are more recent than those in the training set, and the test set contains exams of patients whose latest exams are the most recent. For patients in the training and validation sets we utilized all the exams available for each patient; for test patients we dropped all but the latest exam for each test patient. After this procedure there were 186,816, 28,462 and 14,148 exams in the training, validation and test sets respectively. By partitioning the data in the way described above, we aim to obtain an estimate of the generalization error for patients that we have not seen before whose exams may use more recent imaging equipment and/or a different vendor. There is no significant difference of the age distribution between training, validation and test sets (see Figure 2).

B. Breast-level cancer labels. To obtain labels indicating whether each breast of the patient was found to have malignant or benign findings at the end of the diagnostic pipeline, we used pathology reports from biopsies. We have 6,305 breasts (stemming from 5,832 exams) with at least one biopsy performed within 120 days of the screening mammogram. Among these, for 985 (8.4%) breasts, a biopsy confirmed malignancy, for 5,556 (47.6%) breasts, a concerning imaging finding was biopsy-proven benign and 236 (2.0%) breasts had both malignant and benign findings. The remaining breasts were not matched with any biopsy and assigned a label corresponding to the absence of malignant or benign findings in both breasts. Please refer to Section 3.A for more details on the process of extracting these labels.

C. Pixel-level cancer labels. To collect pixel-level labels, we asked a group of radiologists (provided with the corresponding

[*]Some exams contain more than one image per view as mammographic technologists may need to repeat an image or provide a supplemental view to completely image the breast in a screening examination.

[†]Details of how the images were extracted are in section 2.

[1]To whom correspondence should be addressed. E-mail: k.j.geras@nyu.edu.

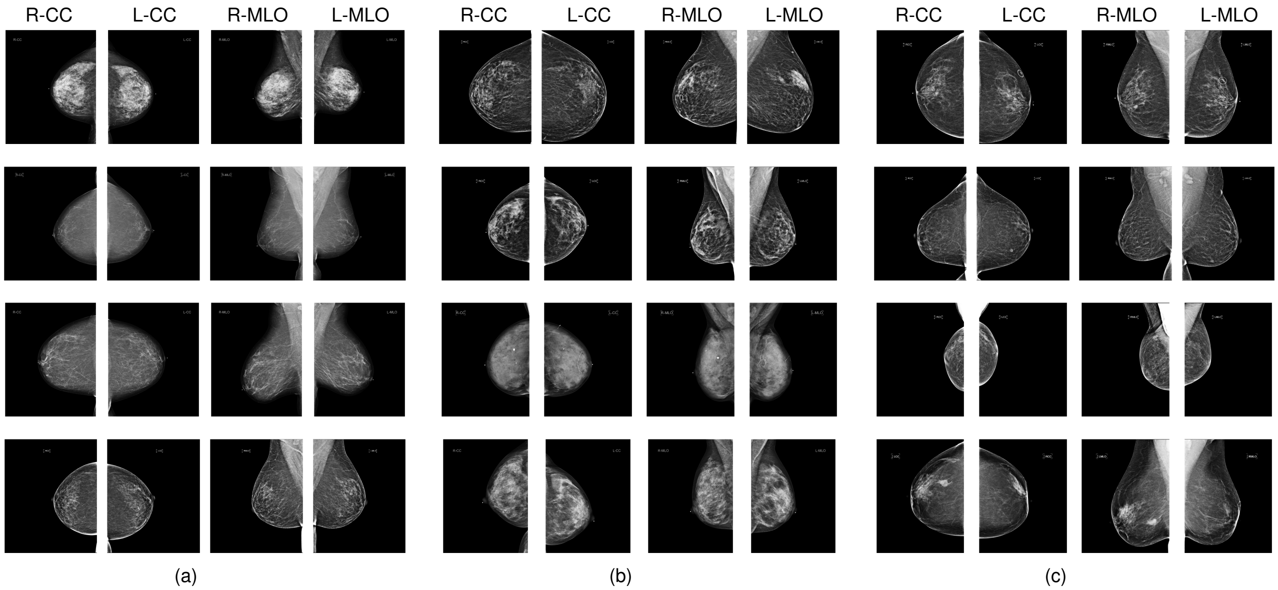


Fig. 1. Twelve exams with four standard views (R-CC, L-CC, R-MLO and L-MLO). **(a)** Exams without any related biopsy. **(b)** Exams with biopsies, which proved the finding to be benign. The breasts labeled as benign were: left in the first row, right in the second row, left in the third row, and right in the fourth row. **(c)** Exams with biopsies, which yielded diagnoses of malignancy. The breasts labeled as malignant were: right in the first row, left in the second row, left in the third row, and left in the fourth row.

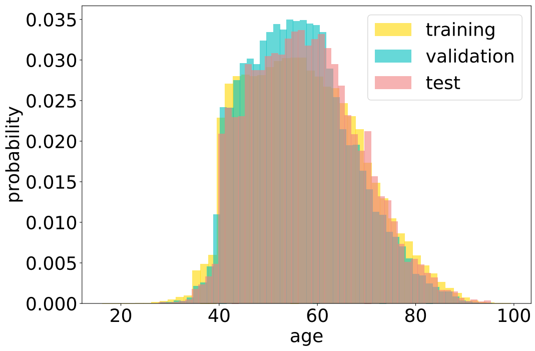


Fig. 2. Distribution of patients' age at the time of the exam in training, validation and test sets.

pathology reports) to retrospectively indicate the location of the biopsied lesions on a pixel level in all images in which they were visible. Each radiologist was working separately on the subset of data. The findings were marked with three colors: green for benign findings (e.g. cyst, fibroadenoma, fibrocystic change), yellow for high-risk[‡] findings that increase the risk of cancer but are not cancer (e.g. atypical ductal hyperplasia, LCIS) and red for malignant findings (e.g. IDC, ILC, DCIS). Figure 3 shows ten examples of such segmentations produced by radiologists. Among the 5,832 exams we provided to the radiologists, there were 3,917 exams (8,080 images) returned with at least one marked finding. Some breasts had multiple benign and/or malignant findings. See Table 1 for the statistics of collected segmentations in detail. We found that, according to the radiologists, approximately 32.8% of exams were mammographically occult, i.e., the lesions that were biopsied were not visible on mammography, even retrospectively, and were identified and biopsied using other imaging modalities, ultrasound or MRI. More details about the occult cases on the breast level are in Table 2.

Table 1. Number of segmentations for biopsied findings for benign, high-risk benign and malignant categories as indicated by the radiologists.

	category		
	benign	**high-risk benign**	**malignant**
training	4753	849	1790
validation	629	93	128
test	386	78	136
overall	5768	1020	2054

Table 2. Number of breasts with malignant and benign findings based on the labels extracted from the pathology reports, broken down according to whether the findings were visible or occult. We found that much fewer malignant findings than benign findings were occult.

	malignant		**benign**	
	visible	**occult**	**visible**	**occult**
training	750	107	2,586	2,004
validation	51	15	357	253
test	54	8	215	141
overall	855 (86.8%)	130 (13.2%)	3,158 (56.84%)	2,398 (43.16%)

D. Exam-level BI-RADS and breast density labels. In addition to the labels extracted from pathology reports, for each exam, we also extracted BI-RADS (Breast Imaging Reporting and Data System) labels (3) reflecting initial diagnosis made by a radiologist at the time of the screening mammography and breast density labels reflecting the amount of fibroglandular tissue in the breast. Both types of labels were extracted

[‡] Unless stated explicitly otherwise, high-risk benign findings were elsewhere considered as a subset of benign findings.

Table 3. Breakdown of data distribution with respect to breast density and BI-RADS. “Unknown” indicates exams with missing or ambiguous information on breast density.

	Breast density					BI-RADS			
	1	2	3	4	Unknown	0	1	2	Total
training	17967 (9.64%)	78532 (42.12%)	78857 (42.30%)	11085 (5.94%)	375 (0.00%)	24230 (12.97%)	86238 (46.16%)	76348 (40.87%)	186816
validation	2517 (8.86%)	11798 (41.53%)	12684 (44.65%)	1410 (4.96%)	53 (0.00%)	3314 (11.64%)	13223 (46.46%)	11925 (41.90%)	28462
test	1260 (8.92%)	5964 (42.23%)	6338 (44.88%)	560 (3.97%)	26 (0.00%)	2246 (15.87%)	6083 (43.00%)	5819 (41.13%)	14148
overall	21744 (9.50%)	96294 (42.05%)	97879 (42.75%)	13055 (5.70%)	454 (0.00%)	29790 (12.98%)	105544 (46.00%)	94092 (41.02%)	229426

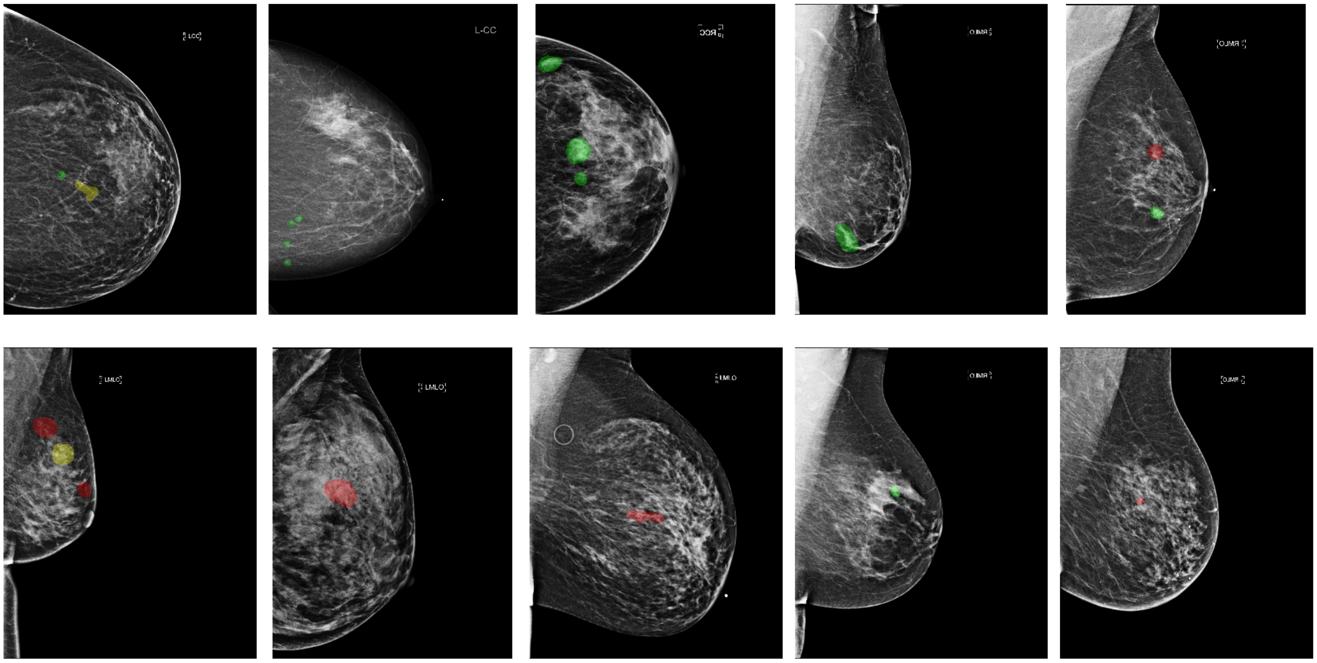


Fig. 3. Examples of segmentations provided by radiologists. The biopsied findings were marked with three colors on the original image: green for benign findings (e.g. cyst, fibroadenoma, fibrocystic change), yellow for high risk benign findings that increase the risk of cancer but are not cancer (e.g. atypical ductal hyperplasia, LCIS), red for malignant findings (e.g. IDC, ILC, DCIS).

from reports produced by a radiologist following breast cancer screening. We considered three BI-RADS categories used in screening mammography: BI-RADS 0 (“incomplete, needs additional imaging”) BI-RADS 1 (“normal”) and BI-RADS 2 (“benign”). Please refer to Section 3.B for more details of the process of extracting these labels. Breast density was qualitatively categorized into four types used in clinical practice in the United States: almost entirely fatty (1), scattered areas of fibroglandular density (2), heterogeneously dense (3) and extremely dense (4) per BI-RADS criteria (3). Please refer to Section 3.C for more details of the process of extracting these labels. Detailed statistics for BI-RADS and breast denisty labels are in Table 3.

E. Scanner information. We extracted information about scanner used for acquiring each image using `ManufacturerModelName` attribute in DICOM files. The images in the dataset are coming from four types of scanners: Mammomat Inspiration, Mammomat Novation DR, Lorad Selenia and Selenia Dimensions. Distribution of types of scanners in training, validation and test data can be found in Table 4.

Table 4. Distribution of types of scanners with respect to images in training, validation and test data.

	Mammomat Inspiration	Mammomat Novation DR	Lorad Selenia	Selenia Dimensions
training	203939	114157	343083	154497
validation	22477	8335	51412	41700
test	3011	657	21694	36131
overall	229427	123149	416189	232328

2. Image collection and preprocessing

In this section we explain in detail the complete processing pipeline for extracting the screening mammography images from raw DICOM data. This pipeline consists of four major phases: (A) data extraction, (B) image cropping and (C) filtering.

A. Image collection. We collected metadata from all DICOM files. Out of 3,050,238 DICOM files, 941,094 files without necessary fields (`PatientID`, `AccessionNumber`, `PatientsName`, `PatientsBirthDate`, `StudyDate`, `SOPInstanceUID`, `ViewPosition`, `ImageLaterality`, `Rows`, `Columns`, `EstimatedRadiographicMagnificationFactor`, FieldOfViewHorizontalFlip, ImageType) were excluded. Manual inspection revealed that excluded images were not mammograms and were representing either some type of metadata or were representing other imaging modalities.

B. Image Cropping. We apply a rigorous procedure to crop all valid mammography images to contain only the relevant breast regions, as shown in Figure 4. As a result of this procedure, we were able exclude large regions of background, significantly reducing image loading time.

Because the background is zero-valued in all correctly processed screening mammograms, we took the largest nonzero connected component of each image to be the breast region. Unfortunately, there were many cases where nonzero artifacts in the background were connected to the largest nonzero component, leading to the cropped image still containing large portions of background. In addition, mammography images occasionally capture patients' chin at the top of the image or abdominal wall at the bottom. These prevent accurate and tight cropping of images.

To address the above issues and correctly identify the location of the breasts, we applied morphological binary erosion and dilation to the mammograms (4, 5). We applied erosion and dilation to the nonzero mask (binary mask of strictly nonzero pixels, not necessarily connected) of the images. Formally, erosion of image A by structural element B is defined as:

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

and 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 E denotes Euclidean space and B_z denotes the translation of B by the vector z and $B^s = \{x \in E | -x \in B\}$. We use the following structural element B, where only the pixels directly connected to the center are considered neighbors:

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

With our choice of structural element, one iteration of erosion shrinks nonzero shapes in an image by one pixel, and dilation expands nonzero shapes in an image by one pixel. By performing erosion for some number of iterations followed by performing dilation for the same number of iterations (also referred to as morphological opening), we were able to eliminate small nonzero artifacts from the binary masks of mammograms. This procedure is often used for noise removal in image processing. Our procedure slightly differs from standard morphological opening as we only dilate the nonzero largest connected component after erosion and discard the rest. As we were only interested in the breast region, which is most often the largest connected component, we chose to ignore irrelevant regions such as arms or the stomach which are typically disconnected from the breast after a sufficient number of iterations of erosion.

Afterwards, we cropped the region of the image which contains the dilated largest connected component, with a small pixel buffer (in case of imperfect restoration of small regions such as the nipple or very small breasts). We found that performing 100 iterations of erosion and dilation and including a 50-pixel buffer in all directions worked well. The full algorithm is summarized in Algorithm 1.

Examples of the images produced as an outcome of this cropping procedure are shown in Figure 5. As a result, we reduced the total number pixels in our images by about 50%, which also reduced loading time during training an inference by approximately 50%.

We used the results of this cropping algorithm to further filter out incorrectly acquired images. We discarded 77 images that had no nonzero component after our erosion-and-dilation procedure–these cases were either entirely blank images, or consisted only of uninformative visual artifacts.

In total, we processed 2,109,067 valid images. In the next stage, images were rejected based on the following cropping-related reasons. If the largest connected component did not contain the midpoint of the image in y-axis (i.e. the largest connected component is contained entirely in the top half or bottom half of the image), this usually indicated that the image captured an object that is not a breast. We discarded 269 such images as shown in Figure 6(f). In addition, the first column of pixels on the side opposite to the direction the breast is oriented must contain nonzero values. For example, if the breast is oriented to the right, then the first pixel column on the left side of the image should contain nonzero values because that is the side where the breast region begins from the body. If this is not the case, then the image might have been incorrectly processed or flipped in a different direction from what we expect from the metadata. We discarded 37,361 such images.

C. Filtering. We further filtered the remaining 2,109,067 images to discard invalid images, as well as sets of images which did not form valid exams.

1. We discarded 127,179 images with duplicate SOPInstanceUID, as these indicated duplicates of the same image.
2. We discarded 161,392 images which contain 'ORIGINAL' in ImageType, as these images have gone through different types of post-processing from the rest of the images which contain 'DERIVED' in ImageType. An example of such image is shown in Figure 6(e).
3. We discarded 142,516 images with ViewPosition different than 'CC' or 'MLO' and with ImageLaterality other than 'L' or 'R'. The three most common discarded views are 'XCCL' (43,592), 'LM' (38,510) and 'ML' (37,721)[§].
4. We discarded 9,353 images with EstimatedRadiographicMagnificationFactor outside the range of $[1, 1.1]$. Rejected values ranged from 1.4 to 1.8.
5. We discarded 952 images with PatientSex different than 'F' (female).
6. We discarded 87,587 with BreastImplantPresent different than 'NO'. An example of an image with implants is shown in Figure 6(d).
7. We discarded 2,804 images whose original image matrix size before cropping were smaller than 2290 × 1890 pixels.

[§] We acknowledge that, for some exams, these additional views which are non-standard for screening mammography might show findings not visible in the standard four views. This is a simplifying design choice we made.

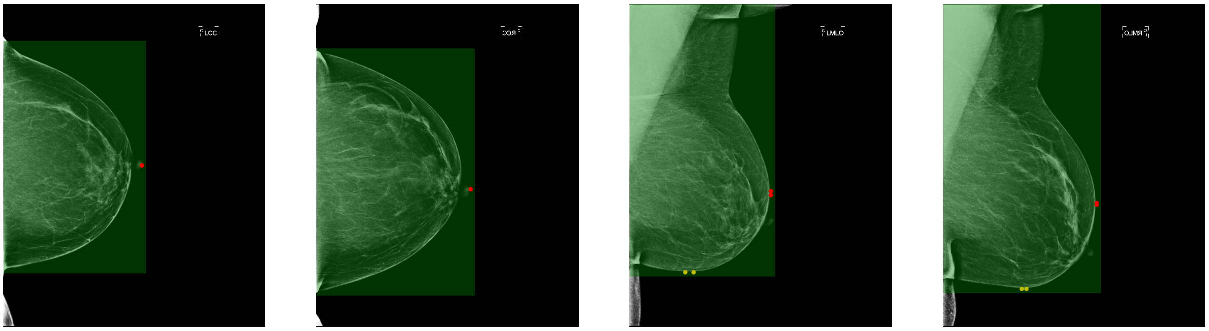
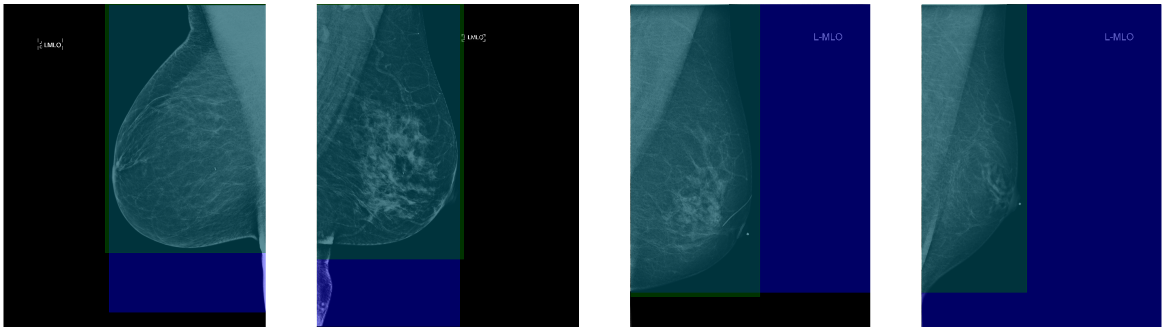


Fig. 4. Example of cropping windows containing breast regions from exam images. The green overlay is the window used in cropping. Red points are the rightmost nonzero pixels after the breasts are flipped to be oriented to the right. Yellow points are the bottommost nonzero pixels for MLO views. These points are used in calculating the size and location of cropping windows.



(a) (b) (c) (d)

Fig. 5. Effects of using erosion and dilation. The blue area is the bounding box of the largest connected component of the image. The green area is the bounding box of the nonzero mask of the image after selecting largest connected component, dilating, and adding buffers, which is the final product of our preprocessing algorithm. The teal area is the overlap between the two bounding boxes. Green bounding boxes in (a), (b), (c), (d), have areas smaller by 17%, 18%, 43%, and 61% compared to corresponding blue bounding boxes, respectively.

The most common invalid size was 1458 × 1458 pixels with 2,070 occurrences.

8. We discarded 703 images for which the cropped breast region were smaller than 350 pixels in width or 1000 pixels in height. These images were incorrectly acquired images or from male patients.
9. We discarded 846,720 diagnostic mammography images and tomosynthesis images based on the values of `PerformedProcedureStepDescription`, `SeriesDescription` and `ProtocolName`. An example of a diagnostic image is shown in Figure 6(b). The most common cause was 234,335 occurrences of 'MAMMO TOMOSYNTHESIS SCREENING BILATERAL' in `PerformedProcedureStepDescription`.
10. We discarded 45,999 images whose ratios of nonzero pixels were more than 95% and 682 images with less than 5%, as they usually indicate empty images or images with invalid processing.
11. We discarded 11,293 images whose `PresentationLUTShape` were not 'identity', as they usually indicate images with inverted pixel intensities. We hypothesize that these images were created in error by technologists. An example of an inverted image is shown in Figure 6(c). All 11,293 rejected images had `PresentationLUTShape` value set to 'INVERSE'.
12. We discarded 319 images whose `ExposureStatus` were not 'NORMAL', as they were usually incorrectly taken images (i.e. the process of taking image is aborted in the middle). An example of an aborted image is shown in Figure 6(a). All 11,293 rejected images had `ExposureStatus` value of 'ABORTED'.
13. Exam-level filtering:
    (a) All images in an exam must have the same `FieldOfViewHorizontalFlip` value: either uniformly 'YES' or 'NO'. We discarded exams that did not, as we empirically found that in most of these cases the value of this field was not set correctly. Three exams were discarded.
    (b) The BI-RADS label (details of how the BI-RADS labels were extracted are in Section 3.B) for that exam must be present. 89,713 exams had missing or incorrect labels.
    (c) All four standard views must be present in an exam: R-CC, L-CC, R-MLO, and L-MLO. 38,422 exams had some views missing.

Some images were discarded for more than one of the reasons above. At the end of this filtering procedure, we found 242,498 valid exams with 1,057,978 valid images.

D. Selecting optimal image windows for model input. Even after cropping, the images remain very large, which poses a

Algorithm 1 Cropping mammograms

```
procedure CROP_MAMMOGRAM(numIter, bufferSize)
    Given image $A$, Get the mask $m$ where $A > 0$
    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 dilated mask $d$.
    Select a window $w$ from $A$ which contains $d$ and buffers of bufferSize in all directions.
    Record the location of $w$ and save the image cropped by $w$.
```

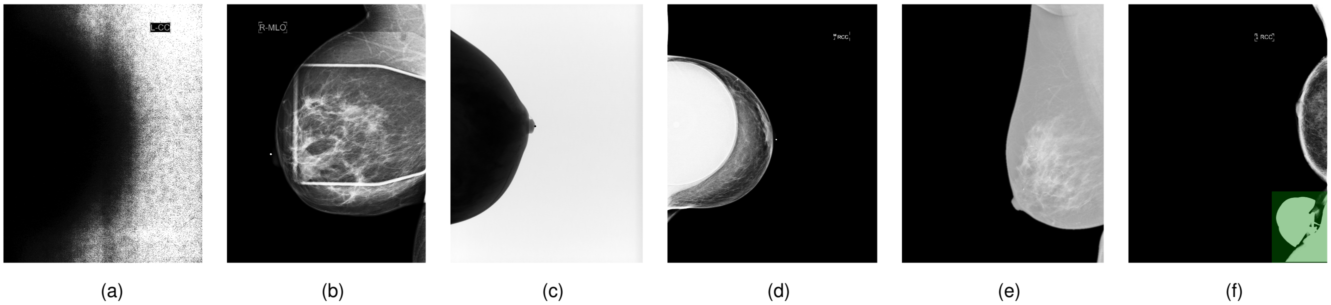


Fig. 6. Examples of rejected mammograms. (a) `ExposureStatus` is 'ABORTED'. (b) `ProtocolName` contains 'Diagnosis'. (c) `PresentationLUTShape` is 'INVERSE'. (d) `BreastImplantPresent` is 'YES'. (e) `ImageType` contains 'ORIGINAL'. (f) cropped image is in the bottom half of the original image.

challenge for the limited GPU memory when processing images with a deep neural network. To find the most informative region of the image, we calculated heuristically the optimal image matrix size for each of the two views (CC and MLO). Furthermore, we also calculated optimal center points based on which cropping window contained the most nonzero pixels in each image, assuming a fixed window size. Keeping the number of pixels for both views approximately the same, we chose to adopt different aspect ratios for the CC and MLO views; shapes of nonzero pixels in MLO views tend to be longer in the y-axis than those of CC views since they contain the upper breast. In addition, we sought to have the cropping window contain the outer breast region rather than other tissue such as the pectoralis major.

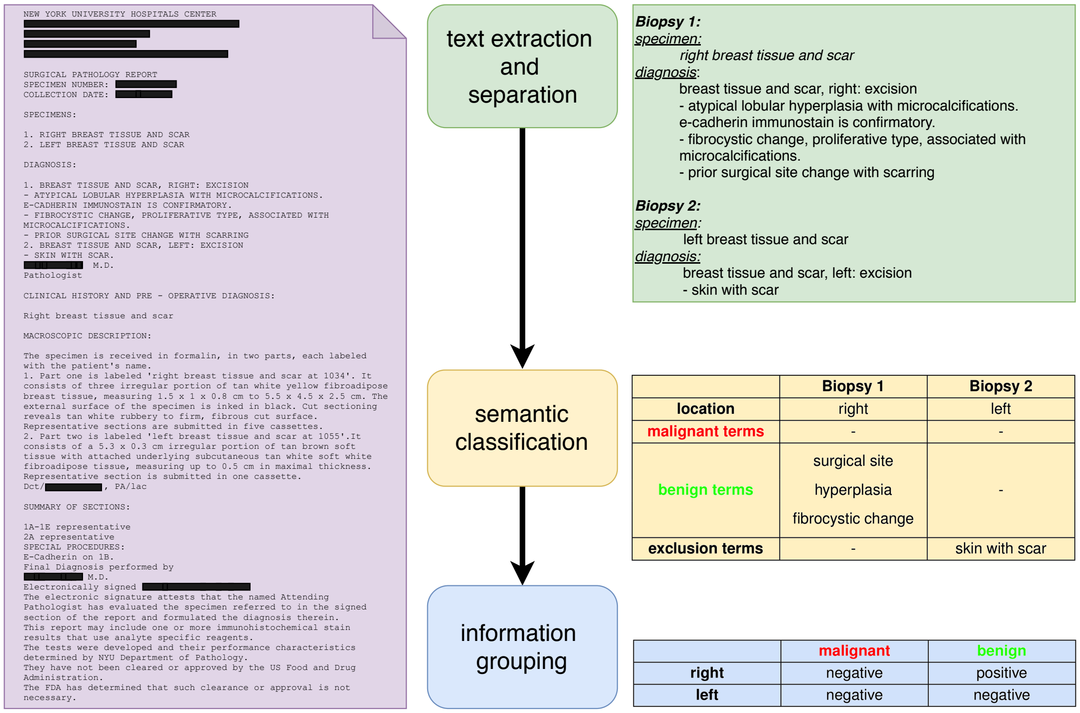
For CC-view images, it sufficed to have the windows contain the rightmost points of the breasts (nonzero pixel) when the breast flipped accordingly to be oriented to the right. For MLO-view images, because of the significant vertical asymmetry, we enforced that the windows contain the bottommost points of the breasts. This is necessary in order to avoid prioritizing non-breast regions such as arms. The rightmost points are calculated from the bottom two-thirds of the image for the same reason. If an MLO view image was large enough that the cropping window cannot contain both rightmost and the bottommost points, we instead constrained the window to contain the rightmost points of the image, and then selected the new bottommost point obeying that constraint. As a result, when we calculated the optimal image matrix sizes for each view to capture the most nonzero pixels for the entire training data, the window for MLO view ended up being narrower than that of CC view.

By selecting for maximal nonzero pixels above a threshold over the training set, we determined optimal image matrix sizes of 2677 × 1942 pixels for the CC view and 2974 × 1748 pixes for the MLO view. These optimal window size and locations are necessary to apply data augmentation such as random shifts and rotations around the most informative portions of the mammograms.

3. Label extraction

A. Extraction of breast-level cancer labels. To obtain breast-level labels indicating whether a patient had breast cancer at the time of the screening exam, we extracted information from pathology reports created following biopsies of suspicious findings. A standard pathology report contains information about the location, size, appearance and type of cells found in the biopsied tissue which helps determine a diagnosis. We used information from two sections of the reports: "specimens" and "diagnosis". Using these, we extracted diagnostic information on each biopsied finding separately, and the extracted information was summarized in four binary labels: (i) whether there was at least one benign finding in the left breast, (ii) whether there was at least one benign finding in the right breast, (iii) whether there was at least one malignant finding in the left breast, (iv) whether there was at least one malignant finding in the right breast. The three main phases of our algorithm for automatic cancer label extraction: (i) text extraction and separation, (ii) semantic classification, and (iii) information grouping are explained below in detail. Figure 7 shows a pathology report and the outputs we obtained after each phase.

A.1. Text extraction and separation. Although pathology reports are written following certain common conventions, there remains significant variability in their structure. These differences arise if the reports were collected from different hospitals or even if were written by different pathologists in the same hospital. Despite these differences, for a vast majority of correctly written reports, the "specimens" and "diagnosis" sections contained all the information necessary to determine in which breast (left or right) the listed biopsied findings were found, and what the diagnostic conclusion is for each or both. To extract these two sections, we manually collated a list of common section headers and footers, and extracted



NEW YORK UNIVERSITY HOSPITALS CENTER
[redacted]

SURGICAL PATHOLOGY REPORT
SPECIMEN NUMBER: [redacted]
COLLECTION DATE: [redacted]

SPECIMENS:

1. RIGHT BREAST TISSUE AND SCAR
2. LEFT BREAST TISSUE AND SCAR

DIAGNOSIS:

1. BREAST TISSUE AND SCAR, RIGHT: EXCISION
- ATYPICAL LOBULAR HYPERPLASIA WITH MICROCALCIFICATIONS.
E-CADHERIN IMMUNOSTAIN IS CONFIRMATORY.
- FIBROCYSTIC CHANGE, PROLIFERATIVE TYPE, ASSOCIATED WITH
MICROCALCIFICATIONS.
- PRIOR SURGICAL SITE CHANGE WITH SCARRING
2. BREAST TISSUE AND SCAR, LEFT: EXCISION
- SKIN WITH SCAR.
[redacted] M.D.
Pathologist

CLINICAL HISTORY AND PRE - OPERATIVE DIAGNOSIS:

Right breast tissue and scar

MACROSCOPIC DESCRIPTION:

The specimen is received in formalin, in two parts, each labeled with the patient's name.
1. Part one is labeled 'right breast tissue and scar at 1034'. It consists of three irregular portion of tan white yellow fibroadipose breast tissue, measuring 1.5 x 1 x 0.8 cm to 5.5 x 4.5 x 2.5 cm. The external surface of the specimen is inked in black. Cut sectioning reveals tan white rubbery to firm, fibrous cut surface.
Representative sections are submitted in five cassettes.
2. Part two is labeled 'left breast tissue and scar at 1055'.It consists of a 5.3 x 0.3 cm irregular portion of tan brown soft tissue with attached underlying subcutaneous tan white soft white fibroadipose tissue, measuring up to 0.5 cm in maximal thickness.
Representative section is submitted in one cassette.
Dct/[redacted], PA/lac

SUMMARY OF SECTIONS:

1A-1E representative
2A representative
SPECIAL PROCEDURES:
E-Cadherin on 1B.
Final Diagnosis performed by
[redacted] M.D.
Electronically signed [redacted]
The electronic signature attests that the named Attending Pathologist has evaluated the specimen referred to in the signed section of the report and formulated the diagnosis therein.
This report may include one or more immunohistochemical stain results that use analyte specific reagents.
The tests were developed and their performance characteristics determined by NYU Department of Pathology.
They have not been cleared or approved by the US Food and Drug Administration.
The FDA has determined that such clearance or approval is not necessary.

Biopsy 1:
specimen:
right breast tissue and scar
diagnosis:
breast tissue and scar, right: excision
- atypical lobular hyperplasia with microcalcifications.
e-cadherin immunostain is confirmatory.
- fibrocystic change, proliferative type, associated with microcalcifications.
- prior surgical site change with scarring

Biopsy 2:
specimen:
left breast tissue and scar
diagnosis:
breast tissue and scar, left: excision
- skin with scar

	Biopsy 1	**Biopsy 2**
location	right	left
malignant terms	-	-
benign terms	surgical site hyperplasia fibrocystic change	-
exclusion terms	-	skin with scar

	malignant	**benign**
right	negative	positive
left	negative	negative

Fig. 7. An illustration of the three main phases in the process of extracting breast-level cancer labels: text extraction and separation, semantic classification, and information grouping. The violet box on the left shows an example of a pathology report (anonymized). The corresponding output of each phase is shown on the right. In this example, there are two biopsies (Biopsy 1 and Biopsy 2). In the **text extraction and separation** phase, we obtain the relevant specimen and diagnosis information for each biopsy. The output of the the **semantic classification** phase consists of four pieces of information for each biopsy: location, malignant terms found, benign terms found and exclusion terms found. Finally, in the last stage (**information grouping**) the above information is consolidated and the final binary malignant and benign labels for each breast are obtained. In this example, one of biopsied specimens is from the left breast one is from the right breast. Only benign terms were found for the right breast. No terms concerning breast cancer were found for the left breast. At the end of the label extraction process, the right breast was labeled as benign and as not malignant, while the left breast was labeled as not benign and as not malignant.

text between detected header-footer pairs. For each section, we partitioned the extracted text into subsections if multiple biopsies were enumerated in the report. After completing this extraction, we obtained a list containing information on all biopsied findings separately.

A.2. Semantic classification. In this phase, we extracted location and diagnosis information for each biopsy. To identify a location, we searched for words and phrases indicating the side of the breast, such as "left", "right" and their variants. To determine the diagnosis, we built two comprehensive lexicons for benign and malignant related terms, including words and phrases. We also collected a list of phrases that indicate a specific biopsy is not breast cancer related, and used it as a separate lexicon. The ten most frequent terms in each lexicon are listed below, along with the number of occurrences:

- **Malignant terms:** *ductal carcinoma (2046), ductal carcinoma in situ (1464), invasive ductal carcinoma (1149), invasive carcinoma (557), metastases (414), metastatic (224), invasive lobular carcinoma', 180), adenocarcinoma (160), invasive mammary carcinoma (128), metastatic carcinoma, (117).*

- **Benign terms:** *fibrocystic change (5842), fibroadenoma (3768), hyperplasia (2569), cyst content (2882), benign breast tissue (1364), fibrocystic changes (1279), fibrosis (1235), negative for malignancy (1049), adipose tissue (1026).*

- **Exclusion terms:** *benign skin (150), explant (93), non-diagnostic (80), no mammary epithelium is identified (59), breast capsule (53), breast implant (48), fibrous capsule (48), no benign or malignant epithelial cells seen (48), no mammary epithelial cells (45), dermal scar (43).*

In addition, we collected two kinds of prefixes, indicating prior breast cancer history and negation of malignancy. The relationships between phrases in all three lexicons are complex. For example, *intraductal papilloma* is an indicator for benign findings but *intraductal papilloma with ductal carcinoma in situ* indicates malignancy. Simply checking for the presence of each term in the text could lead to a significant number of incorrect labels. Instead, we organized all the terms in the three lexicons as a forest, and designed a custom algorithm for searching for terms based on an overall lexicon forest. Given the above lexicon forest and prefixes of history and negation of malignancy, the algorithm is designed to search for all semantically valid terms in the lexicon forest in the input text. Terms are searched in an increasing order of their degree in the lexicon forest. To ensure validity over semantics, terms were removed if there was any prefix before it, or if it is included in another longer term which has already been found. The returned terms are grouped into three classes, malignant, benign and excluded[¶], according to the lexicon it is in. In summary, for each biopsy, we collected the location

[¶] Exclusion terms indicate that the biopsy was from location other than the breast, therefore they should not be taken into consideration for breast cancer.

```
benign skin with mild perivascular lymphocytic infiltrate (benign)
    0 benign skin (excluded)

intraductal papilloma with ductal carcinoma in situ (malignant)
    0 ductal carcinoma in situ (malignant)
        0 ductal carcinoma (malignant)
    1 intraductal papilloma (benign)
```

Fig. 8. Two examples for trees in the lexicon forest. The roots of trees in the forest are phrase not including in any other ones. Parent in a tree always contain its children. Terms with different degrees are in different colors and their type (malignant, benign or excluded) is indicated in the brackets.

Algorithm 2 Searching for terms in the text about diagnosis based on lexicon forest.

```
term is a class with two instance variables:
    term.term is the term (phrase or word) itself;
    term.type ∈ {BENIGN, EXCLUDED, MALIGNANT}, is the corresponding type of the term.
lexicon_forest.terms is a dictionary of terms grouped by their degree.
    lexicon_forest.terms[degree] is a list of instances of term.
lexicon_forest.degree is the maximum degree of all the trees in the forest.
history_prefixes is a list of strings that serve as prefixes before terms indicating breast cancer history.
negation_prefixes is a list of strings that serve as prefixes before terms indicating negating malignancy.
negations is a class with three instance variables:
    negations.BENIGN is an empty list
    negations.EXCLUDED is an empty list
    negations.MALIGNANT is a list including strings in both history_prefixes and negation_prefixes

function TEXT_CLEANING(term, text, negations)
    cleaned_text = text
    for prefix in negations do
        cleaned_text.remove(prefix + ' ' + term)
    return cleaned_text

function SEARCH_IN_LEXICON(text, lexicon_forest)
    for degree in [0, lexicon_forest.degree] do
        for term in lexicon_forest.terms[degree] do
            text_to_search = TEXT_CLEANING(term.term, text, negations.term.type)
            if degree $\neq$ 0 then
                for phrase in terms_found do
                    if term.term in phrase then
                        text_to_search = TEXT_CLEANING(phrase, text_to_search, [])
            if term.term in text_to_search then
                terms_found + = term
    return terms_found
```

and identified all terms indicating whether the tissue was benign, malignant or that it was not correctly acquired (and excluded). Examples of trees in the lexicon forest are shown in Figure 8, and the algorithm is summarized in Algorithm 2.

A.3. Information grouping. We grouped the information extracted by the above procedure and translated them into binary labels. The possible initial classes for a single biopsy sample is one of benign, malignant and excluded. The logic of assigning the initial class was as follows. (i) Excluded, if any term indicating necessity of exclusion are found. (ii) Malignant, if there is no term indicating the need of exclusion and at least one malignant term is found. (iii) Benign, if only benign terms are found. Next, we overwrote excluded cases to be benign if we found any specially designated benign terms (such as *scant benign-appearing ductal cells* and *proteinaceous debris*). We then aggregated the information for different biopsies into a breast-level label in the following manner. If there was a benign or malignant biopsy for a given breast, it was labeled as positive for the corresponding class. We consider benign and malignant labels separately and allowed for a breast to be classified as potentially both benign and malignant. Hence, in an exam with two breasts, this led to four binary labels: left-benign, right-benign, left-malignant, and right-malignant.

After applying the above automatic cancer label extraction procedure, we tasked an attending fellowship-trained breast radiologist with providing labels for excluded cases. We also asked an attending radiologist to verify cases labeled as both malignant and benign. In summary, out of the 15,152 pathology reports we collected, 1,219 cases had labels provided or refined by the radiologist. We successfully collected breast-level cancer labels for all the pathology reports we have. Among them, we had 1,182 reports including malignant biopsied findings of left breast and the same number of reports including malignant biopsied findings of right breast. There were 6,710 reports in which we found benign findings in left breast and 6,587 reports in which benign findings were found in right breast. And there are also 48 reports where we found both malignant and benign findings in the left breast and 85 reports indicating both types of findings in the right breast.

Finally, we matched the cancer labels extracted from pathology reports with the screening mammography exams to construct our dataset. We assigned a cancer label to an exam if the biopsy was performed within 120 days after the screening mammogram. In cases where there are multiple biopsies within that window, if any findings are positive (i.e. presence of a benign finding in left breast) for any report in that window, we applied the positive label to that exam. Screening mammography exams that were not matched to any cancer label were given negative labels for all four labels.[‖]

B. Extraction of exam-level BI-RADS labels. To obtain exam-level BI-RADS labels for the screening exams, we used a conservative, keyword-based heuristic to parse 419,739 reports and extract the corresponding labels for each exam. We used the following metadata from reports: `RawReport`, which contains textual diagnosis, `ExamDescription`, which describes the type of exam, `Acc`, which contains accession number, and `Mrn`, which contains patient identifier. In many cases, the screening exam reports contain BI-RADS information out of the scope of our analysis, such as BI-RADS determined from an ultrasound exam. We retained only the exams for which we could successfully extract a valid screening mammography-based BI-RADS category.

First, we rejected all reports pertaining to diagnostic, tomosynthesis, or ultrasound exams based on the corresponding keywords in the `ExamDescription` field in the exam metadata. Next, we excluded sections of the report text based on section headers and textual markers, such as 'DIAGNOSTIC' section headers and identifiable strings indicating the start of a standard legend describing each label class. After isolating the relevant text, we proceeded to identify BI-RADS labels contained within the remaining text. We performed an exhaustive search for strings exactly matching one of the following label formats:

(a) 'birads: label'

(b) 'bi-rads: label'

(c) 'bi-rads: label'

(d) 'bi-rads category: label',

wherein the label corresponds to either the number or exact text string representing one of the following BI-RADS categories: 0 ('incomplete'), 1 ('negative'), 2 ('benign'), 3 ('probably benign'), 4a ('low suspicious'), 4b ('moderate suspicious'), 4c ('high suspicious') and 5 ('highly suggestive of malignancy'). After obtaining exact matches based on the above, we only kept exams for which there was exactly one successful match. If there were zero or more than one matches, we conservatively excluded that exam, as it indicates a failure mode of our heuristic.

As BI-RADS 0 and BI-RADS 1 and BI-RADS 2 should be the only BI-RADS categories used in screening mammography, we condensed all BI-RADS categories into three classes for the purposes of training our model. BI-RADS 0, 4a/b/c and 5 were mapped to a new 'BI-RADS 0' as each indicates a possibility of malignancy. BI-RADS 1 is retained at 'BI-RADS 1'. BI-RADS 2 and 3 are mapped to a new 'BI-RADS 2', as they both indicate benign findings. This procedure resulted in a dataset consisting of a single BI-RADS label over three classes for each of our valid screening mammography exams.

C. Extraction of exam-level breast density labels. We used a similar keyword-based heuristic to extract the breast density labels for each exam from `RawReport`. Keywords considered for each category of breast density were the following:

- **almost entirely fatty (1):** *predominantly fatty, entirely fatty, breasts are comprised of fatty tissue;*
- **scattered areas of fibroglandular density (2):** *scattered areas of fibroglandular tissue densities, scattered areas of fibroglandular density, scattered fibroglandular, scattered nodular densities;*
- **heterogeneously dense (3):** *heterogeneously dense, heterogeneously dense with a nodular parenchymal pattern*;
- **extremely dense (4):** *extremely dense, breasts are very dense;*

Exams coming with a report that only contains one type of keyword were kept and labeled as the corresponding category of breast density. All other exams, including exams without any matching keywords and exams matching with more than one type of keywords, are considered as 'Unknown', indicating missing or ambiguous information on breast density. There are 228,972 exams successfully aligned with a reliable breast density class, which is over 99% of the entire dataset (see Table 3).

1. Geras KJ, et al. (2017) High-resolution breast cancer screening with multi-view deep convolutional neural networks. *arXiv:1703.07047.*
2. Wu N, et al. (2018) Breast density classification with deep convolutional neural networks in *ICASSP.*
3. 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).
4. Jamil N, Sembok TMT, Bakar ZA (2008) Noise removal and enhancement of binary images using morphological operations in *International Symposium on Information Technology.*
5. Al-Ghaib H (2016) Morphological procedure for mammogram enhancement and registration in *Applied Imagery Pattern Recognition Workshop.*

[‖] We acknowledge that assigning negative labels to exams with no follow-up biopsies is based on a simplifying assumption that 100% of patients, who were recommended further imaging after screening mammography, returns to complete it. In reality, a small fraction of patients decides not to continue or to continue at a different institution. Although this phenomenon introduces some small level of noise to our labels, comparisons between different predictive models or human readers utilizing these labels as ground truth are still fair as all are affected equally by this noise.