Annotation Parsing in Clinical Breast Ultrasound

Kailee Hung
Mentors: Arianna Bunnell, Peter Sadowski

Project Goals

- Develop software to identify and parse useful information from breast ultrasounds
- Aims to aid data cleaning for development of AI in breast cancer detection

Text Identification Challenge

- Detects text in 7 categories with over 90% accuracy
- However, text can be cut off in the image
- Utilized the EasyOCR tool and Regex expressions to match patterns in text and account for variability

Tested on a set of 100,000 images

2.41 cmeft   Breast   1:00 4 CM FN A Rad

4.5. 241 cm**eft** Rreast 1.00 4 CM FN 4 Rad
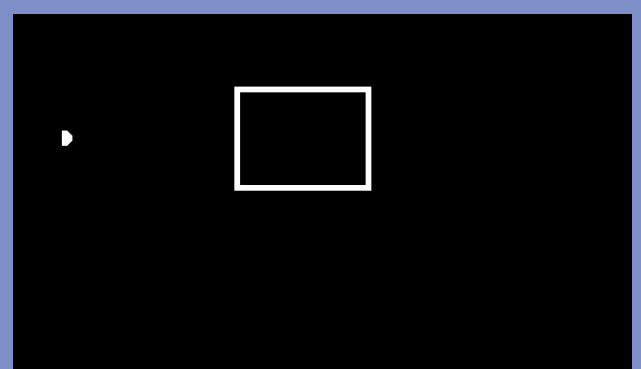[**LEFT**, **ANTIRADIAL**, , **4 CMFN**, **1:00**, 4.5. 241]
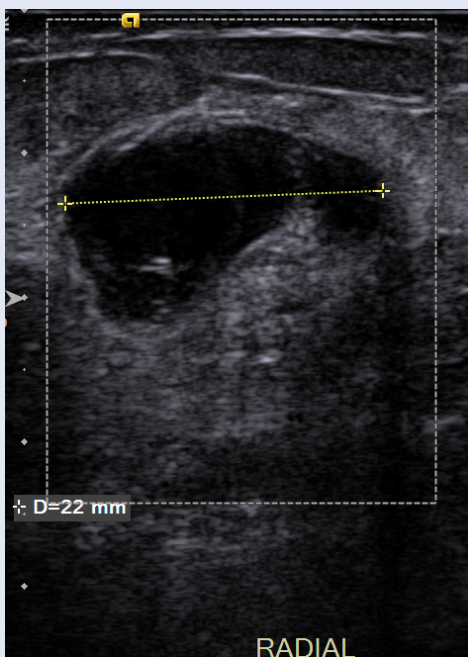
Non-B-Mode Identification

- Non-B-mode scans include those with doppler highlighting and those using ultrasound elastography
- These can be identified with colored highlighting or presence of boxes indicating regions of interest (ROI)
- Used OpenCV library for computer vision tasks
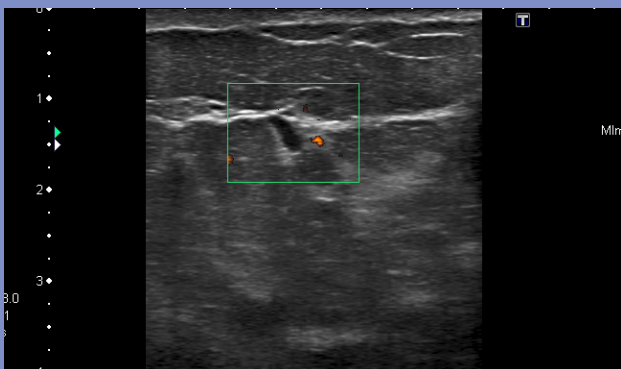
Original Image

Dilated Green Threshold

Methodology

- Split the image based on color thresholds
- For white and green thresholds, dilated the image, then parsed the resulting contours to determine if there was a box
- To check for blood flow highlighting, cropped edges out to eliminate any colored text and detected any remaining color in the image