

Annotation Parsing in Clinical, Real-World Breast Ultrasound Imaging Data

Kailee Hung¹, Arianna Bunnell¹², Peter Sadowski¹, John Shepherd²
University of Hawaii, Honolulu, HI, USA¹, University of Hawaii Cancer Center, Honolulu, HI, USA²



Introduction

- Ultrasound (US) is a viable imaging modality to mammography for the detection of breast cancer in resource-limited settings.
- Clinical US images often contain annotations from examining sonographers that contain information about scanning protocol and conditions.
- Identification of sonographer text annotations may aid in data cleaning for Artificial Intelligence (AI).
- Text annotations in clinical US images is often partially cut off due to Protected Health Information (PHI) removal protocol.
- The goal of this research is to establish a pipeline for identifying and parsing annotations in clinical breast US scans.

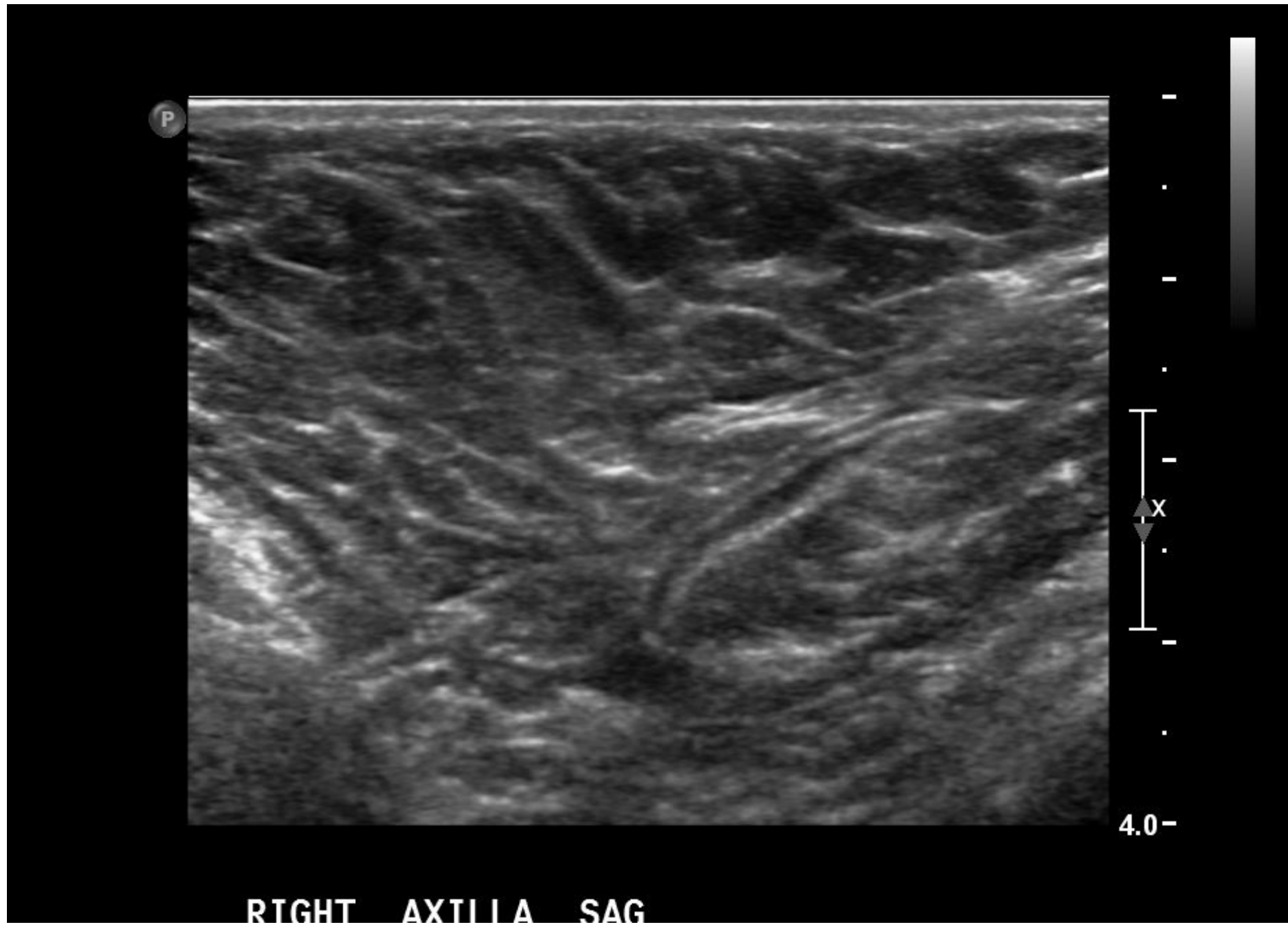
Methods

Text extraction methods were developed through observation of a set of over 100,000 breast US images from the Hawaii & Pacific Islands Mammography Registry (HIPIMR).

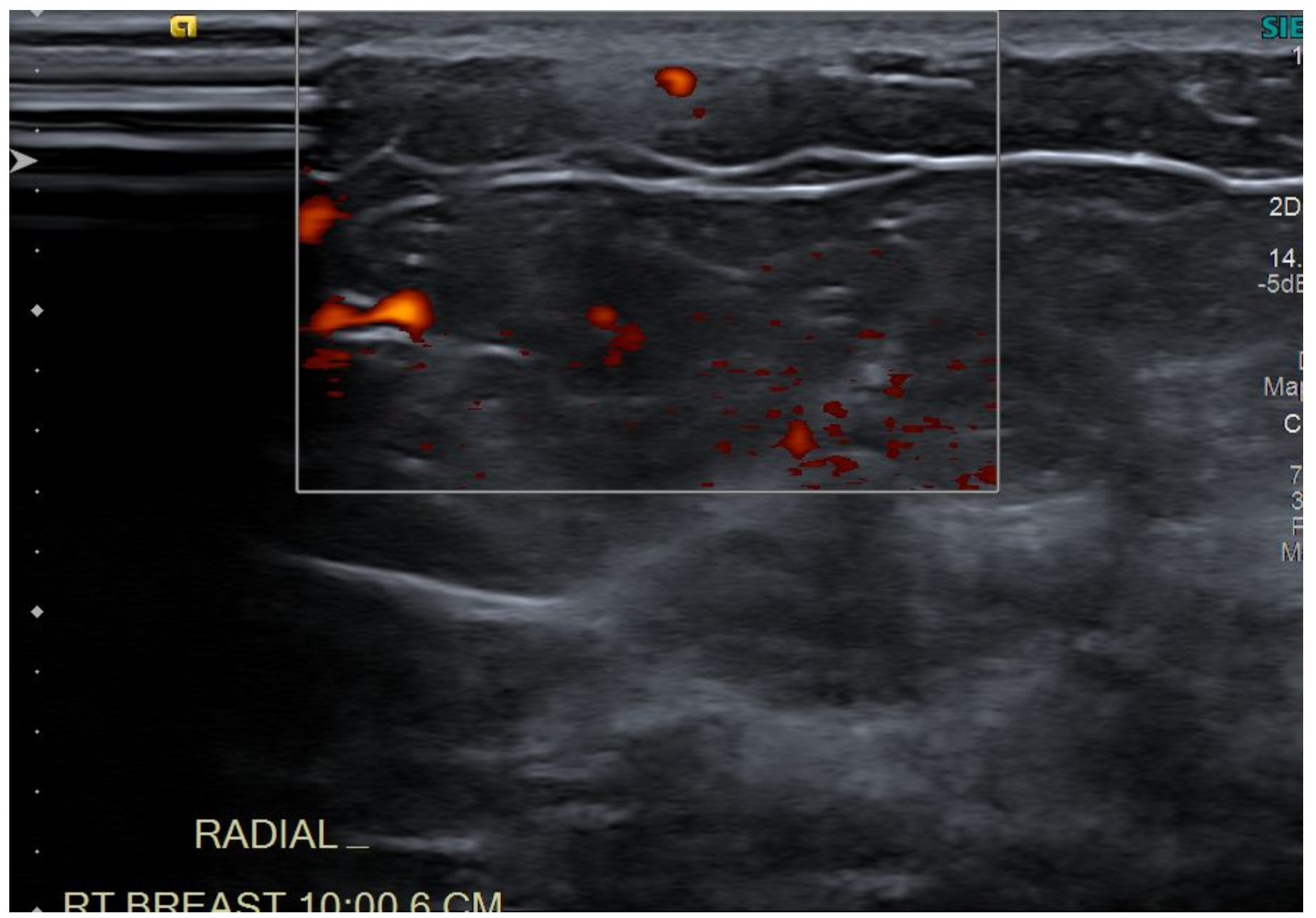
1. Black padding of 70 pixels was added to the bottom of the image to aid in identifying cut-off characters.
2. The EasyOCR (Jaidev AI; Bangkok, Thailand) optical character recognition tool was applied to each scan.

Sonographer annotations describing the structured scanning protocol was categorized into **5 types**: **laterality**, **axilla presence**, **transducer orientation**, **distance from nipple (CMFN)**, **clock position**. A variety of regex patterns were also employed to account for the variety of expected text and incomplete text.

1. For each of the 5 categories: check for the pattern in the detected text from the image. See **Figure 2** for reference.
2. If there is a match, reformat the text as needed. See **Figure 1** for reference.
3. Remove the matched pattern.
4. After checking for all the patterns, the remaining text is classified as miscellaneous.



4.0 - RtGHT AXTI 4 SAG
[RIGHT, SAGITTAL, AXILLA, , , 4.0 -]



9 SII 2D 14. ~SdE Mal C 1 RADIAL RT RRFAST 10*nn 6 CM
[RIGHT, RADIAL, , , , 14. ~sde mal c 10*nn]



4.5. 241 cmeft Breast 1.00 4 CM FN 4 Rad
[LEFT, ANTIRADIAL, , 4 CMFN, 1:00, 4.5. 241]



AL RT BR 9:00 8.5 CM FN ARADI
[RIGHT, , , 5 CMFN, 9:00, 8. aradi]

Figure 2: Visualization of how text was read and parsed into meaningful categories. Under each ultrasound image, the first line is the raw string read by EasyOCR and color coded depending on type of text. The second line is the return array with the text formatted and with the same color coding

Figure 1: ACR Labeling and Measurement Standards For Ultrasound [1]

Labeling for breast US images may contain the following descriptive fields:

- **Laterality:** Designation of left or right breast being examined
 - Rt, Right > RIGHT Lt, Left > LEFT
- **Axilla:** Refers to the armpit region and indicates examination of lymph nodes
 - Axilla, Axillary > AXILLA
- **Transducer Orientation:** Angle that the ultrasound transducer is positioned
 - Rad > RADIAL Arad, Antirad > ANTIRADIAL Sag > SAGITTAL
 - Trans, Trns, Trv > TRANSVERSE
- **CMFN:** Distance from the nipple to the abnormality or the area being scanned in cm
 - 8 cmfn, 8cm fn > 8 CMFN 7-8cmfn > 7-8 CMFN
- **Clock Position:** Anatomic location using clock-face notation
 - 7:00, 7o'clock > 7:00

Results

Table 1. Sensitivity and Specificity of Fine-Tuned EasyOCR on Held-Out Test Set

	Laterality	Axilla	Transducer Orientation	CMFN	Clock Position
Sensitivity (# True Positives)	96.85% (1,632)	97.44% (190)	97.05% (1,351)	93.99% (720)	93.58% (1,167)
Specificity (# False Positives)	100.00% (0)	100.00% (0)	99.34% (4)	97.24% (34)	99.87% (1)
Image Count (% of Images)	1,685 (84.25%)	195 (9.75%)	1,396 (69.80%)	800 (40.00%)	1,248 (62.40%)

The text annotation extraction pipeline was validated on a randomly-selected, hand-labeled subset of 2,000 breast US images from the HIPIMR dataset. Generally, some failures could be attributed to basic scanning errors both due to text cropping or text cursor presence. Other failures were due to oversights in the code for circumstances unaccounted for (cm/n, ftn, fn). Additionally, clock position was excluded if there were two instances leading to false negatives.

Conclusion

These results show the efficacy of our domain-specific text recognition pipeline and may improve breast US data for AI model development.

Improvements

- Further refinement of the pipeline, namely in CMFN, and clock position when it comes to handling multiple values (6:00-7:00) and cm/n, ftn, fn.
- For laterality, significant improvement would be seen after accounting for the "+ breast" pattern.

Future Developments

- Bounding box coordinates were returned, so a system needs to be developed to crop text out of the images.
- Detection and removal of lesion annotations.
- Plans to release code via an open-source license for research use.

References

1. CJ DO, EA S, EB M, Morris EA, al. e. ACR BI-RADS[®] Atlas, Breast Imaging Reporting and Data System. Reston, VA: American College of Radiology; 2013.



Rreast 2.00 Akad MTD
[, , , 2:00, akad mtd]



3.5. RIGHT BREAST 4:00 5-7 CMFN TRV 2.37 cm
[RIGHT, TRANSVERSE, , 5-7 CMFN, 4:00, 3.5.]

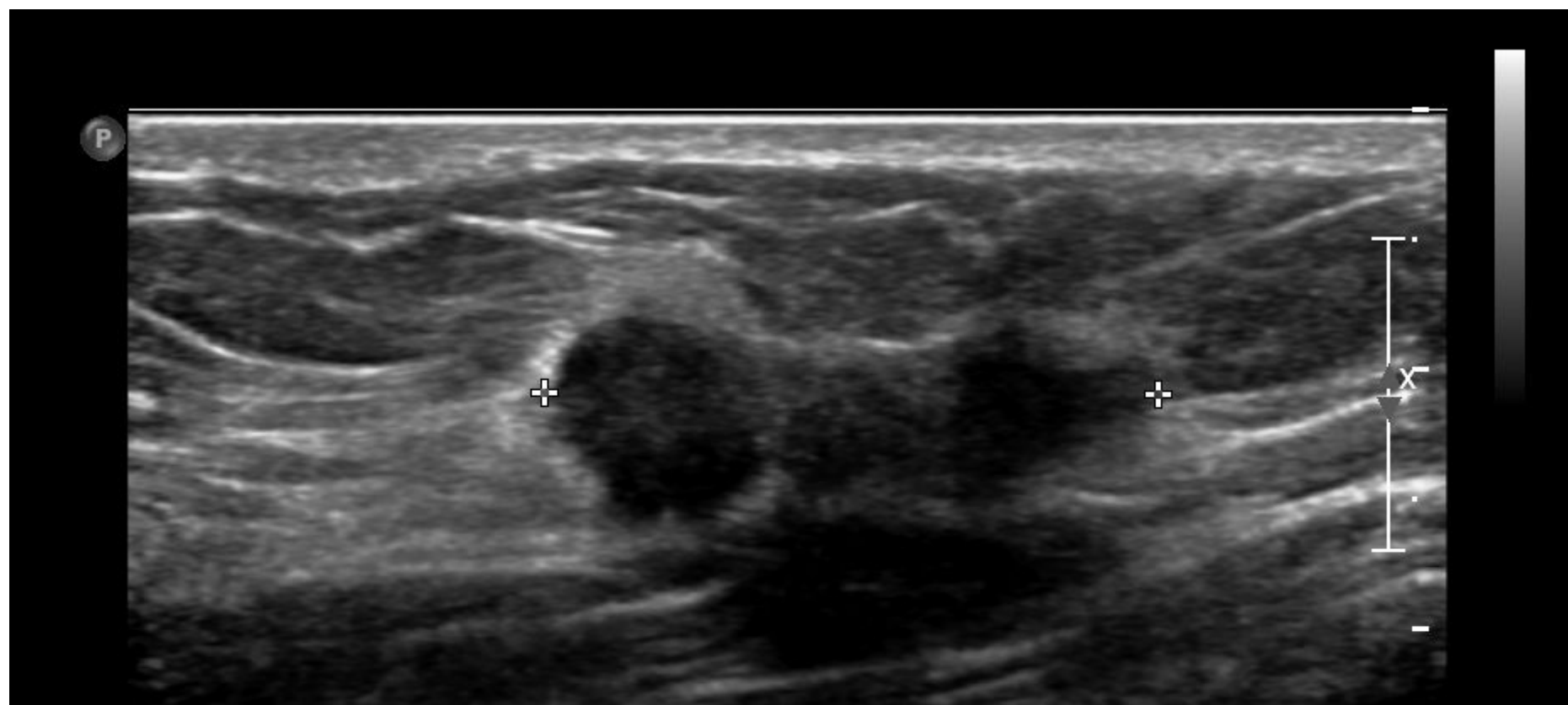


3.0- Left Breast 6 0'clock 3 CMFN Rad
[LEFT, RADIAL, , 3 CMFN, 6:00, 3.0-]

258f6d670522bdace89b8ab56cd9c8bbc1da7aab.png
50fbc2780e8fd65f681619fd60b6cca139d3bc27.png
0df96da1395afbd9980dadf1444919ad2df8b9e1.png

b00c19ab9fc1aef42a313dfe7d50de451bfd4848.png

1d4a0c63343602928b2022e744b56d83a7e140f6.png
1474d7b6c9e2a2f9fc2502da1d79945dc46e2781.png
3233777d24201a081c298a7bf0701fc46ca0b880.png



Data Standardization of Clinical, Real-World Ultrasound Imaging Data

Arianna Bunnell^{1,2}, Dustin Valdez^{1,2}, Thomas Wolfgruber², Peter Sadowski¹, John A. Shepherd²
University of Hawaii, Honolulu, HI, USA¹, University of Hawaii Cancer Center, Honolulu, HI, USA²



ARTIFICIAL INTELLIGENCE
PRECISION HEALTH INSTITUTE
UNIVERSITY OF HAWAII

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 $[y_{top}, \frac{1}{3}h + y_{top}]$, $[\frac{1}{3}h + y_{top}, \frac{2}{3}h + y_{top}]$, and $[\frac{2}{3}h + y_{top}, y_{bottom}]$.
7. Crop the width of the scan according to the median white pixel in $[x_{left}, \frac{1}{3}w + x_{left}]$, $[\frac{1}{3}w + x_{left}, \frac{2}{3}w + x_{left}]$, and $[\frac{2}{3}w + x_{left}, x_{right}]$. 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.

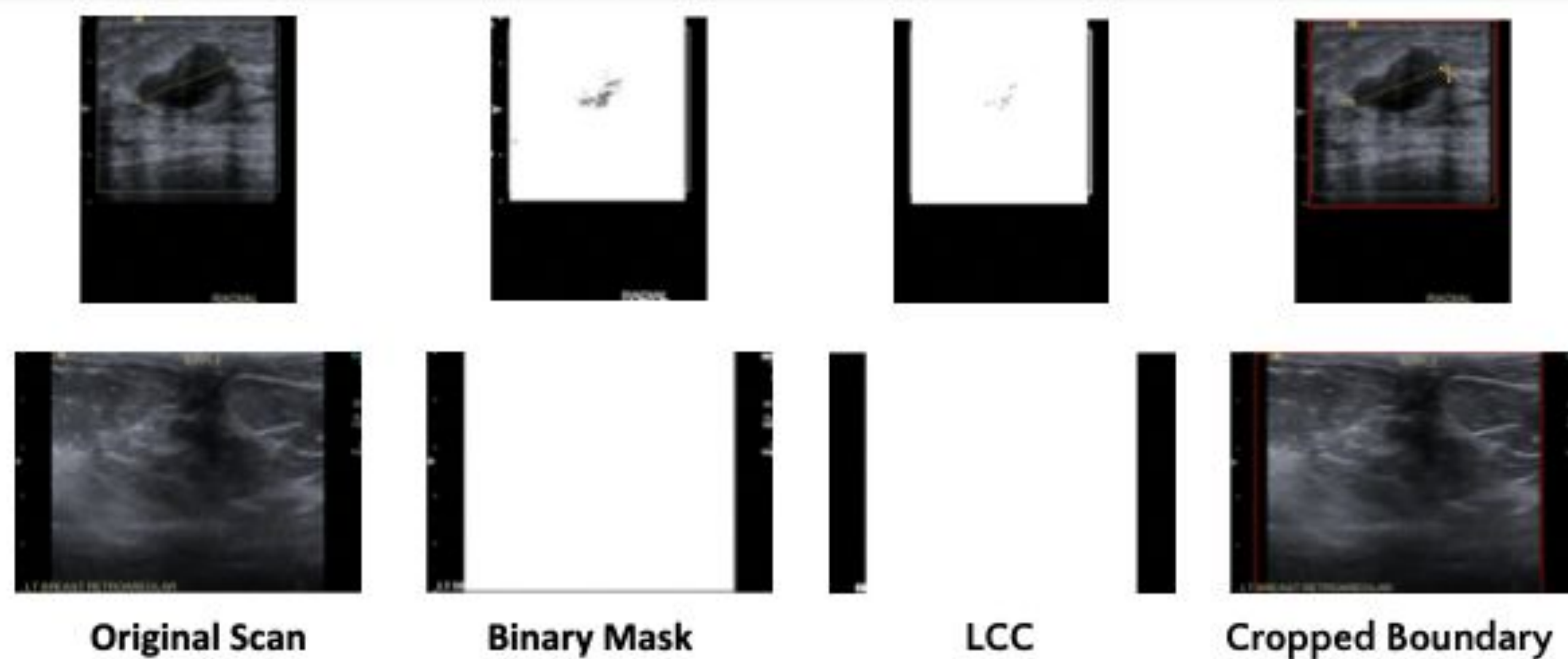


Figure 1: Illustration of the first cropping procedure. The leftmost column shows images as they were extracted from the HIPIMR. The second column shows masks based on thresholding mode-valued pixels. The third column shows the largest connected component of the scan. The rightmost column shows the cropping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

Anything that is flagged as Doppler is Doppler, but there are a bunch of scans which are not Doppler which we need to find

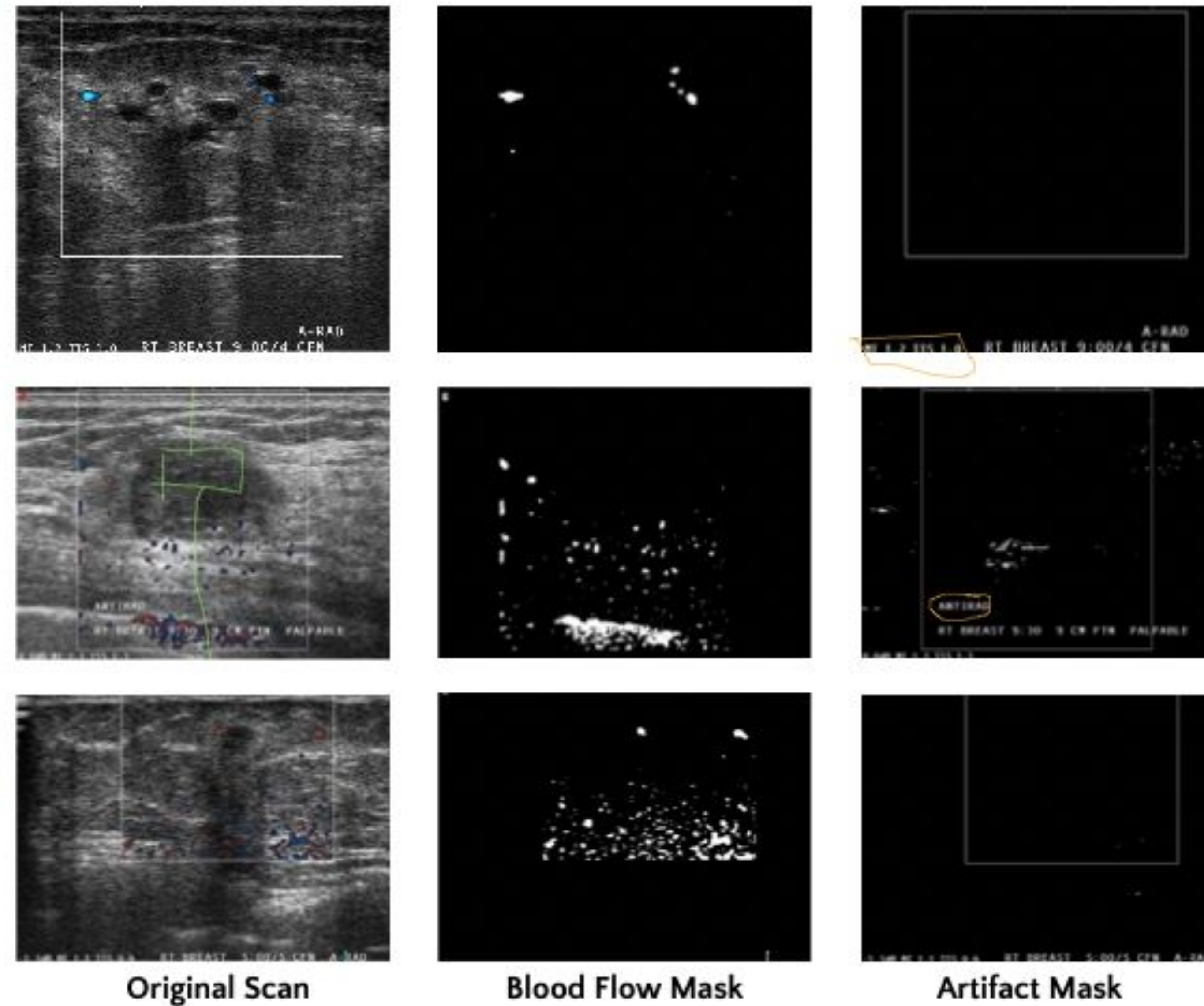


Figure 3: Illustration of the types of HSV color masks used for identifying Color Doppler scans. The leftmost column shows cropped images from the HIPIMR. The middle column shows the mask for red, orange, green, and blue tones. The rightmost column shows the mask for white tones.

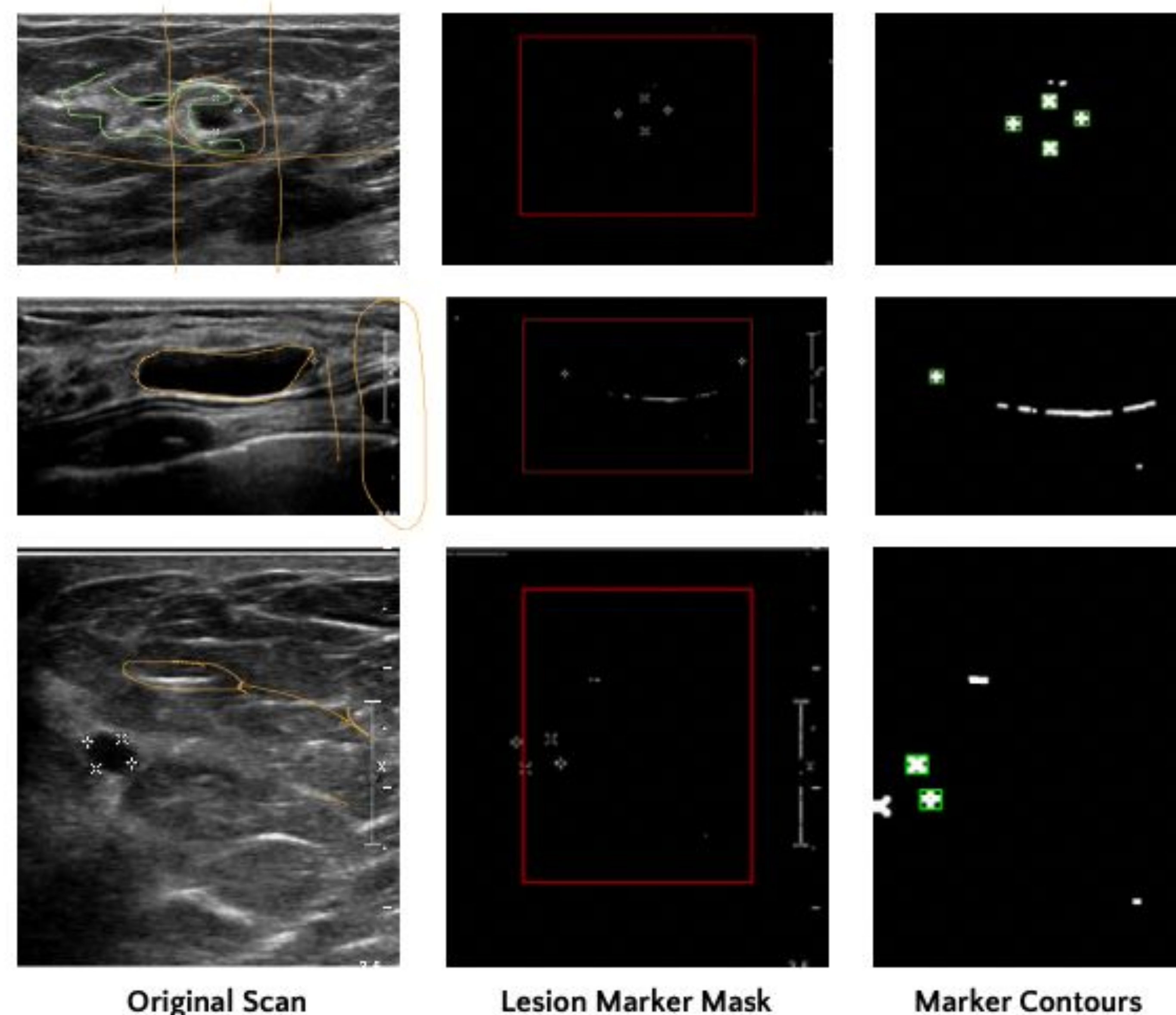


Figure 4: Illustration of the process used for identifying scans with lesion markers. The leftmost column shows cropped images from the HIPIMR. The middle column shows the mask for green, yellow, white, and blue tones as well as the cropping boundary. The rightmost column shows the contours identified 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			
Predicted Classifications		Lesion Markers	Color Doppler	Text Annotation	Unenhanced
	Lesion Markers	301	0	0	0
	Color Doppler	0	0	3	0
	Unenhanced	47	0	242	1,407
	Total	348	0	245	1,407
		Total			
		301	0	3	1,407

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, Beatru, Reig, Moy L, Heacock L, Geras KJ, editors. The NYU Breast Ultrasound Dataset v1.02021.

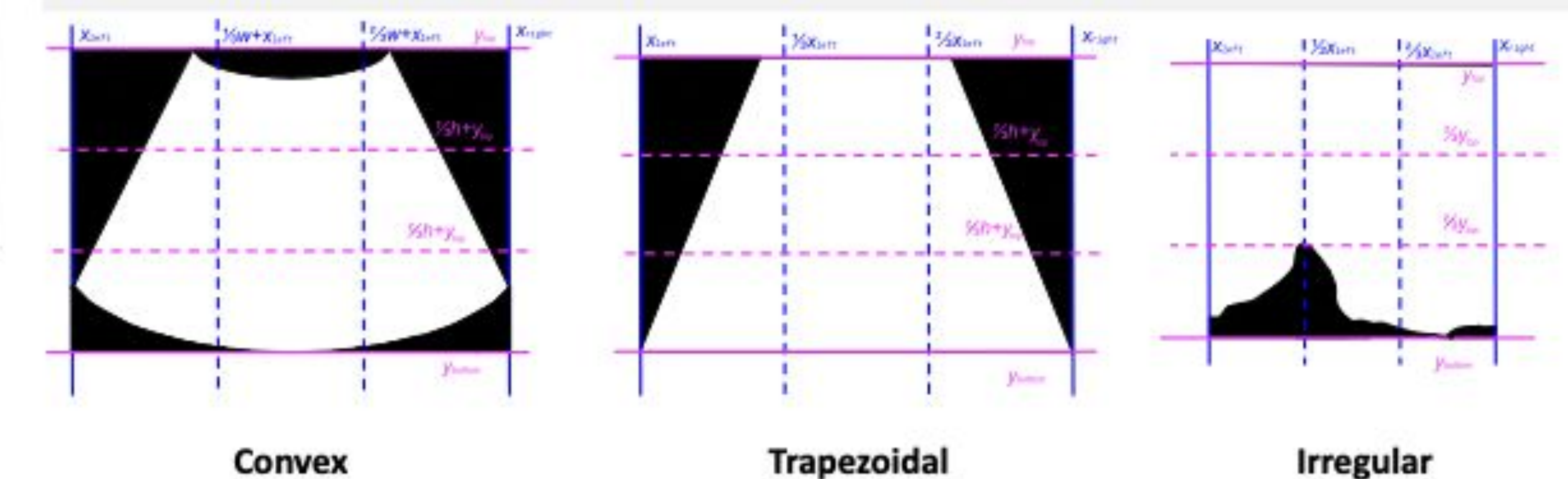


Figure 2: Illustration of the coordinate system used for defining our scan shape-based cropping procedure for convex (leftmost), trapezoidal (middle), and irregular (rightmost) scans. Pink and blue lines represent the borders of our horizontal and vertical image slices used to determine median pixel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

A pilot study to evaluate breast cancer screening in low resource areas of the Pacific using portable ultrasound and artificial intelligence

Dustin Valdez¹, Arianna Bunnell¹, Thomas Wolfgruber¹, Jami Fukui¹, Peter Sadowski¹, John A. Shepherd¹
University of Hawaii Cancer Center, Honolulu, HI, USA¹

Background

- Advanced stage breast cancer (stages III and IV) rates in the Pacific are much higher than in the USA mainland, especially where mammography services do not exist: Marshall Islands (61%), Palau (94%), and Samoa (79%). Little is known about the performance of portable ultrasound when compared to clinical ultrasound for use in breast cancer screening.
- Introducing a portable handheld ultrasound (US) system coupled with an artificial intelligence (AI) detection algorithm and operated by a trained healthcare worker may be able to reduce advanced stage cancer rates in areas where mammography is not available.
- We developed an AI model using breast ultrasound images from the Hawaii Pacific Island Mammography Registry as well as a reader study tool.
- In this preliminary study, we report the preliminary reader study results from two graduate students (surrogate health worker).

Methods

- A case-control study design was used to minimize variability and data selection was done at both the patient and scan level. The matching was 1:3 by birth year (113 cases to 339 controls). **Figure 1** shows a CONSORT style flowchart of our data selection process.
- The ultrasound images were then stripped of PHI and had to be preprocessed (scan area cropping, removal of burnt-in annotations, and splitting of dual-view scans) to be compatible for AI model training. The model selected for our transfer learning was DenseNet121 with a 70%-20%-10% split for training-validation-testing.
- The reader study interface used a VGG Image Annotator software [1] that we modified for our reader study. An example case can be seen in **Figure 2**.
- The preliminary reader study was run with a random sample of 400 breast ultrasound images comprised of 301 benign and 99 malignant scans from a publicly available breast ultrasound dataset [2].
- Readers were asked to provide a BIRADS score for each image using the reader tool.
- Sensitivity and specificity analysis was done based on a BIRADS 4a threshold (2% likelihood of malignancy).

Results

Figure 1. CONSORT style flowchart showing the data selection process at both the patient and scan levels. The red section presents patient-level selection, and the blue section represents the 4812 breast ultrasound scans collected from the 452 case-control patients. After all data selection criteria were met, 3871 total breast ultrasound scans were used for model training and evaluation.

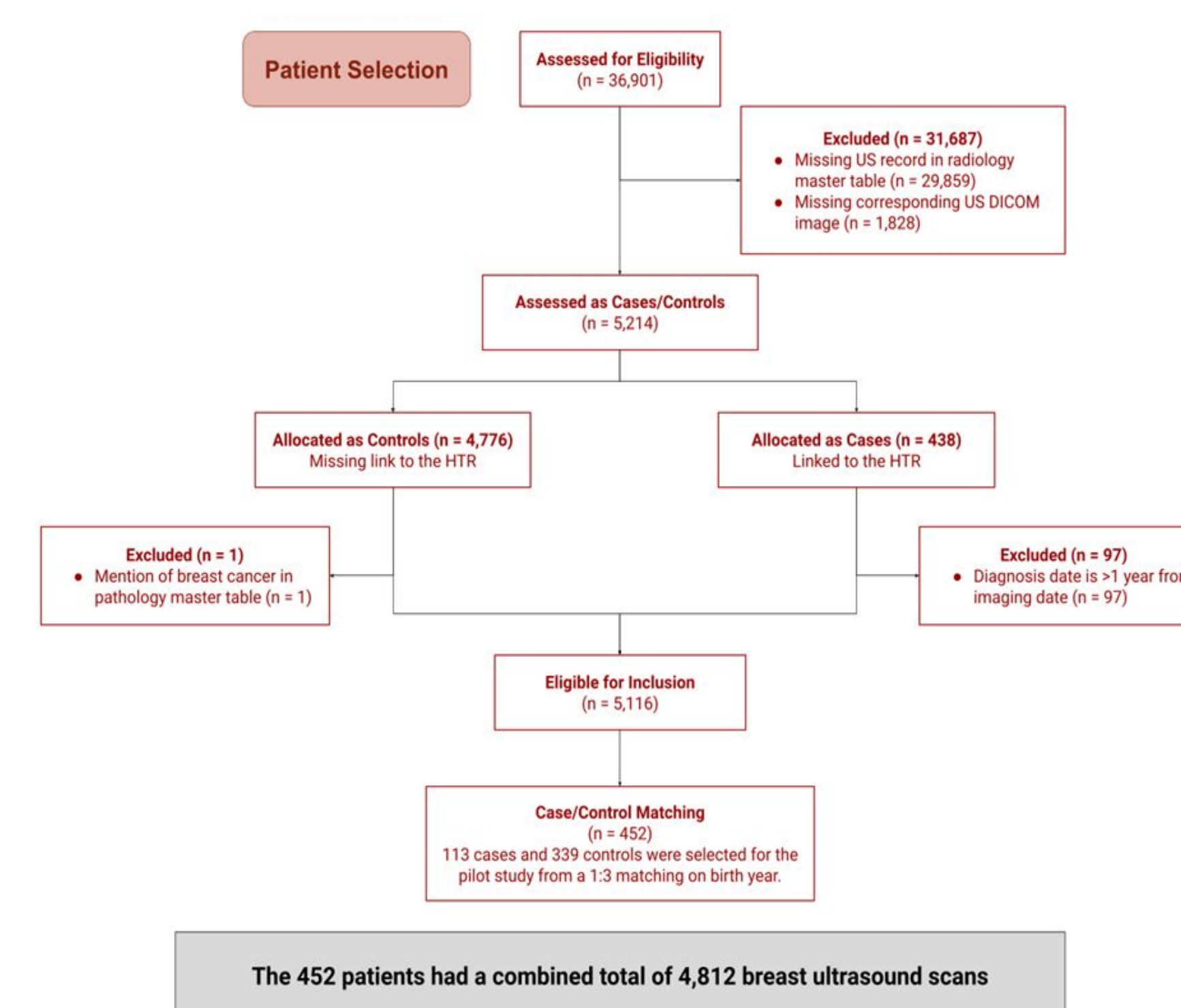


Figure 2. Reader User Interface. Screen capture of the reader user interface from the VGG VIA image annotator tool. Highlighted in blue is the AI output section presented as a percentage likelihood of malignancy and a translated BIRADS score. Highlighted in purple is where the reader indicates the mass-specific characteristics from the ultrasound lexicon.

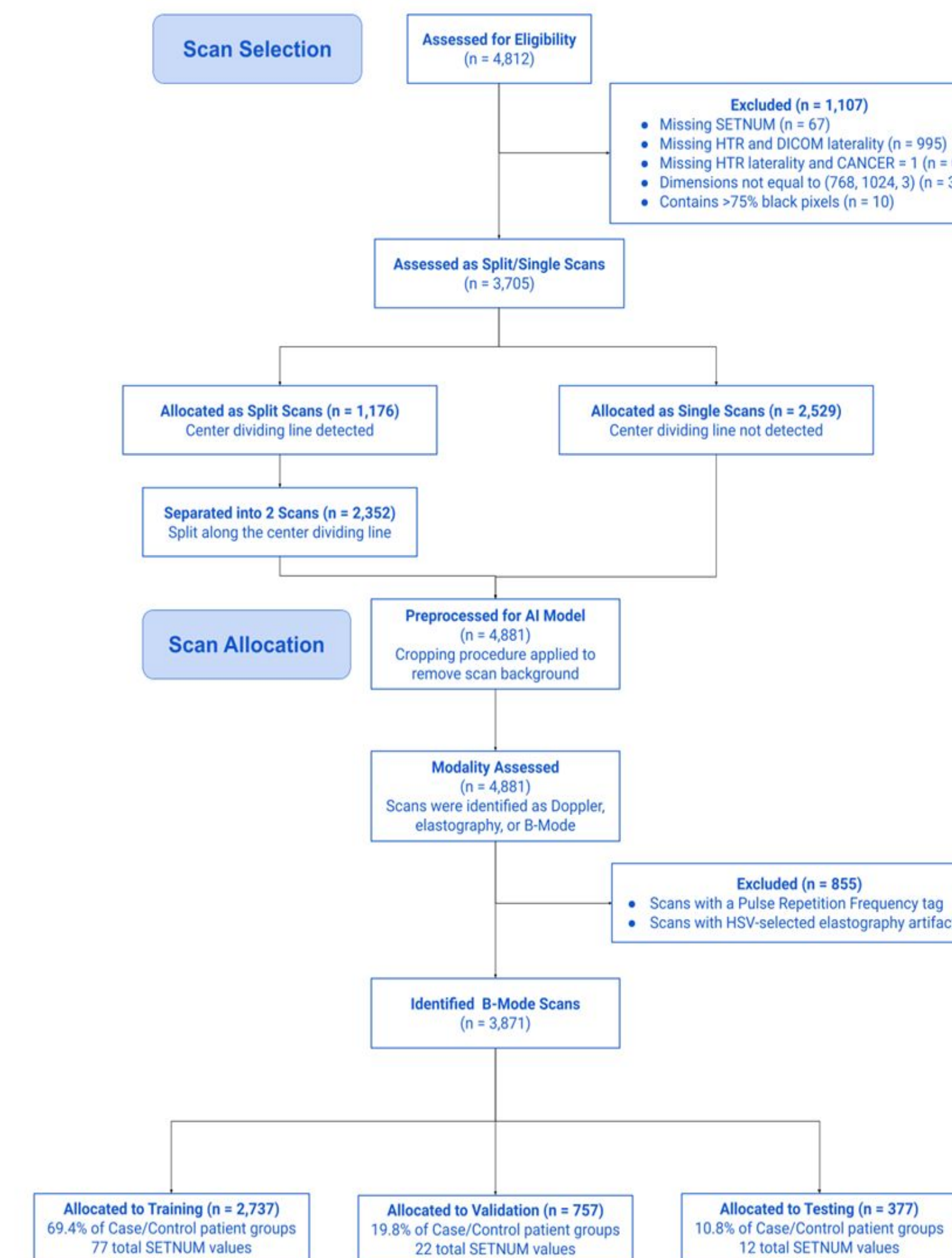
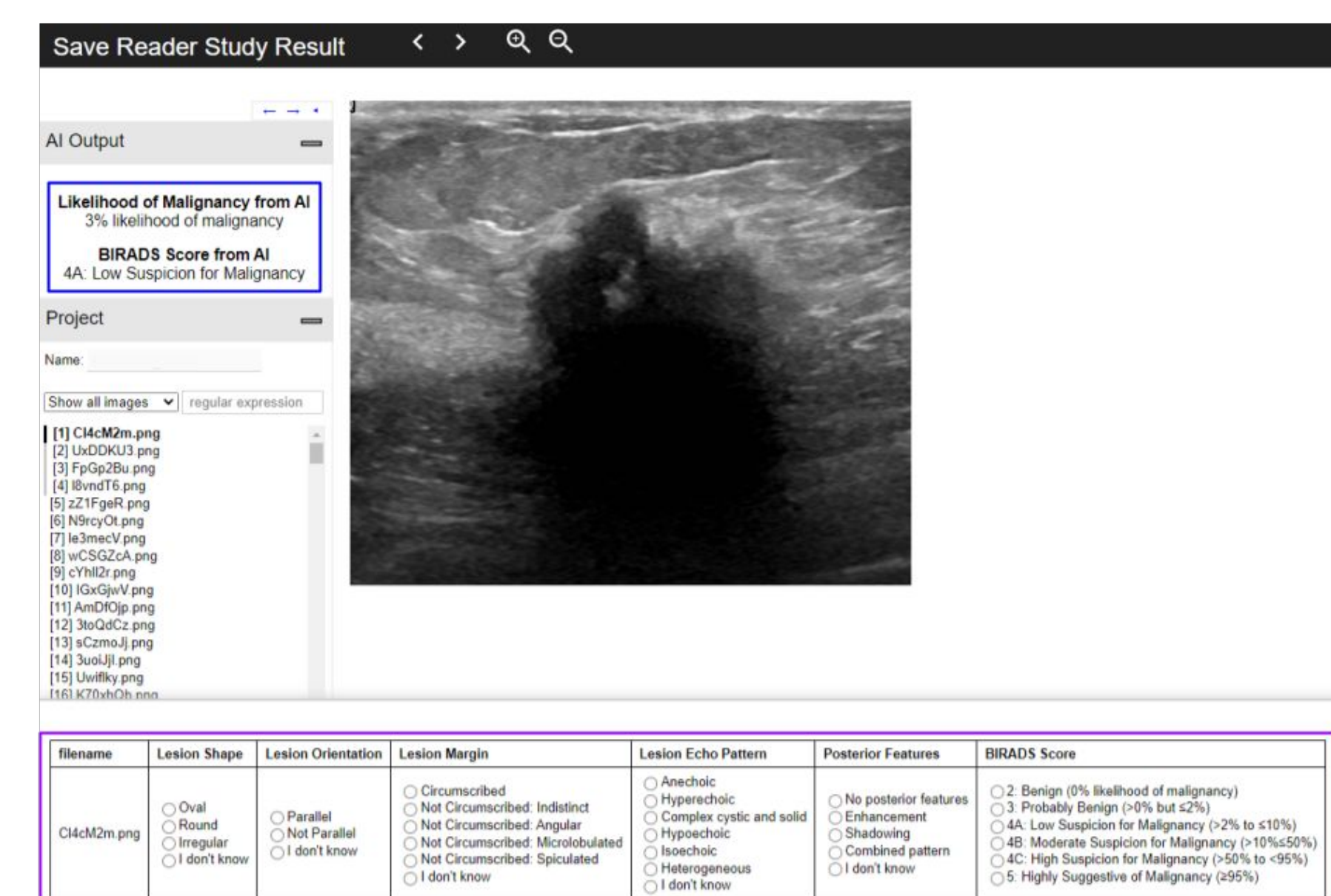
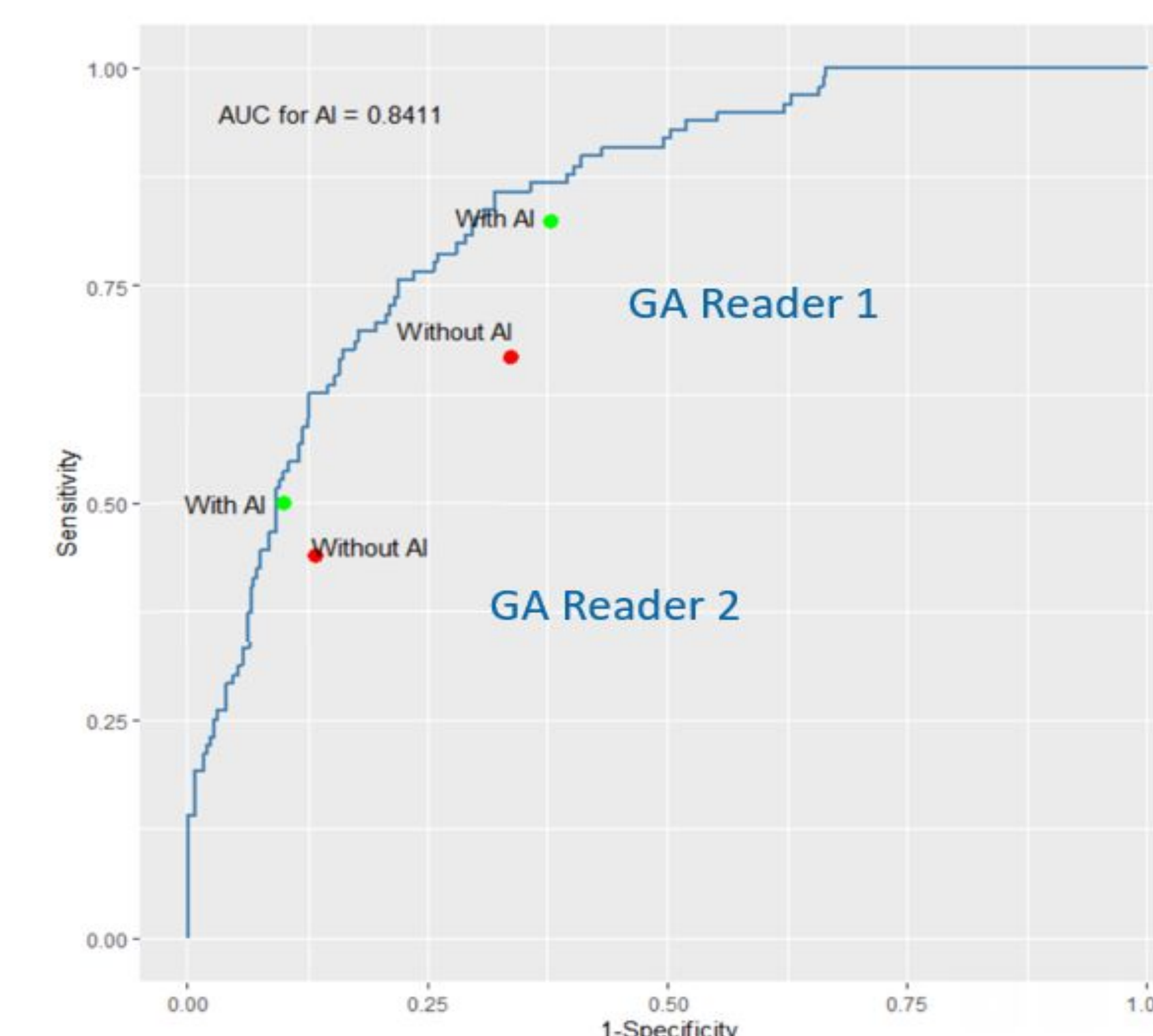


Figure 3. Reader Study Preliminary Results. ROC for the AI model is shown as the blue line. The readers Cancer/No Cancer classifications were determined from a BIRADS score of at least 4a for cancer (corresponding to a 2% likelihood of malignancy).



Results (continued)

- The AI model had an AUC=0.826 on the validation set and AUC=0.776 on the internal held-out test data set. On the external public breast ultrasound data set used for the preliminary reader study the model had an AUC=0.8411.
- The average sensitivity saw an improvement of 0.165 with AI, but a small decrease in specificity of 0.04. The full sensitivity and specificity results with and without AI can be seen in **Table 1**.

Table 1. Reader Study Sensitivity and Specificity Results.

	Sensitivity (Without AI)	Sensitivity (With AI)	Sensitivity Difference	Specificity (Without AI)	Specificity (With AI)	Specificity Difference
Reader 1	0.7	0.83	0.13	0.638	0.601	-0.037
Reader 2	0.667	0.867	0.2	0.664	0.621	-0.043
Averages	0.684	0.849	0.165	0.651	0.611	-0.04

Conclusion

- The preliminary reader results are very promising as both readers saw noticeable gains in sensitivity while maintaining only minor losses in specificity. There is potential for AI to be used with ultrasound for non-radiologist healthcare workers, but a full reader study still needs to be done to show if AI can improve performance of a healthcare worker to that of a radiologist.

Future Work

- We are working on improving our AI model before performing a full reader study where different types of readers (radiologist, MDs, and general healthcare workers) are asked to assign a BIRADS score to breast ultrasound images with and without the aid of an AI system.
- Currently working with a breast radiologist to improve our training materials for non-radiologist readers.

References

1. Abhishek Dutta and Andrew Zisserman. 2019. The VIA Annotation Software for Images, Audio and Video 2019, Nice, France. <https://doi.org/10.1145/3343031.3350535>.
2. Al-Dhabyani, W., et al., *Dataset of breast ultrasound images. Data in brief*, 2019. 28: p. 104863-104863.

Identifying Linguistic Anti-Patterns in Data Science Projects: An Exploratory Study

Carol Wong, Micah Tilton

Mentor: Dr. Anthony Peruma



Undergraduate Research Opportunities Program

Office of the Vice Provost for Research & Scholarship
University of Hawai'i at Mānoa



INFORMATION & COMPUTER SCIENCES

UNIVERSITY of HAWAI'I at MĀNOA

Abstract

Similar to traditional/industrial software applications built by professional software developers, code written by data scientists is also susceptible to bad programming practices that negatively impact system quality and maintenance activities. While previous research has examined code quality from the point of coding standards, there is a gap in understanding the importance of identifier naming, including variables, functions, and classes. In particular, there is limited knowledge regarding the linguistic antipatterns (LA) that data scientists often unknowingly inject into their code. The goal of this research study is to identify naming violations in data science code; this includes identifying LAs specific to data science code. To this extent, we perform a grounded theory-based approach analyzing identifier names within 125 open-source data science Jupyter Notebooks. As part of our initial findings, we propose a taxonomy of 6 LAs. At a high level, we observe that many LAs arise from identifiers not conveying enough information to describe a variable's information or the intended purpose of a function. These issues are commonly observed with identifiers that heavily use abbreviations or documentation that conflicts with implementation. Our results serve as a valuable foundation for creating tools and IDE plugins that provide data scientists with real-time identifier name appraisals and recommendations. Additionally, these results can guide revisions to the existing curriculum, enabling future data scientists to write high-quality and maintenance-friendly code.

Introduction

Data science plays a crucial role in influencing decision-making and innovation across various industries, from banking and healthcare to retail and social media. A key success of the field is the ability of data scientists to build and execute code-based algorithms that translate raw data into meaningful insights.

Code quality - Similar to traditional software applications, data science code is also subject to poor programming practices by data scientists. High-quality code should be consistent, well-documented, tested, and easy for others to understand. Such code is essential to ensure the project's maintainability and facilitate collaboration between team members.

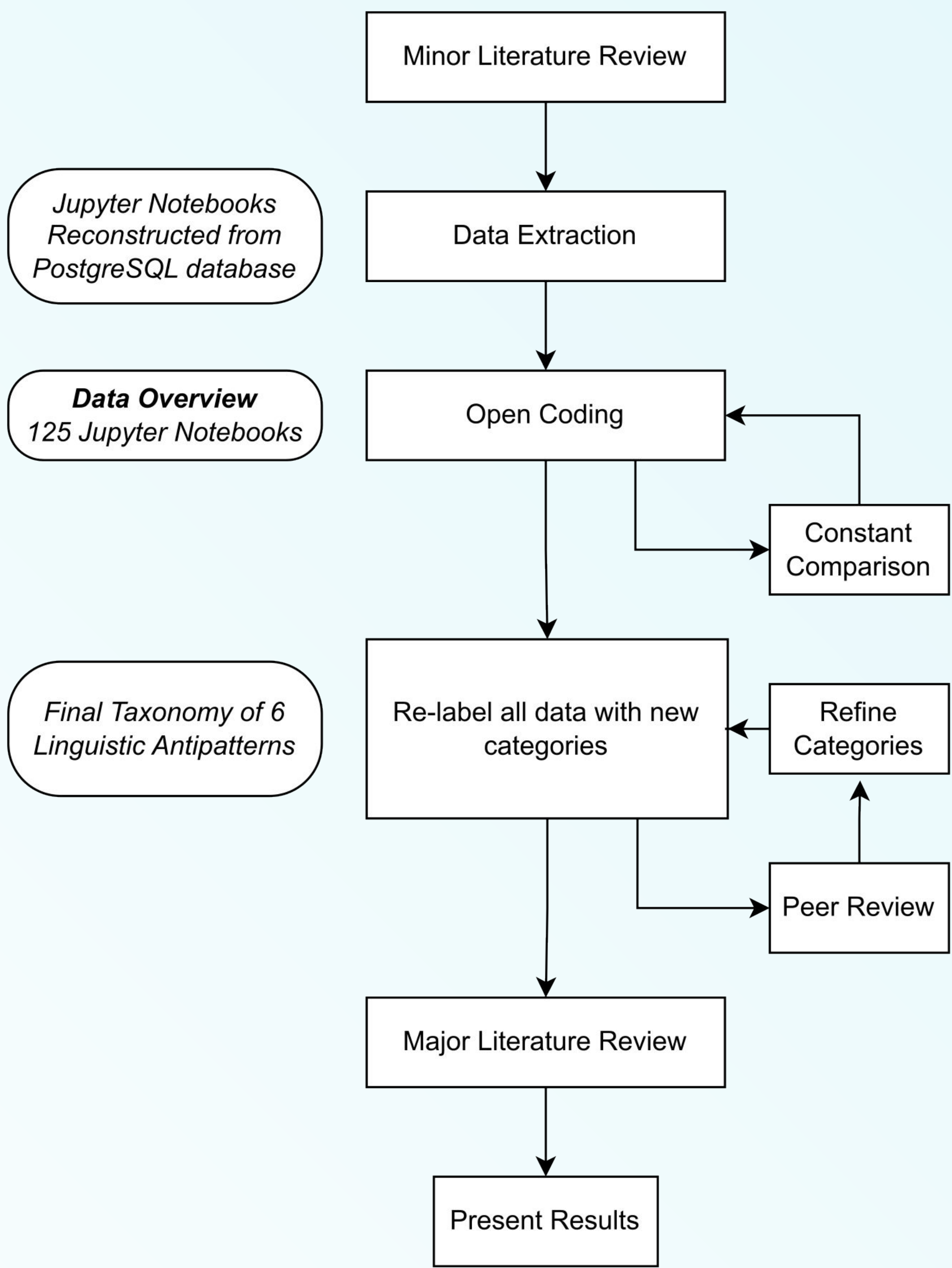
Linguistic Antipatterns - A key activity for any software maintenance task is reading and understanding the code. While code may seem readable, it might not reflect its intended behavior. Knowingly or unknowingly, code is written using identifiers (i.e., class, method, variables, etc.) with low-quality names. Such instances of poor programming practices are known as Linguistic Antipatterns (LA) and are known to impede software maintenance activities.

Motivation

Unlike traditional software, data science code is frequently written by data scientists from diverse backgrounds who might not necessarily have the knowledge of coding best practices. Therefore, this wide array of coding practices may unknowingly introduce LAs; some of which are unique to data science code. This study aims to address the research gap by identifying LAs in data science code. The study expands the field of program comprehension by extending the catalog of known LAs.

Methodology

Grounded-Theory Based Approach



Results

Inconsistent Information: Function/Variable operates differently than identifier implies

Code-Comment: Code inconsistent with comment

Example:

```
# Randomly generate adjacency matrix representation
def generateGraph(nPoints, dimension=2):
```

Comment notes that method will randomly generate adjacency matrix, but method name is much vaguer

Proposed Fix:

```
# Randomly generate adjacency matrix representation
def generateRandomAdjacencyGraph(nPoints, dimension=2):
```

Code-Purpose: Identifier implies code performs different functions than it actually does

Example:

```
def round_up(number):
    return np.ceil(number / 10.) * 10.
```

*Expected Purpose: Function rounds up to nearest integer
Actual Purpose: Function rounds up to nearest multiple of 10*

Proposed Fix:

```
def round_up_tens(number):
    return np.ceil(number / 10.) * 10.
```

Inferred Different Type: Variable identifier implies that it holds a different type than it actually does

Example:

```
grade = class_size['GRADE'] != '09-12'
```

Expected "grade" to hold an integer but it holds a boolean

Proposed Fix:

```
is_highschool = students['GRADE'] == '09-12'
```

Standard Convention: Identifier conflicts with standard naming conventions of common libraries

Example:

```
df_name = data_file[:-4]
data[df_name] = pd.read_csv('../resources/' + data_file)
```

df is commonly used as an abbreviation for pandas Dataframe, so its use as a data_file here may be confusing

Proposed Fix:

```
data_file_name = data_file[:-4]
data[data_file_name] = pd.read_csv('../resources/' + data_file)
```

Insufficient Information: Cannot infer information from function/variable identifier

Cannot Infer Type: Identifier does not give enough information to be able to infer what type of data within it

Example:

```
sch = {}
```

```
def forward(self, x):
```

```
B1 = torch.randn((1, n_hidden))
```

Identifiers are either abbreviations that are not documented or standard convention, single/double-digit characters, or are vague

Proposed Fixes:

```
record_schedules = {}
```

```
def visualize_forward_pass(self, x):
```

```
hidden_layer_bias = torch.randn((1, n_hidden))
```

Expected Parameter: Function name suggests a parameter should exist but it does not

Example:

```
def is_in_image(shape, px=0, py=0):
    return ((px >= 0) and (px < shape[0]) and (py >= 0) and (py < shape[1]))
```

No image is given in parameters

Proposed Fix:

```
def is_in_image(image, px=0, py=0):
    return ((px >= 0) and (px < image.width) and (py >= 0) and (py < image.height))
```

Next Steps

- Create a plugin for the JetBrains DataSpell IDE to help data scientists identify and correct LAs.
- Conducting a survey with data scientists to confirm and refine our findings
- Guide revisions to the existing curriculum by emphasizing the importance of documentation, implementation and choice of identifiers in creating clean, high-quality data science code.

