Biomedical Sciences & Health Disparities Symposium: April 20-21, 2023.

Data Standardization of Clinical, Real-World Ultrasound Imaging Data

Arianna Bunnell[1][2], Dustin Valdez[1][2], Thomas Wolfgruber[2], Peter Sadowski[1], John A. Shepherd[2]
University of Hawaii, Honolulu, HI, USA[1], University of Hawaii Cancer Center, Honolulu, HI, USA[2]

Introduction

- Ultrasound (US) is an alternative imaging modality to mammography for detection and diagnosis of breast cancer in resource-limited settings.
- Clinical US images must be preprocessed before applying Artificial Intelligence (AI) methods.
- We clean and filter breast US data from the Hawai`i Pacific Islands Mammography Registry (HIPIMR) for AI cancer detection and risk prediction.
- We develop our method on over 100,000 breast US scans and provide performance statistics on a hand-labeled, random, held-out subset of 2,000 HIPIMR images.
- The goals of our scan standardization procedure are:
  1. Crop boundary scans to remove irrelevant parts of the image.
  2. Detect and remove scans with Color Doppler highlighting.
  3. Detect and remove scans with lesion markers for training AI detection algorithms.

Methods

Scan Cropping

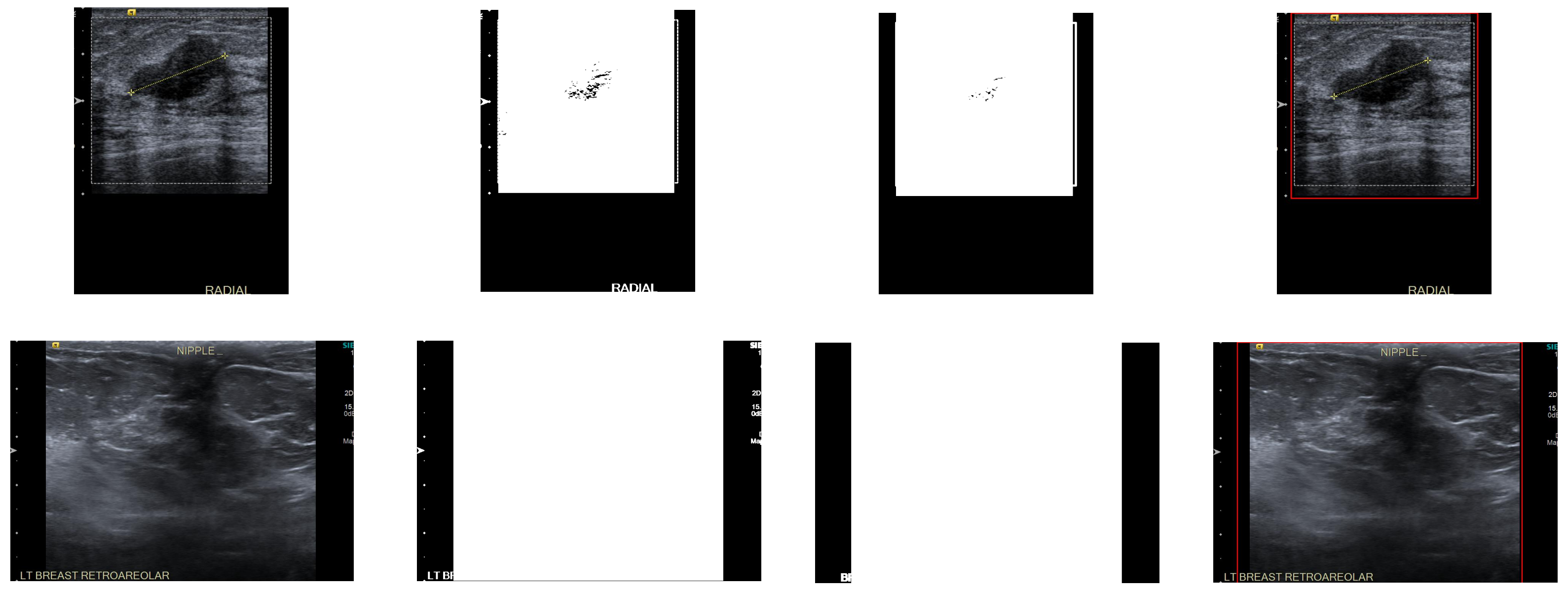
1. Crop scan according to the SequenceOfUltrasoundRegions DICOM value.
2. Apply a color mask for the most common pixel value.
3. Perform *n* rounds of binary erosion and dilation.
4. Identify largest connected component (LCC).
5. Crop the scan to the LCC bounding box [1]. See **Figure 1** for reference.
6. Crop the height of the scan according to the median white pixel in $\left[y_{top}, \frac{1}{3}h + y_{top}\right)$, $\left[\frac{1}{3}h + y_{top}, \frac{2}{3}h + y_{top}\right)$, and $\left[\frac{2}{3}h + y_{top}, y_{bottom}\right]$
7. Crop the width of the scan according to the median white pixel in $\left[x_{left}, \frac{1}{3}w + x_{left}\right)$, $\left[\frac{1}{3}w + x_{left}, \frac{2}{3}w + x_{left}\right)$, and $\left[\frac{2}{3}w + x_{;eft}, x_{right}\right]$. See **Figure 2** for reference.

Color Doppler Highlighting (Figure 3)

1. Does the scan have a PulseRepetitionFrequency value?
2. Apply color mask for white tones and extract the largest contour.
   a) Does the contour have 4 vertices?
3. Apply color masks for red, orange, green, and blue tones.
   a) Is more than 5% of the scan masked?

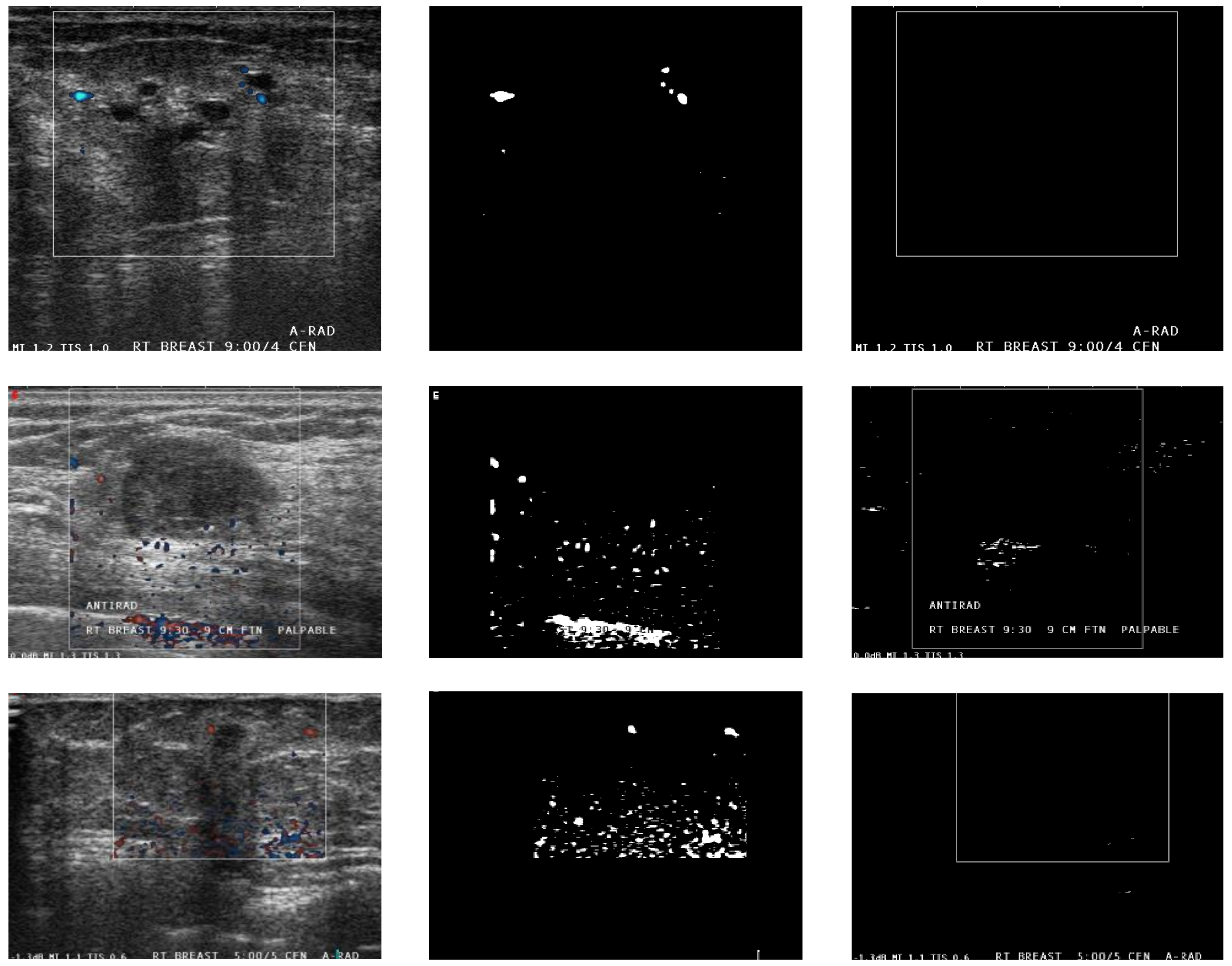
Lesion Markers (Figure 4)

1. Apply color masks for green, yellow, white, and blue tones.
2. Center crop the mask to exclude software artifacts.
3. Perform a single round of binary dilation.
4. Extract contours.
   a) Does the contour have between 14 and 17 vertices?
   b) Is the contour between 10 and 20 pixels?
5. If more than two markers are detected, exclude the scan.



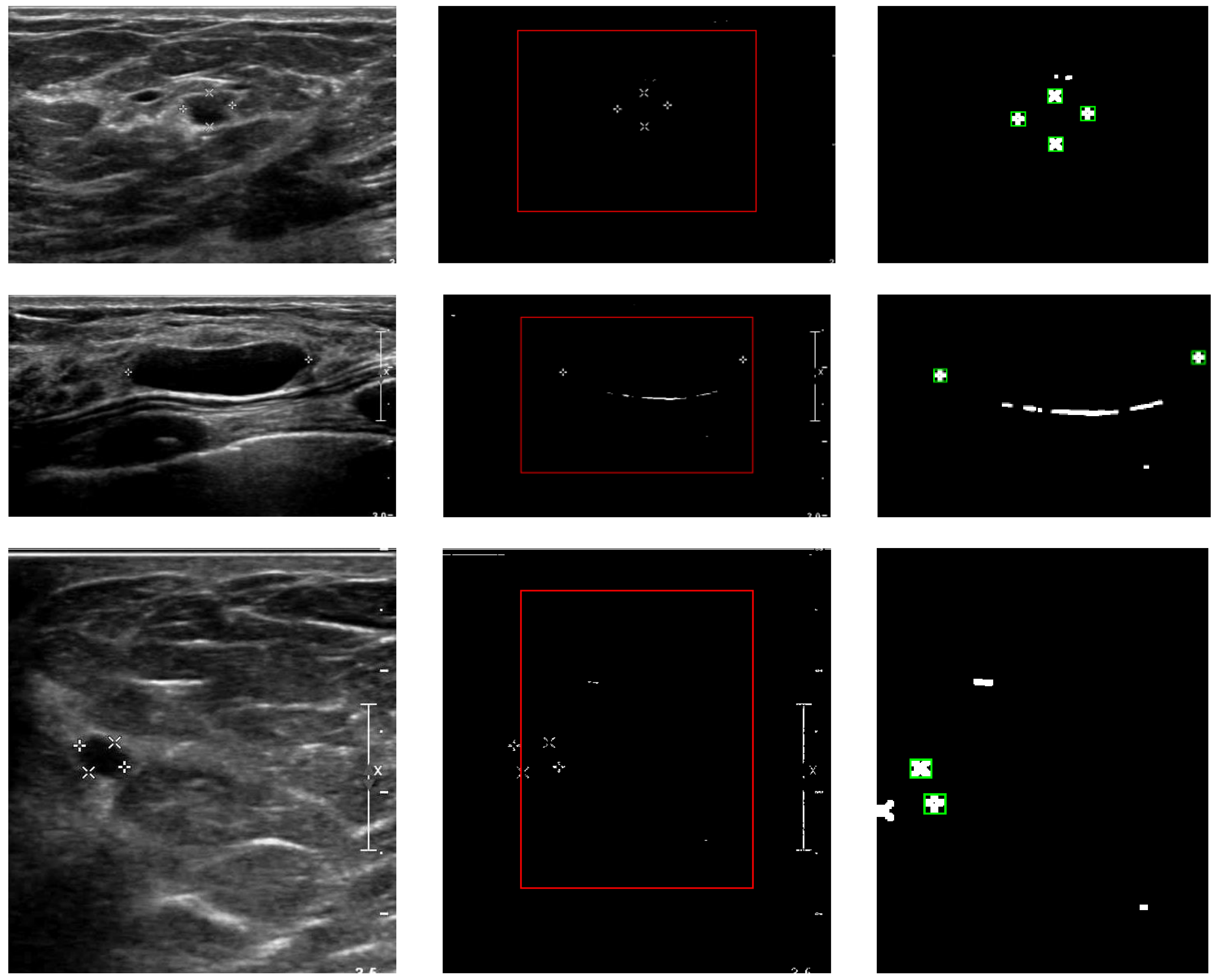
Original Scan   Binary Mask   LCC   Cropped Boundary

Figure 1: Illustration of the first cropping procedure. The leftmost column shows images as they were extracted from the HIPIMR. The second column shows masks based on thresholding mode-valued pixels. The third column shows the largest connected component of the scan. The rightmost column shows the cropping boundary.



Original Scan   Blood Flow Mask   Artifact Mask

Figure 3: Illustration of the types of HSV color masks used for identifying Color Doppler scans. The leftmost column shows cropped images from the HIPIMR. The middle column shows the mask for red, orange, green, and blue tones. The rightmost column shows the mask for white tones.



Original Scan   Lesion Marker Mask   Marker Contours

Figure 4: Illustration of the process used for identifying scans with lesion markers. The leftmost column shows cropped images from the HIPIMR. The middle column shows the mask for green, yellow, white, and blue tones as well as the cropping boundary. The rightmost column shows the contours identified as markers.

Results

From the 4,202 women included in the complete HIPIMR dataset, there were a total of 114,210 breast US scans. Scans with artifacts were identified and removed, including:

- 2,347 elastography and Color Doppler scans
- 17,046 scans with lesion markers

These exclusions resulted in a dataset of 94,817 B-mode ultrasound scans.

The cleaning pipeline was verified on the hand-labeled, randomly selected performance validation subset of 2,000 breast ultrasound scans. 348 scans in the dataset had lesion markers, with 301 flagged correctly and 47 missed (87% sensitivity and 100% specificity). 3 scans were incorrectly flagged as having Color Doppler highlighting. These scans mistakenly identified had colorful text overlaying the scan area, obstructing view of the breast tissue.

Table 1: Confusion matrix showing the predicted and ground truth scan artifact counts in the hand-labeled, random, held-out subset.

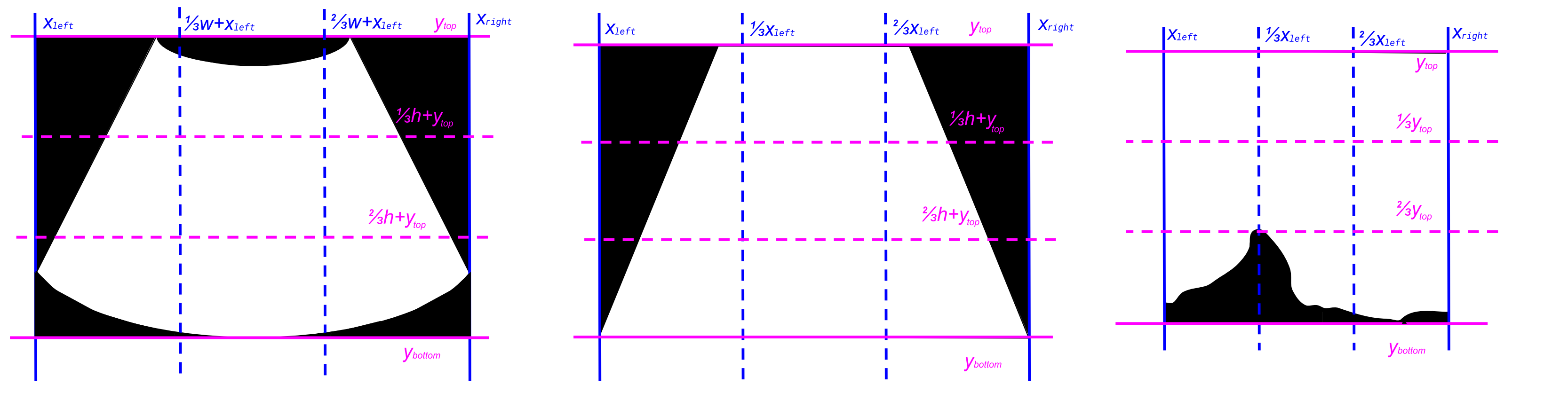
		True Classifications				
		Lesion Markers	Color Doppler	Text Annotation	Unenhanced	Total
Predicted Classifications	Lesion Markers	301	0	0	0	301
	Color Doppler	0	0	3	0	3
	Unenhanced	47	0	242	1,407	1,696
	Total	348	0	245	1,407	2,000

Conclusion

These results demonstrate the efficacy of our breast ultrasound scan cleaning pipeline. Errors in the cleaning pipeline, such as erroneous inclusion of scans with lesion highlighting, add noise to our AI system training. Future work involves further refinement of the pipeline to include other scan artifacts, such as software overlays and operator notes, which may continue to confuse AI model performance.

References

[1] Shamout FE, Shen Y, Witowski JS, Oliver JR, Kannan K, Wu N, Park J, Beatriu, Reig, Moy L, Heacock L, Geras KJ, editors. The NYU Breast Ultrasound Dataset v1.02021.



Convex   Trapezoidal   Irregular

Figure 2: Illustration of the coordinate system used for defining our scan shape-based cropping procedure for convex (leftmost), trapezoidal (middle), and irregular (rightmost) scans. Pink and blue lines represent the borders of our horizontal and vertical image slices used to determine median pixel values, respectively.