Developing a Data Standardization Pipeline for Evaluation of Mammography AI in Hawaii

Kanta Saito[1], Wilson Huynh[1], Elijah Saloma[1], Arianna Bunnell[1,2], Peter Sadowski[1], John Shepherd[2]

[1]Information and Computer Sciences, UH Manoa

[2]University of Hawaii Cancer Center

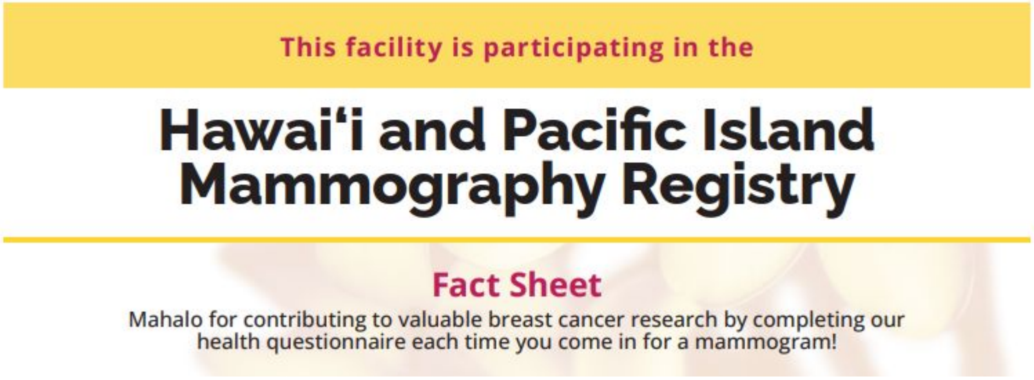
Motivation - Improving Breast Cancer AI for Women in Hawaiʻi

1. Most breast cancer AI is developed on datasets of White women from the continental US and Europe. **In Hawaiʻi, we are powered to look at ANHPI subgroups.**
2. Personal growth: Significant opportunity to use computer vision & AI techniques while actively contributing to the project.

BCSC 5-year risk	White N = 91,308	Filipina N = 6,551	Chinese N = 24,051	Japanese N = 2,485
	% of invasive cancers discovered within 5 years			
<1.67% chance of developing cancer in the next 5 years	54	90	93	89
≥1.67% chance of developing cancer in the next 5 years	46	10	7	11

The Hawai‘i & Pacific Islands Mammography Registry (HIPIMR)

- Prospective cohort of women
- Collects breast imaging and breast health information (2009-present)
- Linked to the Hawai‘i Tumor Registry to identify cases

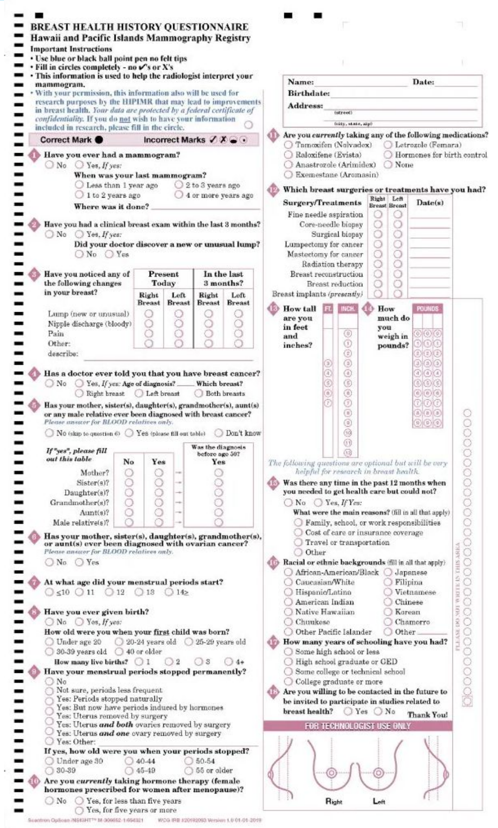


This facility is participating in the

Hawai‘i and Pacific Island Mammography Registry

Fact Sheet

Mahalo for contributing to valuable breast cancer research by completing our health questionnaire each time you come in for a mammogram!



BREAST HEALTH HISTORY QUESTIONNAIRE
Hawaii and Pacific Islands Mammography Registry

Important Instructions
- Use blue or black ball point pen no felt tips
- Fill in circles completely - no ✔'s or X's
- This information is used to help the radiologist interpret your mammogram.
- With your permission, this information also will be used for research purposes by the HIPIMR that may lead to improvements in breast health. *Your data are protected by a federal certificate of confidentiality.* If you do not wish to have your information included in research, please fill in the circle. ○

Correct Mark ● Incorrect Marks ✓ ✗

Name: ____ Date: ____
Birthdate: ____
Address: ____ (street) ____ (city, state, zip)

1. **Have you ever had a mammogram?**
○ No ○ Yes, *If yes:*
When was your last mammogram?
○ Less than 1 year ago ○ 2 to 3 years ago
○ 1 to 2 years ago ○ 4 or more years ago
Where was it done? ____

2. **Have you had a clinical breast exam within the last 3 months?**
○ No ○ Yes, *If yes:*
Did your doctor discover a new or unusual lump?
○ No ○ Yes

3. **Have you noticed any of the following changes in your breast?**

	Present Today		In the last 3 months?	
	Right Breast	Left Breast	Right Breast	Left Breast
Lump (new or unusual)	○	○	○	○
Nipple discharge (bloody)	○	○	○	○
Pain	○	○	○	○
Other:	○	○	○	○

describe: ____

4. **Has a doctor ever told you that you have breast cancer?**
○ No ○ Yes, *If yes:* **Age of diagnosis?** ____ **Which breast?**
○ Right breast ○ Left breast ○ Both breasts

5. **Has your mother, sister(s), daughter(s), grandmother(s), aunt(s) or any male relative ever been diagnosed with breast cancer?**
Please answer for BLOOD relatives only.
○ No (skip to question 6) ○ Yes (please fill out table) ○ Don't know

If "yes", please fill out this table

	No	Yes	Was the diagnosis before age 50? Yes
Mother?	○	○	○
Sister(s)?	○	○	○
Daughter(s)?	○	○	○
Grandmother(s)?	○	○	○
Aunt(s)?	○	○	○
Male relative(s)?	○	○	○

6. **Has your mother, sister(s), daughter(s), grandmother(s), or aunt(s) ever been diagnosed with ovarian cancer?**
Please answer for BLOOD relatives only.
○ No ○ Yes

7. **At what age did your menstrual periods start?**
○ ≤10 ○ 11 ○ 12 ○ 13 ○ 14≥

8. **Have you ever given birth?**
○ No ○ Yes, *If yes:*
How old were you when your first child was born?
○ Under age 20 ○ 20-24 years old ○ 25-29 years old
○ 30-39 years old ○ 40 or older
How many live births? ○ 1 ○ 2 ○ 3 ○ 4+

9. **Have your menstrual periods stopped permanently?**
○ No
○ Not sure, periods less frequent
○ Yes: Periods stopped naturally
○ Yes: But now have periods induced by hormones
○ Yes: Uterus removed by surgery
○ Yes: Uterus *and both* ovaries removed by surgery
○ Yes: Uterus *and one* ovary removed by surgery
○ Yes: Other: ____
If yes, how old were you when your periods stopped?
○ Under age 30 ○ 40-44 ○ 50-54
○ 30-39 ○ 45-49 ○ 55 or older

10. **Are you *currently* taking hormone therapy (female hormones prescribed for women after menopause)?**
○ No ○ Yes, for less than five years
○ Yes, for five years or more

11. **Are you *currently* taking any of the following medications?**
○ Tamoxifen (Nolvadex) ○ Letrozole (Femara)
○ Raloxifene (Evista) ○ Hormones for birth control
○ Anastrozole (Arimidex) ○ None
○ Exemestane (Aromasin)

12. **Which breast surgeries or treatments have you had?**

Surgery/Treatments	Right Breast	Left Breast	Date(s)
Fine needle aspiration	○	○	
Core-needle biopsy	○	○	
Surgical biopsy	○	○	
Lumpectomy for cancer	○	○	
Mastectomy for cancer	○	○	
Radiation therapy	○	○	
Breast reconstruction	○	○	
Breast reduction	○	○	
Breast implants *(presently)*	○	○	

13. **How tall are you in feet and inches?** FT. INCH.

14. **How much do you weigh in pounds?** POUNDS

The following questions are optional but will be very helpful for research in breast health.

15. **Was there any time in the past 12 months when you needed to get health care but could not?**
○ No ○ Yes, *If Yes:*
What were the main reasons? (fill in all that apply)
○ Family, school, or work responsibilities
○ Cost of care or insurance coverage
○ Travel or transportation
○ Other

16. **Racial or ethnic backgrounds** (fill in all that apply)
○ African-American/Black ○ Japanese
○ Caucasian/White ○ Filipina
○ Hispanic/Latina ○ Vietnamese
○ American Indian ○ Chinese
○ Native Hawaiian ○ Korean
○ Chuukese ○ Chamorro
○ Other Pacific Islander ○ Other ____

17. **How many years of schooling have you had?**
○ Some high school or less
○ High school graduate or GED
○ Some college or technical school
○ College graduate or more

18. **Are you willing to be contacted in the future to be invited to participate in studies related to breast health?** ○ Yes ○ No

Thank You!

PLEASE DO NOT WRITE IN THIS AREA

FOR TECHNOLOGIST USE ONLY

Right Left

Modern Breast Radiology & AI

1. Computer-aided diagnosis and detection systems using artificial intelligence (AI) have transformed breast radiology in recent years.
2. **HOWEVER**, clinical imaging data often contains *necessary* artifacts that hinder accurate algorithmic evaluation.

Big Ideas & Significance

✅ Large repository of *clinical* mammogram images for **ANHPI**

❌ Data cleaning → ❌ Reliable verification of AI models with diverse data sets

REAL LIVES DEPEND ON THIS!!!

Research Question & Hypothesis

How can we address this lack of standardization in HIPIMR mammograms?

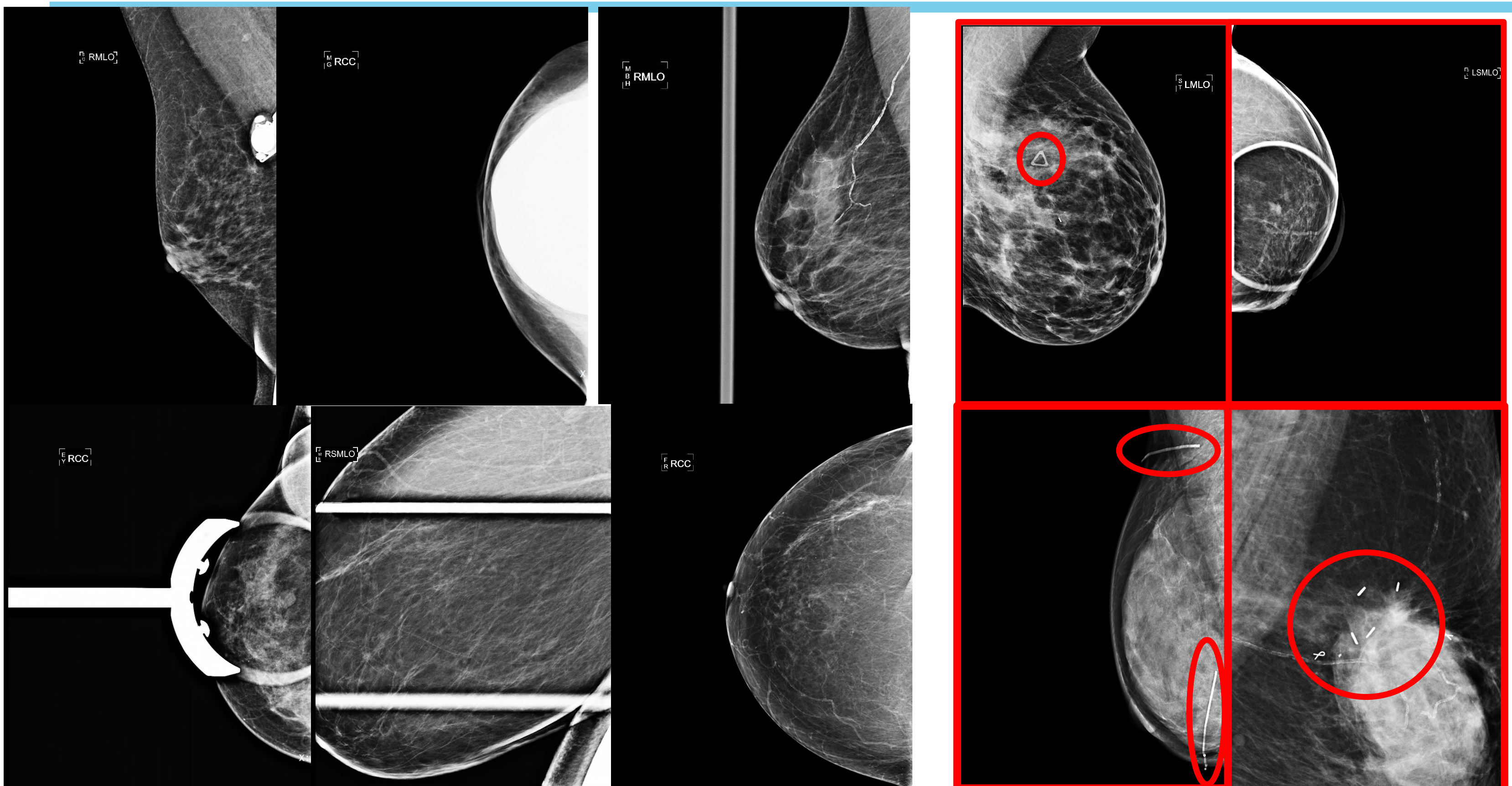
Our Hypothesis:

A fully-automatic pipeline for cleaning mammography images can be developed for the HIPIMR.

MammoClean - Project Overview

- This project focuses on answering how to develop an ***automated pipeline*** for cleaning and standardizing HIPIMR mammograms.
- This project involves:
  - Labelling data with anomalies
  - Developing methods to identify anomalies
    - Computer Vision (CV)
    - Deep Learning

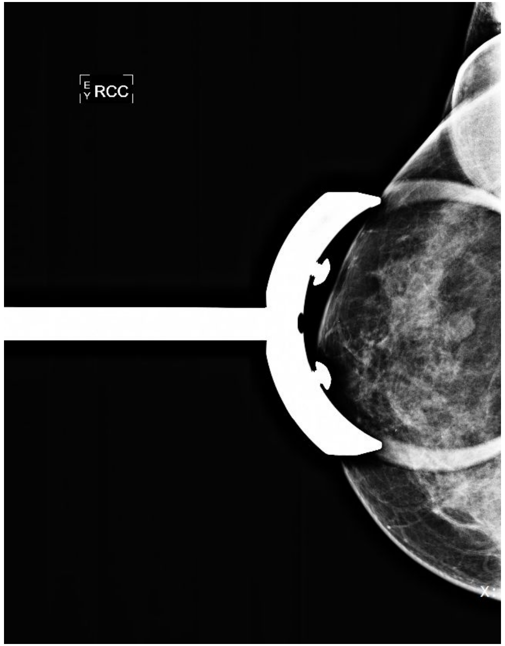
Anomalous Mammography Types

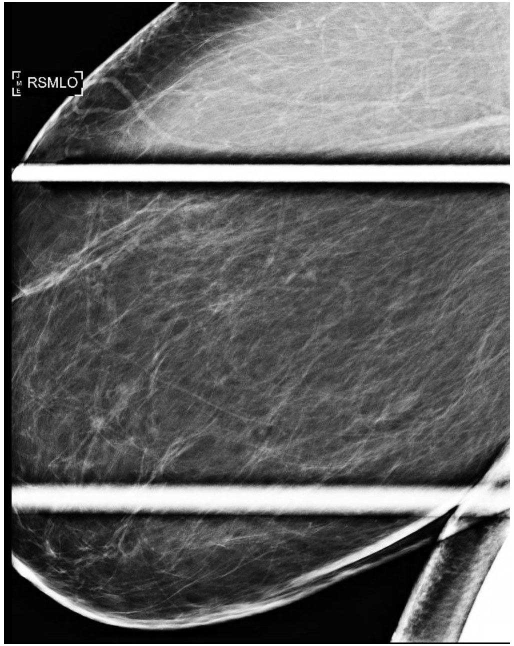


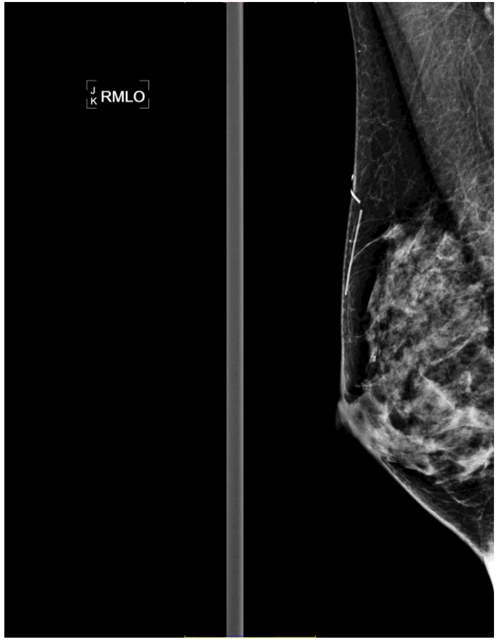
Annotations & Labelling

- We annotated our large dataset, individually identifying regions of interest and labelling them respectively by hand.

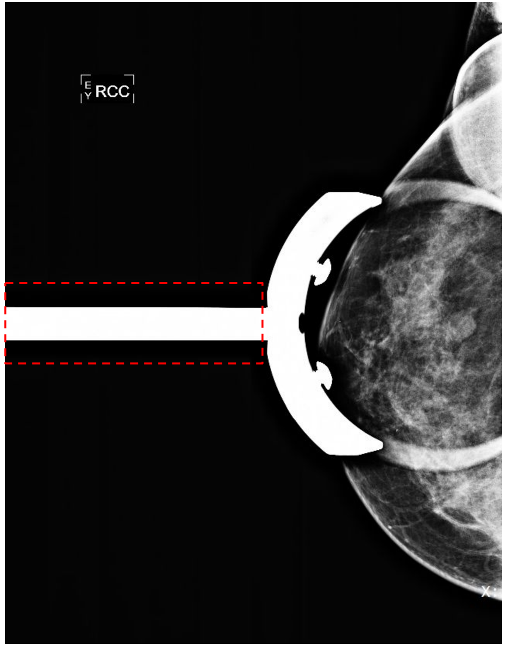
Computer Vision Methods

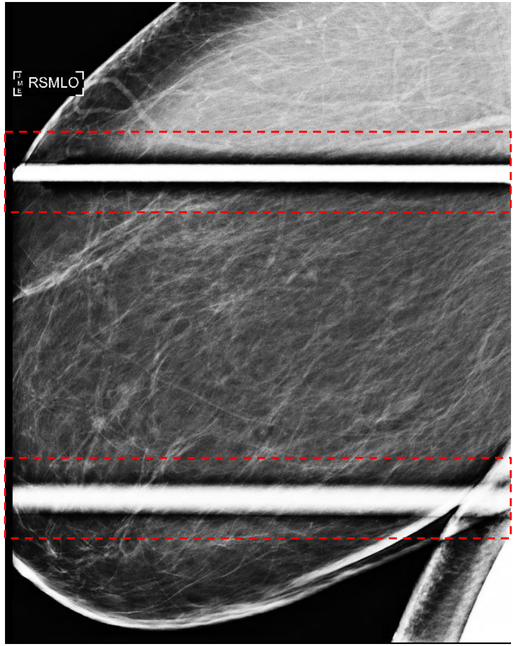


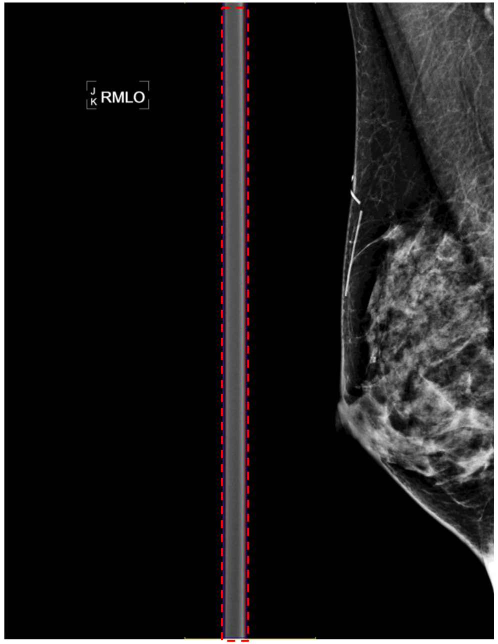




Computer Vision Methods

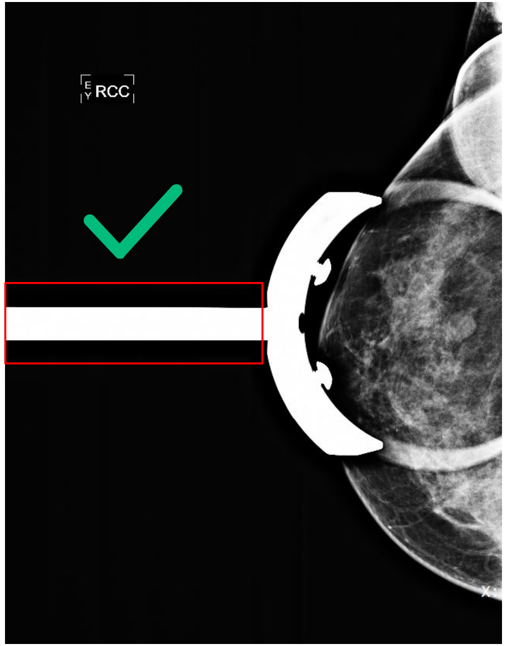




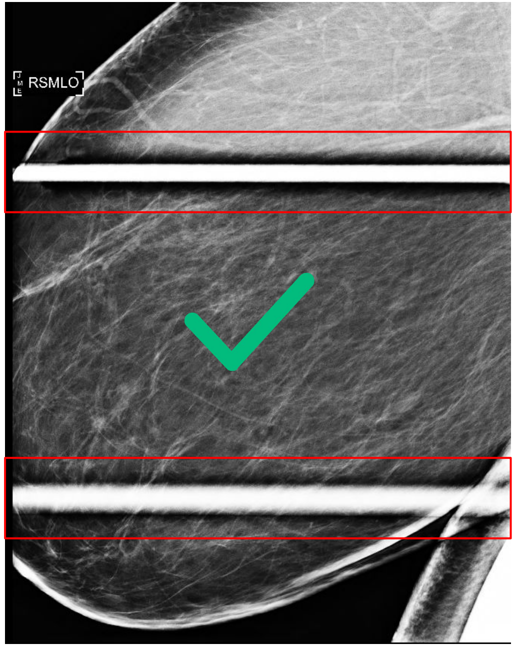


Computer Vision Methods

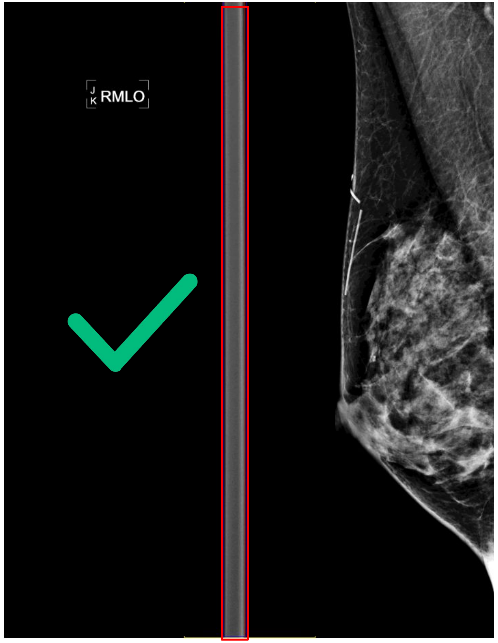
Is it bright?



Is it bright?



Is it straight?

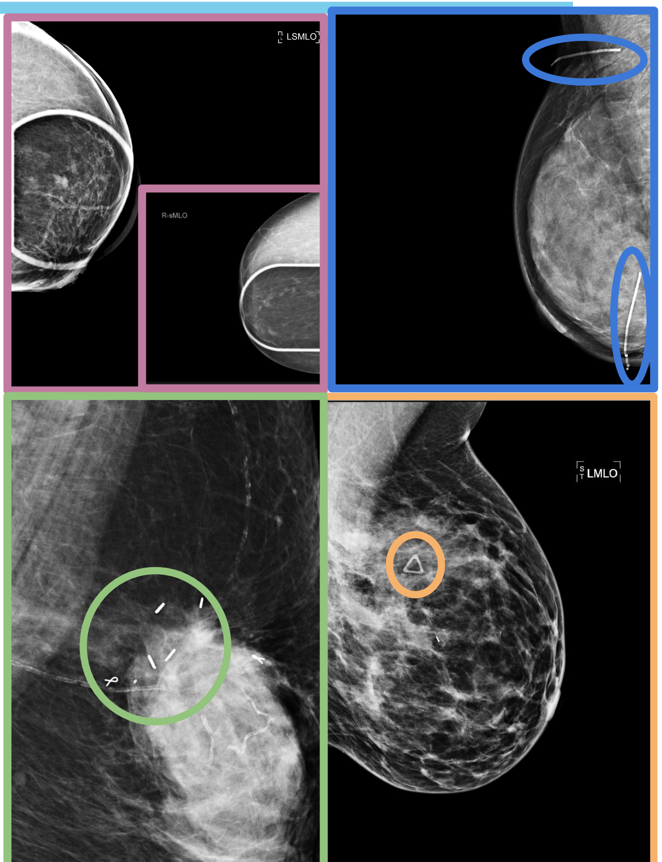


Deep Learning Methods

- **Granular Anomalies**
  - Palpable lump markers
  - Post biopsy clips
  - Wire Marker
- **Handless spot compression paddles**
  - Traditional & Rectangular

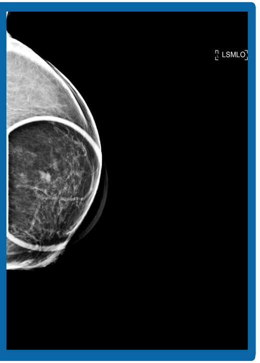
Problem: Too heterogeneous to automate using traditional computer vision methods.

Solution: Train a deep learning model to more effectively identify these anomalies.

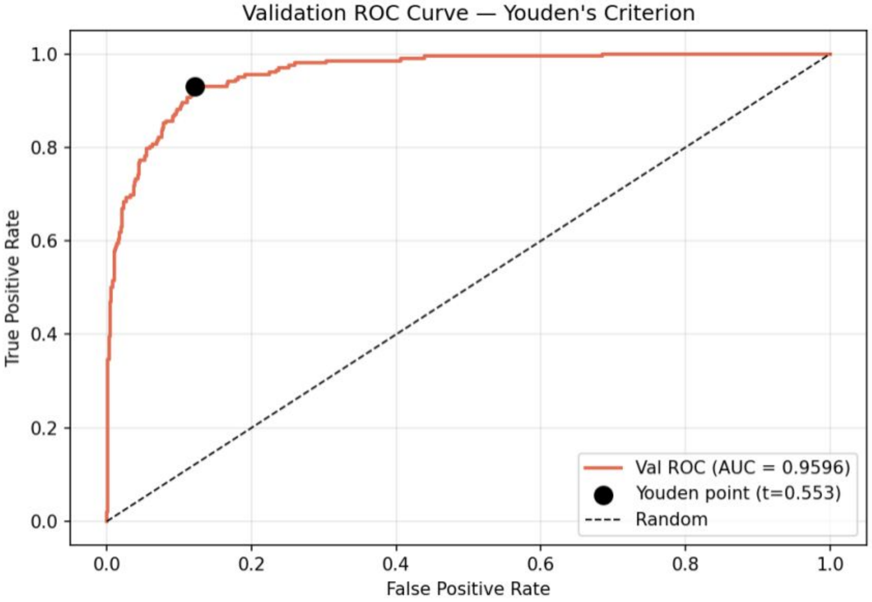


ResNet50 Spot Compression Model

A Binary Classification of **Spot Compression Paddles**

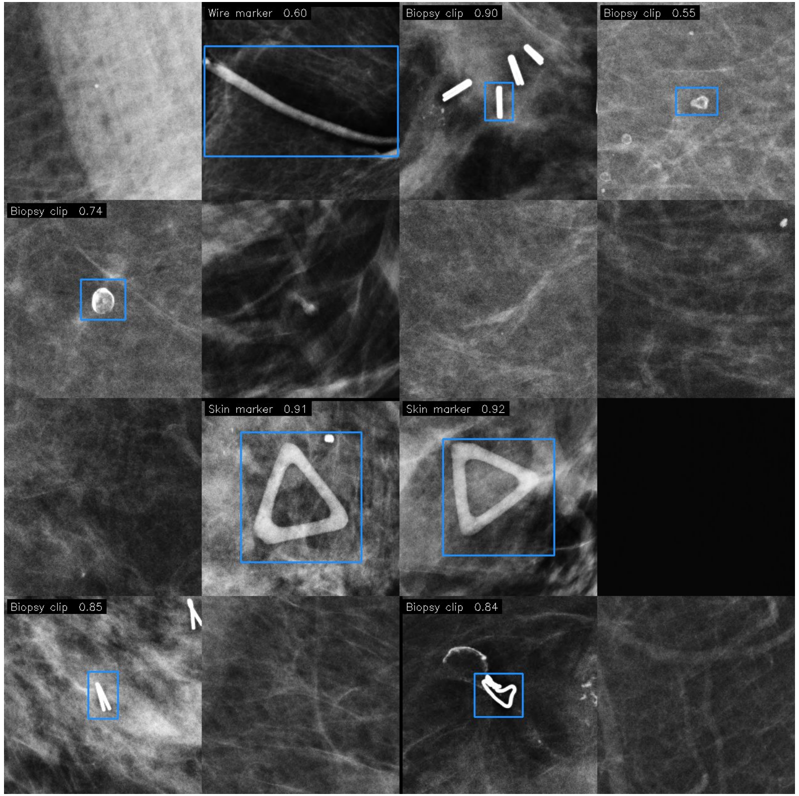


- Trained and compared 4 pre-trained models
  - **ResNet50** achieved the best performance
- Achieved an accuracy of **96.0%**



Skin Marker & Post Bx Clip Detection Model

- Detection of fine grained anomalies
  - **Post-Biopsy Clips**
  - **Wire Markers**
  - **Skin Markers**
- Trained a **RT-DETR** model
- **97% sensitivity** for skin markers and biopsy clips **81% sensitivity** for wire markers



Overall Results

- Performance of the classical CV methods was evaluated on a randomly selected, hand-labeled validation subset of 2,649 images:
  - The CV methods achieved great performance with accuracy and precision **exceeding 95%** for all anomalies.
- The classification model achieved an impressive 96% ROC-AUC score in accurately identifying Spot Compression paddles.
- The detection model achieved a 97% sensitivity for skin markers and post biopsy clips and 81% sensitivity for wire markers.

Quantitative Performance

- Performance metrics of the MammoClean pipeline by artifact category on the validation set.
- Best results are highlighted in bold.

Artifact Type	Precision	Recall	F1-Score	Balanced Acc	MCC
Spot Compression	1.00	0.71	0.83	0.85	0.81
Regular Paddle	**1.00**	0.97	0.98	0.98	0.98
Small Breast Paddle	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Breast Implants	0.96	0.79	0.87	0.90	0.87

Conclusion

- **MammoClean:**
  - **Rule-based computer vision**.
    - Transparent and explainable method.
  - **Deep Learning methods** for more difficult anomalies.
  - **Near-perfect detection** for standard compression artifacts.
- A dataset of 9,000 labelled mammograms to train & validate AI models specifically for **ANHPI** women.
  - Addresses critical limitation where mammograms of ANHPI women are not typically included in training.

Acknowledgement

HIPIMR: NIH R01CA263491

Koa: The technical support and advanced computing resources from University of Hawaiʻi Information Technology Services - Research Cyberinfrastructure, funded in part by the National Science Foundation CC* awards # 2201428 and # 2232862 are gratefully acknowledged.

Undergraduate Research Opportunities Program

Office of the Vice Provost for Research & Scholarship
University of Hawaiʻi at Mānoa

CHANGE HAWAIʻI

References

- Gary Bradski. The opencv library. Dr. Dobb's Journal of Software Tools, 25(11):120–123, 2000. 3
- Arianna Bunnell, Kailee Hung, John A. Shepherd, and Peter Sadowski. Busclean: Open-source software for breast ultrasound image pre-processing and knowledge extraction for medical ai. PLOS ONE, 19(12):e0315434, 2024. 2
- B.N. Hellquist et al. Effectiveness of population-based service screening with mammography for women ages 40 to 49 years: Evaluation of the swedish mammography screening in young women (scry) cohort. Cancer, 117(4):714–722, 2011. 1
- Epimack Michael et al. Breast cancer segmentation methods: Current status and future potentials. BioMed Research International, 2021:9962109, 2021. 1
- Vaijayanthi Nagarajan et al. Feature extraction based on empirical mode decomposition for automatic mass classification of mammogram images. Medicine in Novel Technology and Devices, 1:100004, 2019. 1
- Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464): 447–453, 2019. 1, 5
- S.J. Otto et al. Mammography screening and risk of breast cancer death: a population-based case-control study. Cancer Epidemiology, Biomarkers & Prevention, 21(1):66–73, 2012. 1
- Said Pertuz et al. Open framework for mammography-based breast cancer risk assessment. In 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), pages 1–4. IEEE, 2019. 1
- Laszlo Tabar et al. Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. Radiology, 260(3):658–663, 2011. 1
- World Health Organization. Breast cancer. https :/ / www . who . int / news - room / fact - sheets /detail/breast-cancer, 2025. Accessed: 2025-10-15. 1