Computer-Aided Diagnosis for Breast Ultrasound Imagery Dataset

Tristan P. Hansen[1], Jeffrey S. Baggett[2], Richard L. Ellis[3]

[1]Graduate Research Assistant, Mayo Clinic Health System, La Crosse, WI
[2]Professor of Mathematics & Statistics, University of Wisconsin-La Crosse
[3]Radiologist, Breast Imaging and Intervention, Mayo Clinic Health System, La Crosse, WI

This manuscript was compiled on 2025-06-17

Abstract

The Computer-Aided Diagnosis for Breast Ultrasound Imagery (CADBUSI) dataset comprises 63,769 breast ultrasound exams from 45,284 patients collected between 2002 and 2025 at Mayo Clinic Health System, specifically curated to advance machine learning applications in breast cancer diagnosis. This comprehensive collection includes 440,966 breast ultrasound images and 81,362 videos with BI-RADS® assessments and pathology-verified diagnoses, providing ground truth labels for 55,291 unique breasts (47,942 benign, 7,349 malignant). Our rigorous processing pipeline ensures clinical relevance while standardizing the dataset for research use. This includes extracting text from images with a custom Faster R-CNN model, automatically detecting relevant image regions, removing measurement calipers through our Noise2Noise inpainting approach, and applying HIPAA-compliant anonymization techniques throughout. By addressing challenges in ultrasound image standardization and linking radiological findings with pathological outcomes, this dataset enables the development of computer-aided diagnostic tools with potential to improve breast cancer detection accuracy, reduce unnecessary biopsies, and enhance clinical decision-making in breast imaging. Our code is available at https://github.com/Poofy1/CADBUSI-Database.

1 Statistics of the dataset

The Computer-Aided Diagnosis for Breast Ultrasound Imagery (CADBUSI) dataset consists of 63,769 breast ultrasound exams from 45,284 patients collected between 2002 and 2025 at Mayo Clinic Health System. Following rigorous data processing and quality control measures detailed in Section 2 and Section 4, this includes a total of 440,966 breast ultrasound images and 81,362 breast ultrasound videos. The focus of this dataset is exclusively on breast ultrasound imagery, excluding other modalities or anatomical regions. Each exam contains pixel data from one or both breasts. The data was originally stored using the Digital Imaging and Communications in Medicine (DICOM) Standard [6]. Each exam went through a criterion check found in Section 2 and data cleaning found in Section 4. All data is fully anonymized in the resulting dataset, as described in Section 3. The number of images per ultrasound exam ranges from 1 to 249, with an average of 7.40, as shown in Figure 1 (a). Similarly, the number of videos per exam ranges from 1 to 48, with an average of 2.28, as displayed in Figure 1 (b). Patient ages range from 12 to 101 years with a mean age of 55.59. The average image size is approximately 531 × 808 pixels. The average video size is approximately 552 × 827 pixels, with an average of 180.18 frames for each video (before subsampling in Section 5). There are 28,259 left breast exams, 27,032 right breast exams, and 8,478 bilateral exams. The dataset organization and statistical presentation follow approaches established in previous breast ultrasound dataset work by Shamout et al. [8].

The distribution of BI-RADS® (Breast Imaging-Reporting and Data System) [1] risk assessments and mammographic breast densities found in the radiology reports is detailed in Table 4 and Table 1, respectively. The ultrasound exams were performed using a variety of machines; the distribution of the machine manufacturers is summarized within Table 3.

To develop and evaluate our models, we partitioned the CADBUSI dataset into training, validation, and test sets. All ultrasound exams were first grouped by patient identifier, then patients were randomly assigned to one of the three sets. This patient-based splitting strategy ensures that all exams from a single patient remain within the same set, preventing potential data leakage between splits. The dataset consists of breast-level labels, indicating whether or not at least one malignant lesion was found in that breast.

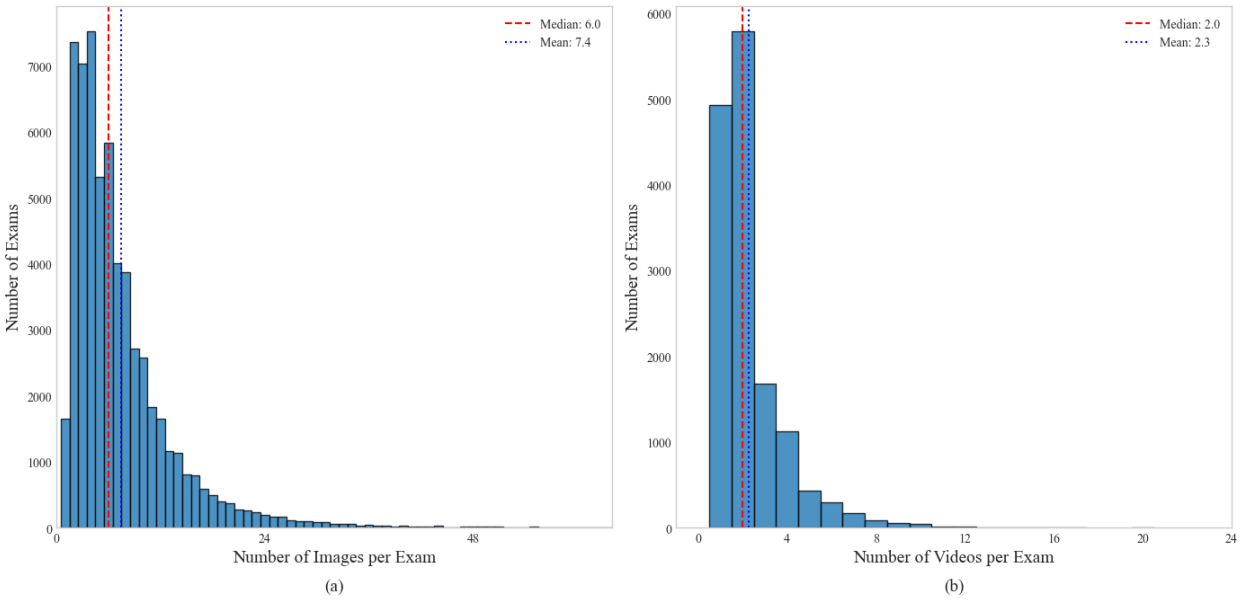


Fig. 1. Distribution of media per exam in the CADBUSI dataset. (a) Number of images per exam. (b) Number of videos per exam.

1.1 Training, validation and test sets

Each patient was randomly assigned to the training set (70% of the patients), validation set (20% of patients), and test set (10% of patients), ensuring that all exams of a given patient were contained within each set. This patient-based splitting strategy, following the approach of Shamout et al. [8], prevents data leakage from patients with multiple exams by keeping all their data within the same subset. Since each breast ultrasound study contains multiple images (average 7.40 per exam), the breast-level splits result in 283,821 training images, 79,795 validation images, and 40,794 test images.

Table 1: Distribution of mammographic breast density categories across dataset splits at the breast level.

Mammographic breast density	**Overall**	**Training set**	**Validation set**	**Test set**
A (breasts are almost entirely fatty)	1,988 (3.1%)	1,414 (3.2%)	405 (3.2%)	169 (2.6%)
B (scattered areas of fibroglandular density)	17,120 (26.8%)	11,991 (26.8%)	3,438 (27.1%)	1,691 (26.4%)
C (breasts are heterogeneously dense)	21,205 (33.3%)	14,793 (33.1%)	4,243 (33.5%)	2,169 (33.8%)
D (the breasts are extremely dense)	3,983 (6.2%)	2,823 (6.3%)	777 (6.1%)	383 (6.0%)
Unknown density	19,473 (30.5%)	13,648 (30.6%)	3,820 (30.1%)	2,005 (31.2%)

Table 2: Number of malignant and benign labels across the left and right breasts in the training, validation, and test sets at the breast ultrasound level.

	malignant		**benign**	
	right	**left**	**right**	**left**
training	2,517	2,558	16,413	17,199
validation	734	770	4,612	4,925
test	388	382	2,368	2,425
overall	3,639	3,710	23,393	24,549

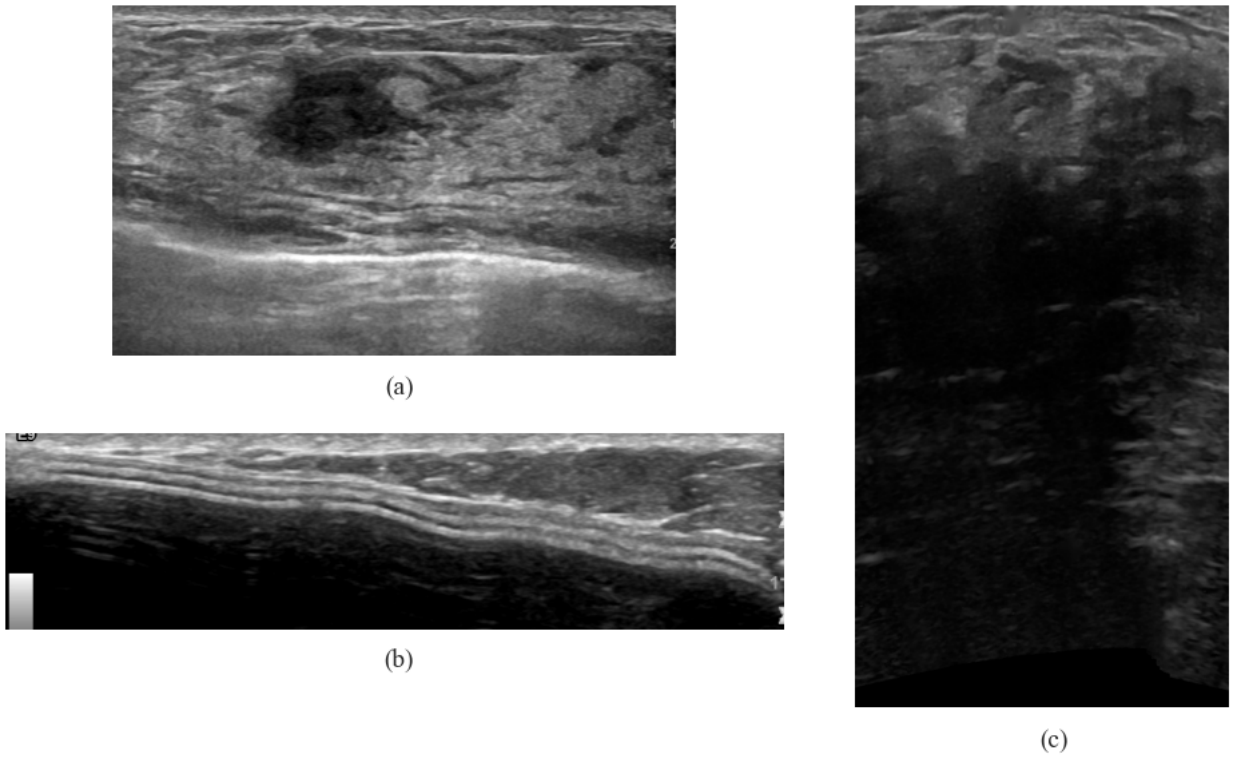


Fig. 2. Visualization of our dataset's cropping procedure applied to images with varying width-to-height proportions. The figure displays three representative examples: (a) an image with an aspect ratio close to the dataset's average, (b) an image with the lowest aspect ratio found in our dataset, and (c) an image with the maximum aspect ratio value present in the dataset.

Table 3: Distribution of ultrasound devices across the training, validation, and test sets at the image level.

Device	**Training set**	**Validation set**	**Test set**
LOGIQE9	139,748	39,276	19,518
EPIQ 5G	51,534	14,199	8,076
LOGIQE10	46,310	13,065	6,974
EPIQ 7G	26,043	7,694	3,301
iU22	8,733	2,229	1,340
EPIQ Elite	7,714	2,130	1,021
S2000	1,883	646	286
Antares	863	198	107
TUS A500	836	300	119
HDI 5000	80	29	51
TUS AI800	35	0	0
RS85	17	1	0
Sequoia	8	25	0
SEQUOIA	8	3	1
XarioXG	5	0	0
S3000	4	0	0

Table 4: Distribution of BI-RADS risk assessment categories across dataset splits at the breast level.

BI-RADS risk assessment	Overall	Training set	Validation set	Test set
1	11,431 (17.9%)	7,958 (17.8%)	2,319 (18.3%)	1,154 (18.0%)
2	27,255 (42.7%)	19,177 (42.9%)	5,390 (42.5%)	2,688 (41.9%)
3	9,068 (14.2%)	6,433 (14.4%)	1,728 (13.6%)	907 (14.1%)
4	9,670 (15.2%)	6,748 (15.1%)	1,956 (15.4%)	966 (15.1%)
4A	1,425 (2.2%)	990 (2.2%)	287 (2.3%)	148 (2.3%)
4B	129 (0.2%)	90 (0.2%)	26 (0.2%)	13 (0.2%)
4C	285 (0.4%)	192 (0.4%)	63 (0.5%)	30 (0.5%)
5	953 (1.5%)	635 (1.4%)	210 (1.7%)	108 (1.7%)
6	3,553 (5.6%)	2,446 (5.5%)	704 (5.6%)	403 (6.3%)

1.2 Data labels

The dataset contains a total of 55,291 unique breasts, 47,942 (86.71%) labeled as benign and 7,349 (13.29%) labeled as malignant. This binary classification approach aligns with clinical decision-making frameworks where the primary diagnostic question is whether malignancy is present or absent within a given breast.

Each breast ultrasound study receives either a benign or malignant label based on comprehensive clinical evidence. A malignant label indicates that at least one malignant lesion was identified within the breast, while a benign label signifies the absence of malignant findings. These labels serve as ground truth derived from pathology verification and radiological follow-up protocols, as described in Section 2.4.

Labels are assigned on a per-breast basis for each exam, meaning bilateral ultrasound exams contribute two separate labeled instances (one for each breast). The distribution shows malignant findings in 3,639 right breasts and 3,710 left breasts, as shown in Table 2.

Figure 3 illustrates the distribution of days between the most recent ultrasound exam prior to biopsy and the biopsy procedure itself, with 95.98% of biopsies occurring within 30 days of the preceding ultrasound exam.

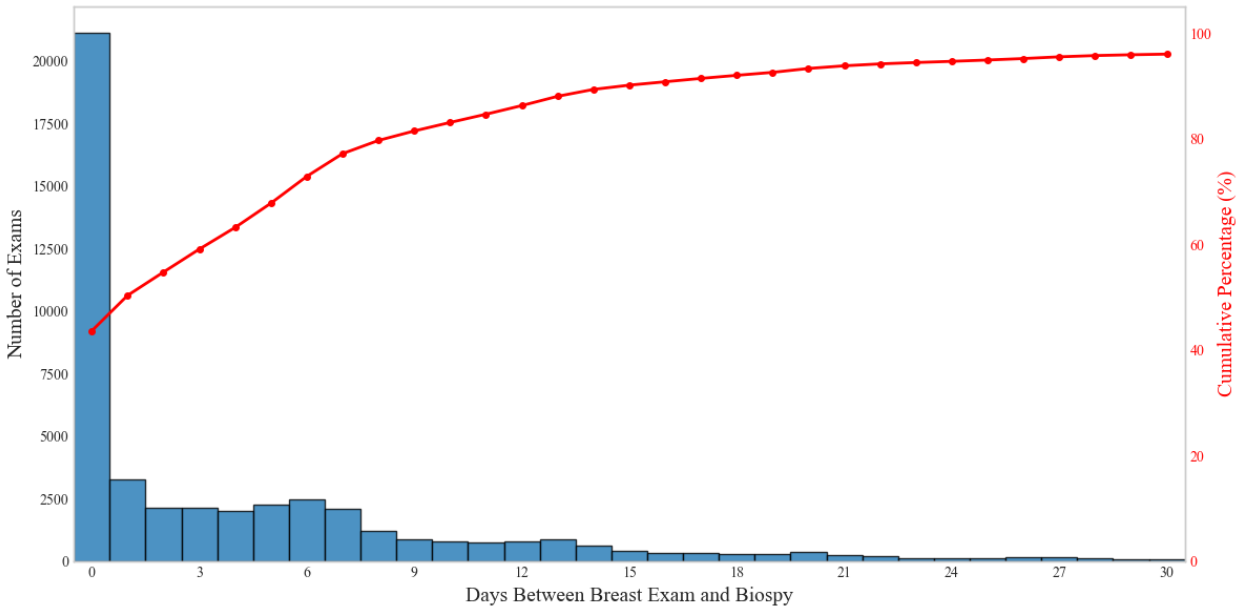


Fig. 3. Distribution of days between breast exams and corresponding biopsy

2 Data collection and preprocessing

Our dataset creation followed a structured pipeline to transform clinical data into a machine learning-ready resource with verified diagnostic labels. Starting with a query of Mayo Clinic's medical imaging database, we developed algorithms to extract and standardize information from unstructured radiology and pathology reports. We addressed challenges in correlating imaging findings with tissue sampling results and implemented a multi-criteria approach for ground truth verification. This process distilled an initial pool of 104,453 exams into 60,179 benign and 9,184 malignant cases with high-confidence labels.

2.1 Breast ultrasound query

The initial data collection phase involved querying the Mayo Clinic Health System internal medical imaging database using Google BigQuery. Our database query began by identifying potential cases using procedure codes associated with breast imaging. We then refined results to exclude males and include only patients who had undergone at least one ultrasound breast exam. Our query retrieved all breast-imaging studies for these qualifying patients, regardless of modality, to establish a complete history of breast imaging. The query retrieved comprehensive study information including accession numbers, procedure descriptions, modality details, full radiology reports, and exam timestamps. We also captured demographic data such as patient age at the time of exam, ethnicity, race, and zip code.

This data collection process involved querying two primary datasets: radiology and pathology. The radiology dataset served as our master query for patient identification because it contains studies of all breast-imaging exams, regardless of whether pathology was subsequently performed. The initial data collection spanned from 1992 to 2025, yielding 1,256,408 radiology studies from 92,971 patients. Of these patients, 26,055 (28.02%) had at least one pathology report, totaling 656,669 pathology reports.

The initial raw data obtained from the database query often consisted of unstructured data within the radiology and pathology reports. This required parsing to extract relevant information and organize the findings into a normalized format, as detailed in the following sections.

2.2 Radiology Parsing

Our radiology parsing system was designed to extract structured information from the highly variable free-text reports generated during breast imaging exams. We developed a parsing system to extract three primary data elements from each report: BI-RADS assessment categories, breast laterality (left, right, or bilateral), and imaging modality. Additionally, we captured supplementary information including breast density classification, documentation of biopsies (including specific detection of ultrasound-guided procedures), and the complete impression and pathology sections when available within the radiology report.

To address the considerable variability in reporting styles across radiologists and time periods, we implemented a set of regular expression patterns to locate common terms and identifiers. For BI-RADS classification alone, our system includes nine distinct complex regex patterns to recognize diverse documentation formats such as "BI-RADS ASSESSMENT: CODE: 4B-SUSPICIOUS", "US BIRADS: 2 benign", and "BI-RADS® Category: 5 - HIGHLY SUSPICIOUS" among many others. Our extraction algorithm searches across multiple report fields (DESCRIPTION, TEST_DESCRIPTION, RADIOLOGY_REPORT, RADIOLOGY_NARRATIVE) using a fallback strategy when information isn't found in primary fields, maximizing data capture. Figure 4 illustrates this extraction process with examples of both successful parsing (a) and a case requiring fallback strategy implementation where laterality and other fields were not successfully extracted from primary fields (b).

The laterality determination algorithm includes keyword detection for standard terms ("RIGHT", "LEFT", "BILATERAL") as well as abbreviated forms ("RT", "LT") with specific pattern matching for contextual awareness (e.g., distinguishing "L BI" from "LT" with trailing space). To enhance extraction accuracy, we incorporated contextual analysis to prevent misclassification when these keywords appeared in unrelated sections of the report.

The parsing process also served as a quality control step, allowing us to apply additional filtering criteria to the dataset. We excluded 128,430 radiology studies (10.22% of the initial dataset) that were recorded

as having been performed outside the Mayo Clinic system to maintain consistency in reporting standards. Studies with incomplete BI-RADS assessments (category 0) were removed from the dataset, representing 80,013 studies (6.37% of the initial studies). We also enforced our ultrasound modality inclusion criterion at this stage, removing 10,113 patients representing 814,016 studies (64.79%) who did not have at least one ultrasound exam after all filtering steps. This parsing and filtering process yielded 233,949 high-quality studies for subsequent analysis.

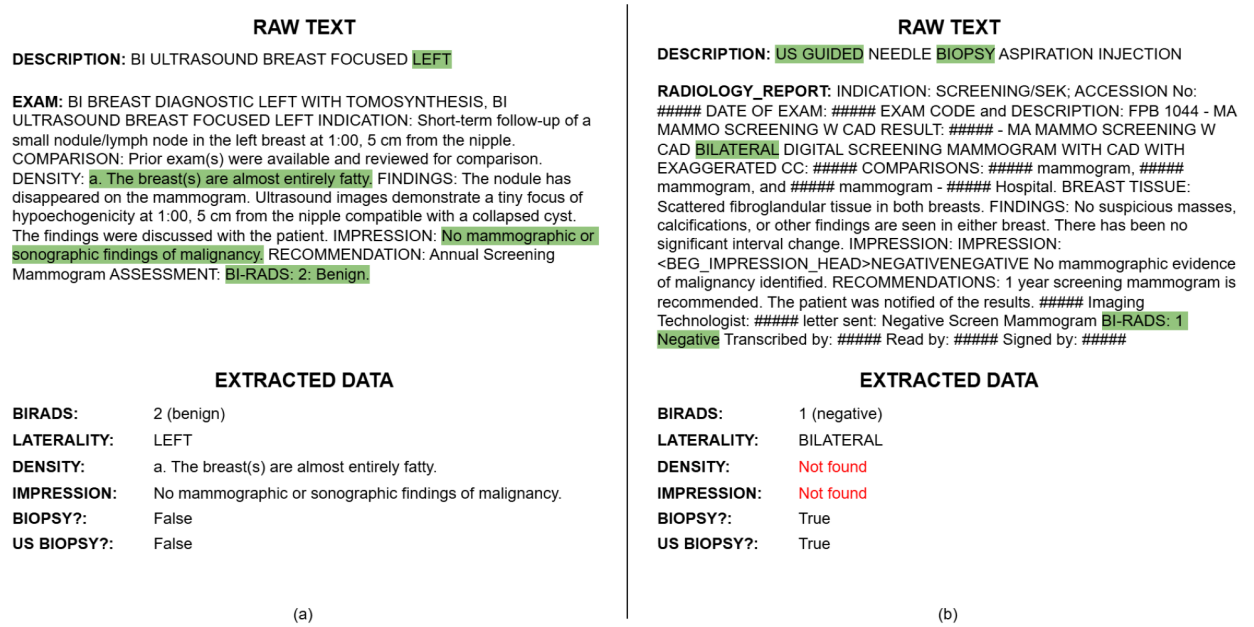


Fig. 4. Breast imaging report parsing and data extraction examples. (a) Successful parsing of a left breast ultrasound report with BI-RADS 2 (benign) assessment. (b) Parsing challenges shown where laterality is missed in the DESCRIPTION field but recovered from the RADIOLOGY REPORT as BILATERAL, while DENSITY and IMPRESSION fields fail to parse due to capitalization issues interrupting the extraction process.

2.3 Pathology parsing

The primary objective of our pathology parsing system was to extract structured diagnostic information from free-text pathology reports and standardize findings on a per-lesion basis. This separate parsing system aimed to identify key diagnostic sections within each pathology report, including laterality and whether the specimen was benign or malignant.

For the laterality determination, we implemented a multi-layered approach that searched across three fields (final_diag, PART_DESCRIPTION, and SPECIMEN_NOTE) with a fallback strategy when information wasn't found in primary fields. The algorithm analyzed each part separately in multi-part reports, tracking mentions of "RIGHT" and "LEFT" to make accurate laterality assignments and avoid misclassification when laterality terms appeared in unrelated contexts.

For diagnostic classification, we implemented a three-tier hierarchy of pattern matching rules:

1. First checking for malignant patterns using nine distinct regex expressions (e.g., "INVASIVE DUCTAL CARCINOMA", "DCIS", "CARCINOMA")
2. Examining the surrounding context (50 characters before matched terms) to identify negated findings (e.g., "NEGATIVE FOR", "NO EVIDENCE OF") that would override an initial malignant classification
3. Applying a comprehensive set of over 25 benign classification patterns (e.g., "FIBROADENOMA", "NORMAL BREAST TISSUE", "FIBROCYSTIC")

A key challenge was handling reports describing multiple specimens or "parts" within a single document. We implemented a splitting function that identified lettered divisions (e.g., "A.", "B.") within the diagnostic text and created separate entries for each part, preserving the relationship to the original report. This specimen-level splitting approach was crucial because some of our data was already organized at the specimen level, while other pathology reports contained multiple specimens per entry - requiring consistent granularity across the dataset. This processing step expanded our initial collection of 656,669 pathology reports to 2,233,473 discrete specimens.

This multi-step classification approach resulted in our final specimen label distribution being 1,341,238 benign (60.05%), 770,589 malignant (34.50%), 121,646 unknown (5.45%). Our laterality determination analysis identified 1,112,806 left specimens (49.82%), 1,117,397 right specimens (50.03%), and 3,270 unknown laterality (0.15%). The distributions of diagnostic categories and specimen laterality are visualized in Figure 5.

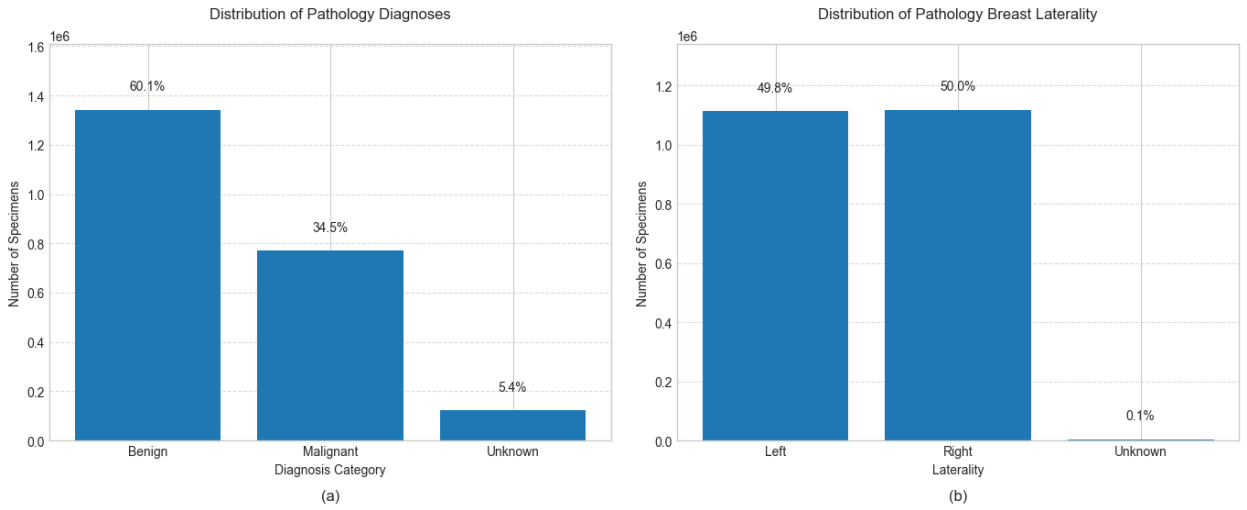


Fig. 5. Pathology data characteristics in the CADBUSI dataset. (a) Distribution of pathology diagnoses showing benign, malignant, and unknown categories with their respective percentages. (b) Distribution of breast laterality in pathology specimens.

2.4 Label generation

Having parsed both radiology and pathology reports, we then developed algorithms to generate ground truth labels for each breast examination. This process aims to provide a ground truth label (benign or malignant) for each breast ultrasound study by systematically evaluating both the extracted radiology findings and the corresponding pathology results. We implement four different algorithms to determine these ground truth labels, with the complete filtering protocol illustrated in Figure 6.

Low-Risk Studies Without Biopsy
For Negative BI-RADS (BI-RADS 1-2) breasts without biopsies, we apply stringent follow-up criteria. These breasts are labeled benign only when no biopsy is performed within a window of 30 days before to 120 days after the ultrasound exam, no non-benign BI-RADS assessments (BI-RADS 4, 4A, 4B, 4C, 5, or 6) appear in the 24-month follow-up period, no malignant pathology is reported within 15 months after the ultrasound, and at least 6 months of follow-up data are available. This approach identifies 41,071 (39.32%) benign breasts.

BI-RADS 3 Studies
For BI-RADS 3 breasts (probably benign), we apply an extended follow-up protocol. Breasts are labeled benign when no biopsy is performed within the -30 to +120 day window, and all subsequent ultrasounds within 36 months have either only BI-RADS assessments of null, 1, or 2, or only assessments of null, 1, 2, or 3 with at least one exam at 24+ months. Additionally, no malignant pathology can be found within 15 months, and a minimum follow-up period of 6 months is required. This protocol identifies an additional 8,860 (8.48%) benign breasts.

BI-RADS 6 Studies
Studies with BI-RADS 6 assessment (known malignancy) are labeled malignant when the patient has at least one confirmed malignant pathology in their record, resulting in 4,039 (3.87%) malignant breasts.

Pathology-Confirmed Studies
For remaining ultrasound studies, we examine pathology reports within an 8-month window. For potential malignant breasts (BI-RADS 4, 4A, 4B, 4C, 5, or 6), we label them malignant if laterality matches between ultrasound and pathology, and at least one ultrasound-guided biopsy is performed. For potential benign breasts (BI-RADS 1, 2, 3, 4, 4A, or 4B), we label them benign if laterality matches and pathology confirms benign findings with no malignant findings. This approach identifies 5,145 (4.93%) malignant breasts and 10,248 (9.81%) benign breasts.

This multi-step classification approach, following rigorous labeling principles established by Shamout et al. [8], results in a final label distribution of 60,179 (57.61%) benign breasts and 9,184 (8.79%) malignant breasts, with 35,090 (33.59%) remaining unlabeled as they did not meet the inclusion criteria of any of the four classification algorithms. These unlabeled breasts were not included in the final database.

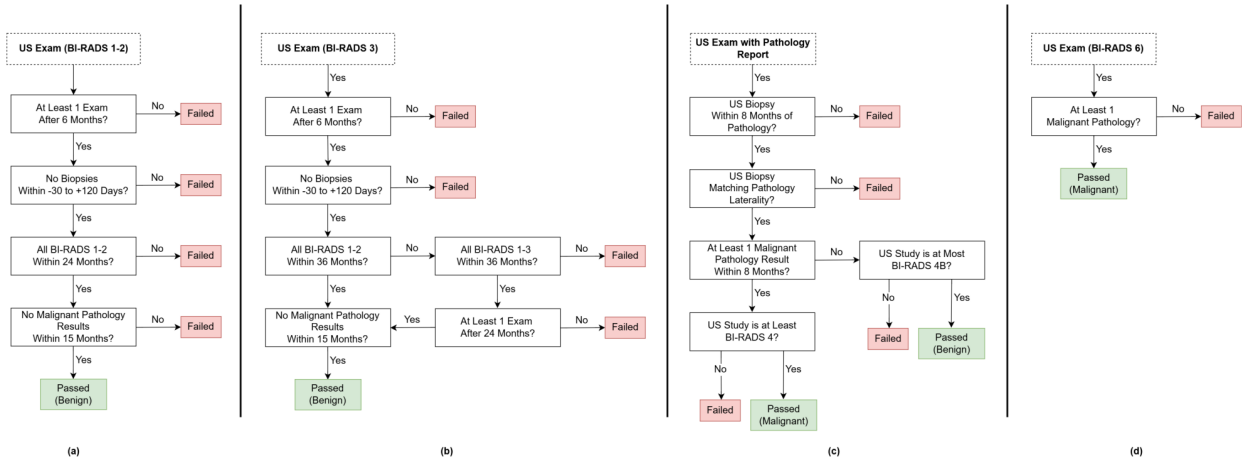


Fig. 6. Filtering protocol for non-biopsied patients whose ultrasound exams had BI-RADS risk assessment 1-2 (a), BI-RADS risk assessment 3 (b), patients with a pathology report (c), and patients with biopsy-proven cancer (d). Inspired by Shamout et al. [8].

3 Anonymization

To protect patient privacy and comply with data protection regulations including HIPAA (Health Insurance Portability and Accountability Act) and institutional IRB requirements, all DICOM files in the dataset undergo a thorough anonymization process. This procedure carefully removes or modifies sensitive patient information while preserving the clinical utility of the images for research.

The process begins by identifying and removing direct patient identifiers within the DICOM files. This includes information such as patient names, referring physician details, institution names, and unique study codes that could potentially link an image back to an individual.

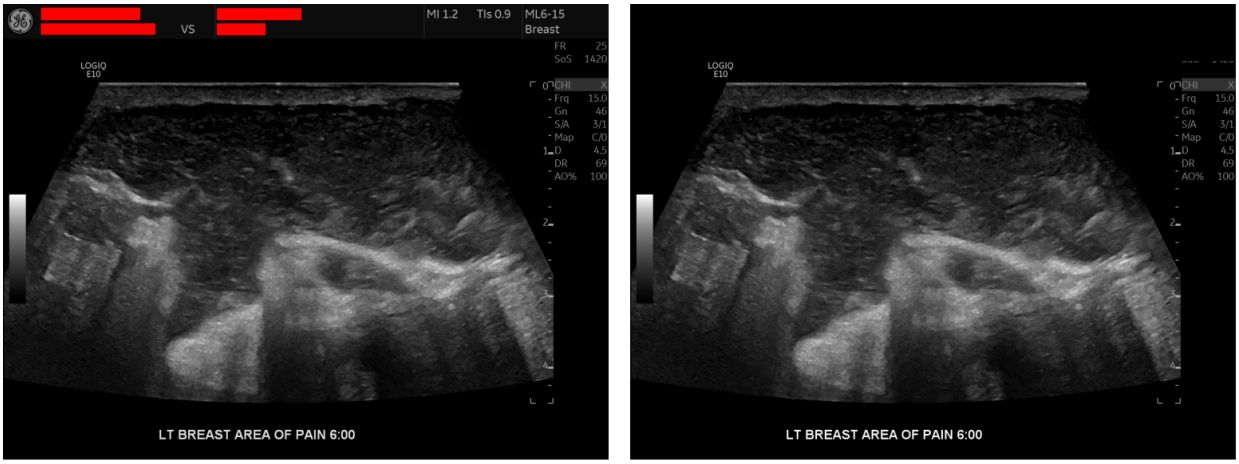
Additionally, dates are standardized to include only the year (removing month and day), and scan times are removed to eliminate exact temporal identifiers while preserving study sequence.

To ensure internal data provenance, original patient and study identifiers are replaced with format-preserving encryption using the NIST SP 800-38G FF1 algorithm with a 128-bit key [2]. The FF1 (Format-preserving, Feistel-based encryption) algorithm is a mode of operation for block ciphers that encrypts data while preserving its original format and length. For example, a 10-digit patient ID remains a 10-digit number after encryption, and alphanumeric identifiers maintain their character composition. This approach maintains the format and length of the original identifiers while providing HIPAA-compliant protection that preserves

referential integrity across the dataset. The encryption keys are secured via Mayo Clinic's data protection infrastructure with access restricted to authorized research personnel.

The final step addresses burned-in identifiers - patient information physically embedded in the pixel data. Our approach leverages DICOM metadata already present in ultrasound files, specifically the 'RegionLocationMinY0' parameter from the SequenceOfUltrasoundRegions tag (0018,6011), which defines the vertical starting position of the clinically relevant ultrasound image. For cases where this parameter is available, we use its exact value; otherwise, we default to a conservative threshold (y0=101) based on analysis of our dataset. This boundary is used to replace all pixel data above this line with black pixels (zero values), effectively removing the header region where identifying information appears while preserving the complete diagnostic content, as demonstrated in Figure 7. This process handles both single-frame images and multi-frame ultrasound video sequences, with special consideration for compressed DICOM formats that require decompression before pixel manipulation.

To validate the effectiveness of this approach, we manually reviewed over 10% of the processed images (804,582) and confirmed complete removal of identifying information with no cases of under-cropping or removal of diagnostically relevant image data. We have no evidence to suggest that the anonymization process introduced any image bias that could affect the training of models designed to differentiate between benign and malignant lesions.



(a) (b)

Fig. 7. Demonstration of the DICOM anonymization process for ultrasound images. (a) Raw ultrasound image with patient identifiers visible in the header region (highlighted in red). These identifiers include patient name, ID, and exam information that could potentially link the image to a specific individual. (b) Processed image after our automated anonymization algorithm identifies and removes all patient-identifying information by replacing the header region with black pixels while preserving the complete diagnostic content of the ultrasound image.

4 Image processing

Following the medical data filtering and label generation process described in Section 2, we obtained 664,015 breast ultrasound images and 105,420 videos from 69,363 exams. To prepare these raw ultrasound images for machine learning applications, we applied a comprehensive image processing pipeline that refined the dataset to 440,966 images and 81,362 videos. This pipeline included optical character recognition (OCR) for metadata extraction, automated cropping to isolate diagnostic regions, detection of image pairs, detection of measurement calipers, inpainting to remove potential data leakage sources, and quality filtering to locate poorly taken ultrasound images. To the best of our knowledge, no bias affecting the differentiation of benign or malignant lesions was introduced during these processes, as equal treatment was applied to all images regardless of their pathology.

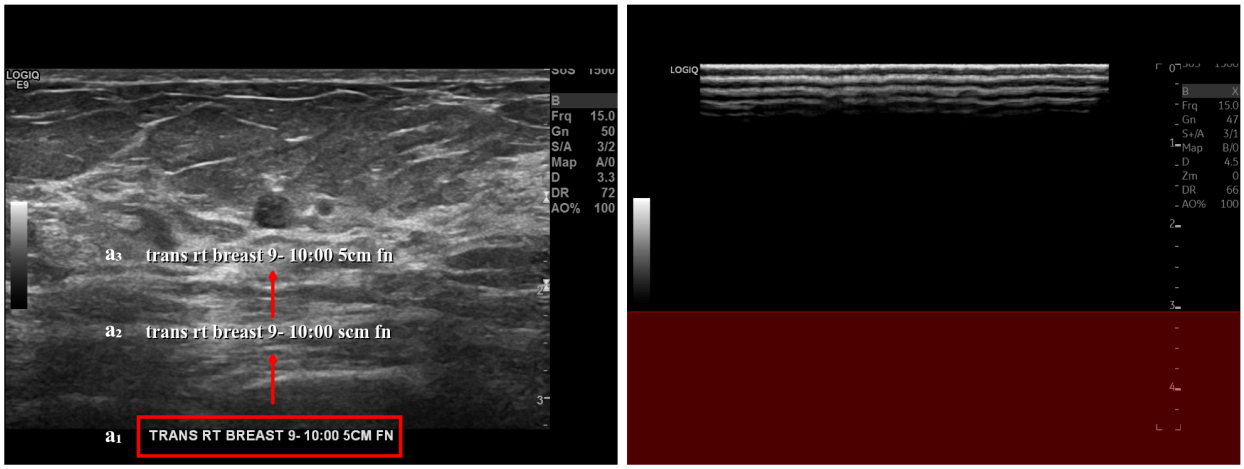
4.1 Text extraction and OCR

Ultrasound images commonly contain vital diagnostic metadata embedded as text within annotation regions. To systematically extract this information, we developed a two-phase approach beginning with precise localization of text-containing regions.

This localization is required because text regions vary in position across different ultrasound images. We implemented a Faster R-CNN model [7] with a SqueezeNet1.1 backbone [4], trained on 1,100 manually annotated ultrasound images from diverse equipment vendors. The model identifies two distinct text-containing region classes that hold critical clinical information such as laterality, orientation, clock position, and distance measurements.

The detection model was configured with multiple anchor scales (32-512 pixels) and aspect ratios to accommodate varying text region dimensions across different ultrasound machines. Evaluation on a test set of 100 ultrasound images demonstrated a precision of 0.83, recall of 0.88, and an F1-score of 0.86. The resulting bounding boxes are used to crop relevant text regions from the original images before proceeding to OCR, significantly improving character recognition by isolating text-dense areas and reducing noise. In cases where the detection model failed to identify text regions, we applied a fallback strategy that automatically cropped the lower third of the image—a region commonly containing descriptive text in ultrasound images. Figure 8 illustrates both our primary text detection approach and the fallback strategy.

For the OCR phase, we implemented a pipeline using EasyOCR [5] to extract text from the cropped regions. Each region was preprocessed with light Gaussian blurring (3×3 kernel) to reduce noise while preserving text clarity. To improve detection quality, we expanded the bounding boxes by 2.5% in each direction to ensure complete capture of text that might intersect with region boundaries. The extracted text underwent post-processing to correct common OCR errors, particularly equipment manufacturer names (e.g., correcting "loc" to "logiq") and standardizing case for consistent downstream parsing.



(a) (b)

Fig. 8. Text extraction process in ultrasound images. (a) Primary approach using Faster R-CNN detection: a_1 shows the detected text region containing key metadata, a_2 describes the extracted text from EasyOCR, and a_3 shows the corrected diagnostic information, fixing the error from 'scm fn' to'5cm fn'. (b) Fallback strategy when detection fails, automatically cropping the bottom third of the image (highlighted in red) where descriptive text commonly appears in ultrasound images.

4.2 Cropping

Ultrasound images typically contain diagnostic information embedded within a visually distinct scan region surrounded by peripheral elements such as instrumentation overlays and borders. Our system automatically isolates this primary diagnostic region through a multi-step process (as shown in Figure 10).

First, we detect and mask text elements in the lower third of the image using our OCR pipeline, preventing them from interfering with boundary detection. The algorithm then applies threshold-based binarization

to create a mask highlighting the high-contrast ultrasound region. To reduce noise and small artifacts, we perform morphological erosion using a cross-shaped structuring element (5 iterations), which effectively removes small bright regions while preserving the main ultrasound boundary.

After erosion, we detect contours in the processed image and identify the largest one as the primary ultrasound region. For consistent cropping, we compute the convex hull of this contour and employ a specialized function to identify top edge points within a 20-pixel vertical range from the highest y-coordinate. This approach accommodates the irregular curved upper boundaries commonly found in sector/fan-shaped ultrasound scans while providing stable detection of the rectangular bounds.

The final crop coordinates are determined by the leftmost and rightmost points along the top edge, with the height extending to the lowest point of the convex hull. This methodology provides reliable separation of the diagnostically relevant ultrasound data from surrounding artifacts across diverse equipment vendors and scan types (as shown in Figure 9). Our approach builds upon techniques described by Shamout et al. [8] for handling various ultrasound image shapes and removing surrounding margins in breast ultrasound datasets.

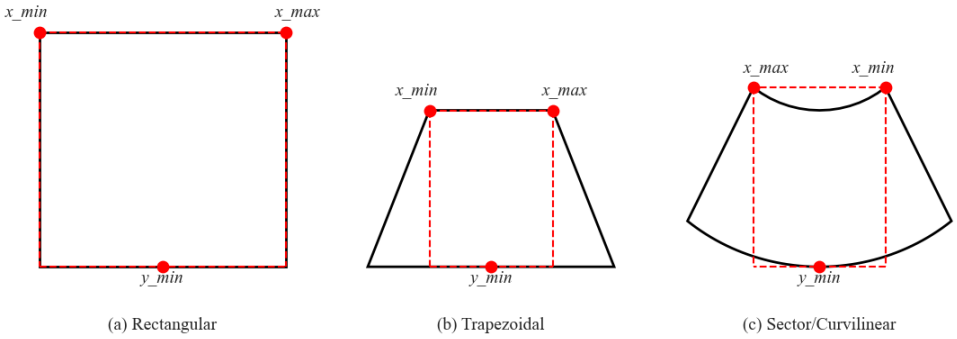


Fig. 9. Common ultrasound imaging geometries: (a) Rectangular, (b) Trapezoidal, and (c) Sector/Curvilinear. Red dots mark key boundary points (x_min, x_max, y_min) that our cropping algorithm detects, with dotted lines showing the resulting rectangular bounding boxes. Methodology adapted from Shamout et al. [8].

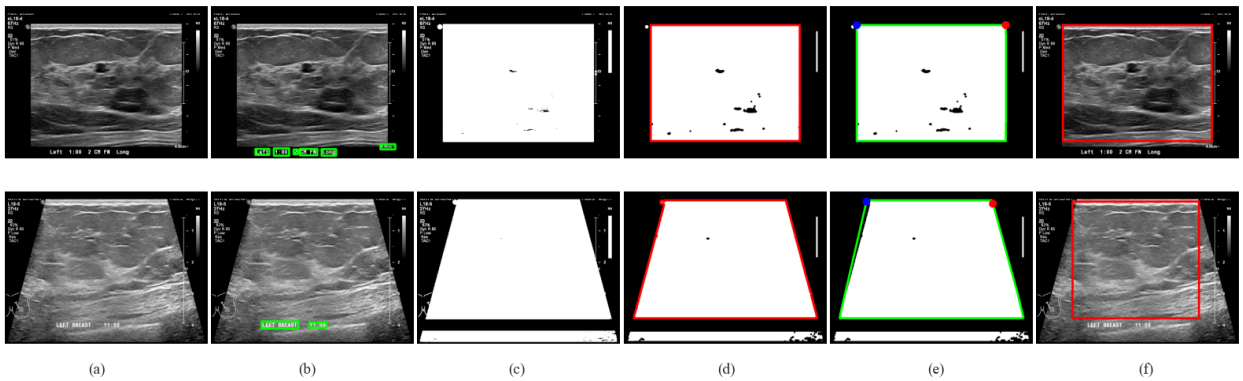


Fig. 10. Image processing pipeline for bus display text extraction: (a) original image, (b) text detection, (c) binary mask with black bar on text, (d) eroded mask with largest contour identification, (e) convex hull with top edge points detection, and (f) final result with extracted region. The pipeline is demonstrated on two different bus images. Methodology adapted from Shamout et al. [8].

4.3 Measurement caliper detection

Measurement calipers are often placed on ultrasound images by sonographers to indicate lesion size. These calipers could potentially create data leakage, as they might inadvertently provide the model with measure-

ment information that correlates with diagnostic outcomes, rather than forcing the model to learn from the tissue characteristics visible in the image.

To address this issue, we developed a custom binary classification model based on a modified ResNet-18 architecture [3] adapted for grayscale ultrasound images. The model was trained on 1,100 manually annotated ultrasound images from diverse equipment vendors, achieving 0.9390 AUC in detecting the presence of measurement calipers.

Rather than excluding caliper-containing images from our dataset, we retain the original images and incorporate caliper presence as an explicit feature in our metadata. This identification also enables selective artifact removal through advanced inpainting techniques described in the following section. We found that 151,568 images (approximately 22.83% of the dataset) contained measurement calipers.

4.4 Finding duplicates

Ultrasound exams frequently contain near-duplicate image pairs, one with measurement calipers and the other without. Identifying these pairs reduces dataset redundancy and eliminates the need for inpainting when clean (no image annotations or markings) versions exist. However, due to clinical workflow variations and human error, many caliper-marked images lack clean duplicates, necessitating inpainting techniques to recover the underlying tissue structures in these cases.

Our duplicate detection algorithm operates by comparing cropped ultrasound regions across images from the same patient study. We then compute pixel-wise differences between each candidate image and all other images from the same patient. A distance metric based on the mean absolute difference between flattened image arrays identifies the most similar image pairs. When this distance falls below an empirically determined threshold of 5.00, the images are flagged as near-duplicates.

For each detected pair, we record the relationship in our database. In particular, when one image contains measurement calipers (as determined in Section 4.3) while its duplicate does not, we can simply use the clean duplicate rather than applying computationally intensive inpainting techniques.

This approach identified 231,030 pairs of duplicate images in our dataset, with 93,148 of caliper-containing images having a corresponding clean duplicate, reducing the need for inpainting.

4.5 Inpainting

For ultrasound images where calipers were detected, an inpainting technique based on a Noise2Noise (N2N) model [9] was applied to remove these artifacts. Through the combined application of our caliper detection and duplicate identification processes (Section 4.3 and Section 4.4), we identified 151,568 images containing measurement calipers. Of these, 93,148 images had corresponding clean duplicates available, leaving 58,420 images that required inpainting due to the absence of unmarked alternatives. This N2N model was trained using a dataset of clean ultrasound images. We first collected and extracted caliper patterns from annotated images to create a library of binary caliper masks (seen in Figure 11).

Fig. 11. Representative subset of extracted caliper patterns.

During training, these extracted caliper masks were procedurally applied to clean images with random transformations (scaling 50-150%, rotation ±45°, and affine skewing) to improve model robustness against variations in caliper appearance. By training the model with noisy input images (clean ultrasound + procedural caliper noise) paired with their corresponding clean targets with its own procedural caliper noise, the

N2N framework learned to differentiate between inherent anatomical structures and superimposed measurement calipers (See Figure 12). The model's inability to predict the random placement of procedural noise forced it to focus on reconstructing the underlying tissue structures, resulting in effective artifact removal while preserving diagnostically relevant features.

Our training dataset consisted of 25,000 ultrasound image pairs, each containing a clean image and its corresponding version with calipers, from which we extracted 16 unique caliper patterns exhibiting diverse shapes and orientations. We observed that certain caliper patterns appeared significantly more frequently in clinical practice than others. By adjusting our procedural caliper placement to reflect these real-world frequency distributions rather than uniform sampling, we improved model performance on validation data. The dataset was split into 80% for training and 20% for validation. On our validation set, the model achieved an average PSNR (Peak Signal-to-Noise Ratio, measuring reconstruction quality) of 28.36 dB and SSIM (Structural Similarity Index Measure, assessing perceptual image similarity) of 0.9866. Visual inspection revealed that the N2N approach produced noticeably cleaner reconstructions compared to traditional inpainting techniques, with little to no visible artifacts in the processed images.

(a) (b) (c)

Fig. 12. Inpainting results demonstrating the Noise2Noise approach: (a) Input ultrasound image with random caliper placement, (b) Target image with random caliper placement, (c) Resulting inpainted image with calipers removed while preserving tissue structures.

4.6 Signal intensity thresholding

The average "darkness" or pixel intensity distribution within the cropped ultrasound region was calculated. This quantitative measure enabled identification of suboptimal ultrasound captures, including those with insufficient penetration depth or poor contrast. While these darker or technically inadequate images could potentially be retained in the dataset, we opted to exclude them to enhance the signal-to-noise ratio and improve model training.

Through this filtering process, we identified and excluded 13,216 ultrasound images (1.99% of our dataset) that exhibited excessive darkness or inadequate tissue visualization. Figure 13 illustrates the darkness distribution across our dataset, with the quality threshold indicated by a red dashed line and an example of a rejected ultrasound shown in the inset.

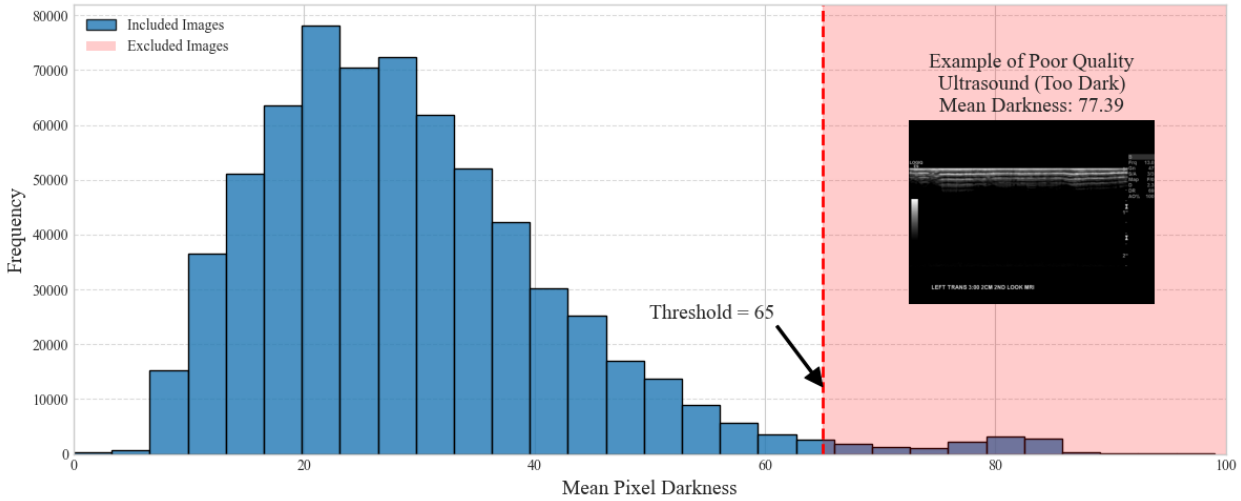


Fig. 13. This figure shows our ultrasound dataset's darkness distribution with the quality threshold (red region). Images exceeding this darkness threshold were excluded to remove inadequate scans with insufficient penetration, optimizing our Multiple Instance Learning approach by improving signal-to-noise ratio. The inset displays an example of a rejected dark ultrasound image (mean darkness: 77.39).

5 Video processing

Processing of ultrasound video sequences shares core methodologies with our image processing pipeline while addressing challenges specific to temporal data. Each video is treated as a sequence of frames, with key processing parameters determined from the initial frame and applied consistently throughout the sequence to maintain coherence.

Processing began by extracting all frames from each DICOM video. The first frame of each sequence underwent our complete detection pipeline to establish parameters for text regions (Section 4.1), cropping boundaries (Section 4.2), and anonymization (Section 3). These parameters were then consistently applied to all subsequent frames to maintain spatial coherence throughout the sequence.

Unlike static ultrasound images, we observed that measurement calipers are not placed on video sequences in clinical practice. This eliminated the need for caliper detection (Section 4.3) and inpainting procedures (Section 4.5) in our video processing pipeline.

To manage data volume and reduce temporal redundancy, only every fourth frame from each video is extracted and retained for the final dataset. This frame sampling approach captures the essential visual information while significantly reducing the overall data size by 75%.

6 Strengths and weaknesses

The methodologies used in creating and processing the breast ultrasound dataset offer significant strengths for AI development. The dataset provides a large, meticulously labeled resource with pathology-verified ground truths, comprising a total of 440,966 images and 55,291 unique breast labels. Preprocessing steps have been implemented to ensure data quality and minimize data leakage, which is crucial for training reliable AI models. These steps include robust anonymization, automated cropping of diagnostic regions, and a novel Noise2Noise inpainting technique to remove measurement calipers.

Despite these advantages, certain weaknesses exist. Imperfections in OCR for text extraction (0.86 F1-score for detection) could lead to metadata errors. While inpainting techniques are advanced (28.36 dB PSNR), they might introduce subtle artifacts or remove fine details. Furthermore, the stringent criteria for label generation, though ensuring high confidence, resulted in a large portion of the initial data remaining unlabeled. Additionally, the final labeled dataset is notably imbalanced, with 86.71% of breasts labeled as benign and only 13.29% as malignant, which may require class-balancing strategies during model training.

The image and video processing and data handling techniques detailed are not limited to breast ultrasound. These methods can be adapted for diagnostic ultrasound of other organ systems like the thyroid, liver, or kidneys, where similar challenges in standardizing image data and extracting relevant information for AI analysis are prevalent. This adaptability makes the described framework a blueprint for broader applications in medical ultrasound imaging.

7 References

1. D'Orsi CJ, Sickles EA, Mendelson EB, Morris EA, et al. ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. Reston, VA, American College of Radiology; 2013.
2. Dworkin M. Recommendation for Block Cipher Modes of Operation: Methods for Format-Preserving Encryption. NIST Special Publication 800-38G. National Institute of Standards and Technology, U.S. Department of Commerce, 2016. Available at: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-38G.pdf
3. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. Available at: https://arxiv.org/abs/1512.03385
4. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. arXiv preprint arXiv:1602.07360, 2016. Available at: https://arxiv.org/abs/1602.07360
5. JaidedAI. EasyOCR: Ready-to-use OCR with 80+ supported languages. GitHub repository, 2020. Available at: https://github.com/JaidedAI/EasyOCR
6. National Electrical Manufacturers Association. Digital Imaging and Communications in Medicine (DICOM) Standard. NEMA PS3 / ISO 12052, 2024. Available at: https://www.dicomstandard.org/current
7. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. Available at: https://arxiv.org/abs/1506.01497
8. Shamout F, Shen Y, Witowski J, Oliver J, Kannan K, Wu N, Park J, Reig B, Moy L, Heacock L, Geras KJ. The NYU Breast Ultrasound Dataset v1.0. Technical report, 2021. Available at: https://cs.nyu.edu/~kgeras/reports/ultrasound_datav1.0.pdf
9. Zhang Y, Jiang N, Xie Z, Cao J, Teng Y. Ultrasonic Image's Annotation Removal: A Self-supervised Noise2Noise Approach. *IEEE Transactions on Computational Imaging*, 2023. arXiv preprint arXiv:2307.04133. Available at: https://arxiv.org/abs/2307.04133