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

Kanta Saito
Dept. of Computer Science
Univ. of Hawaii at Mānoa
kantas@hawaii.edu

Elijah Saloma
Dept. of Computer Science
Univ. of Hawaii at Mānoa
esaloma@hawaii.edu

Wilson Huynh
Dept. of Computer Science
Univ. of Hawaii at Mānoa
huynhw@hawaii.edu

Abstract

Ensuring equitable performance of Artificial Intelligence (AI) in breast cancer screening requires rigorous validation across diverse demographics, particularly underrepresented Asian, Native Hawaiian, and Pacific Islander (AANHPI) populations. However, raw clinical data from these groups often contains confounding artifacts that hinder accurate algorithmic evaluation. To address this challenge, we present MammoClean, a specialized pipeline for the automated cleaning and standardization of mammography data within the Hawaii and Pacific Islands Mammography Registry (HIPIMR). We introduce two key technical contributions to prepare this real-world data for AI development: (1) a taxonomy of anomaly types which may be harmful for AI evaluation; and (2) a multi-class anomaly detection framework that identifies confounding clinical artifacts such as compression paddles, breast implants, and cardiac devices. These enhancements directly mitigate algorithmic bias, facilitating the rigorous testing of AI models on AANHPI women to ensure equitable healthcare outcomes.

1. Introduction

Breast cancer remains one of the most common cancers affecting women globally, with the World Health Organization estimating 2.3 million new diagnoses in 2022 [10]. Early detection through mammography screening is critical for reducing mortality, a task increasingly supported by artificial intelligence (AI) systems [3, 7, 9].

Mammography is the gold standard for breast cancer screening, utilizing low-energy X-rays to visualize the internal structure of the breast. Standard screening protocols typically acquire two views per breast: the craniocaudal (CC) and mediolateral oblique (MLO) views. Radiologists examine these images for signs of pathology, such as masses, architectural distortions, or microcalcifications, which may indicate early-stage carcinoma. Computer-aided diagnosis has further transformed this workflow by assisting in cancer detection, mass classification [5], and risk assessment [8]. However, current major AI models suffer from a critical limitation: they were predominantly developed in Europe or the mainland United States [4]. Consequently, their performance remains unvalidated in populations with large proportions of Asian, Native Hawaiian, or Pacific Islander (AANHPI) women, creating a risk of significant algorithmic bias where models may underperform based on unrepresented demographic features [6].

Validating and adapting these models requires large, high-quality datasets, but raw clinical data from diverse registries like the Hawaii and Pacific Islands Mammography Registry (HIPIMR) is often noisy. Standard AI pipelines struggle with this real-world data because it contains specific anomalies that confound computer vision algorithms. These include palpable lump markers, compression paddles, breast implants, and clips marking past biopsy sites (Figure 1). These artifacts act as more than just noise; they present potential confounding effects that compromise model validation. For example, an AI model might learn to incorrectly associate skin markers with malignancy prediction, or interpret biopsy clips as indicators of high risk due to prior cancer scares, rather than analyzing actual tissue patterns.

In this work, we present MammoClean, an automated pipeline designed to bridge this gap by cleaning and standardizing mammography images from the HIPIMR. Our work builds on the intersection of epidemiological research and computer vision, introducing a novel preprocessing framework specifically designed to handle the unique artifacts found in raw clinical registries. By automating the curation process, we aim to make it feasible to research how AI models can be validated and adapted for the unique population of women in Hawaii.

Our first approach is the development of a multi-class anomaly detection framework that utilizes computer vision techniques to identify clinical artifacts. More specif-

ically, our approach targets non-anatomical features that act as confounders, including compression paddles, breast implants, and biopsy clips. Therefore, using these methods, we are able to pinpoint images that deviate from the standard training distribution expected by AI models, preventing foreign objects from being falsely interpreted as pathological features.

Our second approach is the implementation of a standardization protocol that harmonizes the dataset for downstream AI tasks. The motivation behind this approach is to support equitable model evaluation. To this end, our pipeline filters the dataset to create a "clean" baseline, strictly dropping images containing the anomalies identified in the previous stage. By establishing this high-quality registry, MammoClean provides the necessary infrastructure to rigorously test commercial models and develop new AI tools that work fairly for AANHPI women, reducing health imbalances for this unique population.

2. Data

The dataset for this study was sourced from the Hawaii and Pacific Islands Mammography Registry (HIPIMR), a diverse repository of breast imaging data distinct from standard mainland US datasets. All women included in this study participated in diagnostic mammography imaging at one of three clinical sites in the HIPIMR from 2009 to 2022. The HIPIMR collects data from three distinct clinical partners: Clinic 1 is a nonprofit healthcare network comprising four medical centers; Clinic 2 is a private nonprofit tertiary hospital; and Clinic 3 is a diagnostic medical imaging center [2].

Women were identified by the patient's legal sex in their clinical site's electronic medical record. The HIPIMR prospectively collects breast imaging, demographics, and clinical risk factors. Women were retrospectively selected for inclusion if they met all the following criteria: (a) had at least one diagnostic mammography visit; (b) had some finding indicated on their exam (BI-RADS 2, 3, 4, or 5); and (c) had known clinical breast density. From these, cases were defined as women diagnosed with invasive breast cancer at most 3 months from their extracted exam. Non-cases were selected from women who did not develop cancer following their examination, matched to cases at a 3:1 non-case:case ratio on year of screening mammography, mammography machine type, and patient breast density.

2.1. Data Acquisition and Demographics

The study utilizes a subset of 9,048 full-field digital mammograms collected from these clinical sites. The cohort is uniquely characterized by a high prevalence of Asian, Native Hawaiian, and Pacific Islander (AANHPI) women, a demographic historically underrepresented in AI training sets. Each image was converted from DICOM format to high-resolution PNG format for computer vision processing. Metadata regarding View Position (CC/MLO) and Image Laterality (Left/Right) was extracted to guide region-of-interest (ROI) selection. Standard screening protocols typically acquire two views per breast: the craniocaudal (CC) and mediolateral oblique (MLO) views.

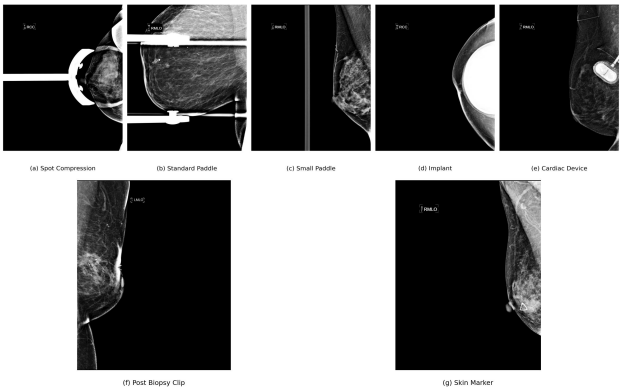

Figure 1. Taxonomy of anomalies in the HIPIMR dataset. (a) Spot compression handle. (b) Standard compression paddle edge. (c) Small breast paddle. (d) Breast implant. (e) Cardiac device. (f) Post Biopsy clip. (g) Skin marker.

2.2. Taxonomy of Clinical Artifacts

To rigorously evaluate the pipeline, we define the specific physical characteristics of the confounding anomalies present in this registry:

- **Breast Implants:** Prosthetic devices used for augmentation or reconstruction, typically appearing as large, radiopaque (white) oval or round shapes that obscure breast tissue and significantly alter the histogram distribution of the image.
- **Cardiac Devices:** Implantable medical devices such as pacemakers or implantable cardioverter-defibrillators (ICDs). These appear as distinct, high-contrast metallic objects with leads (wires) extending into the chest cavity, often overlapping with the breast tissue in the MLO view.
- **Compression Paddles:** Acrylic plates used to compress the breast during image acquisition. While standard paddles are usually transparent to X-rays, their edges or specific "spot compression" handles can appear as sharp, geometric lines or dense structural artifacts at the periphery of the image.

2.3. Ground Truth Annotation

To validate the proposed pipeline, the validation set was manually annotated for ground truth. The development and internal testing datasets were labeled for scan abnormalities by an author (A.B.), with disagreements in labeling resolved through adjudication. This ground truth dataset labels the

presence of the specific acquisition anomalies defined above versus clean (artifact-free) images.

3. Methods

The proposed MammoClean pipeline utilizes a deterministic, rule-based computer vision approach. Unlike deep learning models that require vast annotated training data, this pipeline relies on geometric heuristics and intensity profiling derived from the physical properties of mammographic acquisition artifacts. The system is implemented in Python using the OpenCV library [1].

3.1. Image Preprocessing and ROI Definition

Prior to artifact detection, all DICOM images were converted to 16-bit grayscale PNGs. Original diagnostic mammograms are typically high-resolution (ranging from 2400×3000 to 4000×5600 pixels) with a high dynamic range (12-14 bit depth). To ensure computational efficiency while retaining structural fidelity, images were downsampled to a fixed width of $W_{target} = 400$ pixels while maintaining the aspect ratio.

Crucially, the pipeline utilizes DICOM metadata tags—specifically *ImageLaterality* (Left/Right) and *ViewPosition* (CC/MLO)—to dynamically define Regions of Interest (ROIs). For example, cardiac devices are physically located on the chest wall, whereas spot compression handles appear on the lateral edge of the detector. By flipping the coordinate system based on laterality, the algorithms are invariant to breast orientation.

For tissue-dense artifacts (implants and cardiac devices), Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied to enhance local contrast before thresholding.

3.2. Compression Artifact Detection

Compression paddles appear as distinct geometric structures with straight edges and high-intensity boundaries. We developed three specialized detectors:

3.2.1. Spot Compression Handles

Spot compression views often capture the mechanical handle of the paddle. The algorithm extracts a vertical strip from the middle third of the image height ($H/3$ to $2H/3$). Based on image laterality, the lateral edge (5 pixels wide) is isolated. The detector triggers if the pixel count above an intensity threshold ($I > 150$) exceeds 75 pixels, signifying the presence of the rigid handle structure rather than gradual tissue attenuation.

3.2.2. Standard Compression Paddles

To detect the faint, straight vertical lines characteristic of standard paddles, the algorithm isolates the top $N\%$ brightest pixels (default 7%). A vertical edge detection scan is performed on the top and bottom thirds of the image. The image is classified as containing a paddle if:

1. Vertical bright segments are detected at the same x-coordinate (within a margin of error) in both the top and bottom regions.
2. The vertical midline of the image contains sufficient signal, distinguishing the paddle edge from the skin-air interface of the breast.

3.2.3. Small Breast Paddles

Small breast paddles appear as a distinct rectangular "box" enclosing the breast tissue. This detector utilizes a geometric coincidence check. The algorithm scans horizontal strips at the vertical center of the image. It identifies contiguous horizontal segments of bright pixels. A positive detection occurs only if a segment in the top strip spatially overlaps with a segment in the bottom strip, satisfying a span consistency ratio ($Span_{ratio} < 0.85$). This specifically targets the parallel top and bottom edges of the smaller compression device.

3.3. Foreign Object Detection

Foreign objects are characterized by radio-opacity significantly higher than glandular tissue. Because these objects attenuate X-rays far more strongly than biological tissue, they effectively mask underlying anatomical structures and compress the dynamic range of the image histogram. This disruption hinders standard intensity-based normalization techniques and can cause AI models to fixate on the high-contrast artifact rather than the breast tissue.

3.3.1. Breast Implants

Implants present as large, smooth, high-intensity regions. The detection pipeline follows a multi-stage process:

1. **Segmentation:** A binary breast mask is generated using Otsu's thresholding and morphological closing.
2. **Intensity Extraction:** The top 30% brightest pixels within the breast mask are isolated.
3. **Morphological Analysis:** Connected components are analyzed for geometric properties. A component is classified as an implant if it satisfies thresholds for Area ($> 80,000$ px), Circularity (> 0.35), and Density.

Heuristic overrides were implemented to account for extremely dense silicone implants, which trigger detection based on high-intensity pixel ratios ($> 70\%$ of ROI) even if shape constraints are marginally violated.

3.3.2. Cardiac Devices

Pacemakers and ICDs appear as small, extremely bright metallic objects near the chest wall. The ROI is restricted to the medial third of the image. The algorithm extracts the top 1% brightest pixels (a stricter threshold than implants) and filters connected components based on:

- **Circularity:** > 0.30 (distinguishes from linear surgical clips).
- **Aspect Ratio:** < 4.0 (rejects linear tubing associated with chemo ports).
- **Density:** > 0.5 (ensures the object is solid metal).

This strict filtering helps differentiate cardiac devices from scattered micro-calcifications.

4. Results

We evaluated the performance of MammoClean on the fully annotated validation set of 2,649 mammograms. Performance metrics were calculated against ground-truth labels using a strict binary classification criterion for each artifact category.

4.1. Quantitative Performance

Table 1 summarizes the classification metrics. The pipeline achieved exceptional performance on paddle detection, with the *Small Breast Paddle* detector achieving perfect metrics across all categories (Precision=1.0, Recall=1.0).

We report the Matthews Correlation Coefficient (MCC) and Balanced Accuracy alongside standard metrics. Balanced Accuracy, calculated as the arithmetic mean of sensitivity and specificity, is crucial for our dataset because standard accuracy is misleading when the "Anomaly" class is rare (e.g., only 7 Cardiac Devices vs. 2,600+ negatives).

Table 1. 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
Cardiac Devices	0.25	0.86	0.39	0.93	0.46
Average	**0.84**	**0.87**	**0.81**	**0.93**	**0.82**

4.2. Error Analysis

To understand the limitations of rule-based standardization, we analyzed the confusion matrices (Figure 4) and specific failure cases.

Compression Artifacts: The *Regular Paddle* and *Small Breast Paddle* detectors were highly robust. The *Spot Compression* detector showed high precision (> 0.99) but moderate recall (0.71). As seen in the confusion matrix, there were significant False Negatives. This is attributable to the algorithm's specific focus on the geometric "handle" structure; spot compression views taken without a visible handle mechanism were consistently missed.

Figure 2 illustrates this specific failure mode. To verify the efficacy of the detector on its intended target (paddles with visible handles), we performed an adjusted analysis excluding the "handle-less" cases (Table 2). When adjusting for this definition, the detector achieves perfect recall and precision, confirming that the method is robust for the specific visual feature it was designed to capture.

Figure 2. Example of a "handle-less" spot compression view. These cases account for the false negatives in the spot compression detector, which relies on the geometric signature of the handle mechanism.

Table 2. Performance comparison for Spot Compression detection: Baseline vs. Adjusted (excluding handle-less variants).

Scenario	Precision	Recall	F1-Score	Balanced Acc	MCC
Baseline	0.997	0.706	0.827	0.853	0.811
Adjusted	0.997	**1.000**	**0.998**	**1.000**	**0.998**

Foreign Objects: The *Breast Implant* detector performed reliably (MCC=0.87). False Positives were primarily caused by extremely dense breast tissue in younger patients mimicking the intensity of silicone.

Cardiac Devices: This category presented a unique challenge. While the model successfully identified 6 out of 7 devices (Recall=0.86), it suffered from low precision (0.25). These false alarms were triggered by high-intensity calcifications or surgical clips that satisfied the density threshold but lacked the specific shape of a pacemaker. In a practical cleaning pipeline, high recall is preferred over high precision for rare anomalies. Because the number of flagged images is small ($N = 18$ false positives), the cost of manual review is negligible compared to the risk of allowing foreign objects to contaminate the training distribution.

5. Conclusion and Future Work

In this work, we presented MammoClean, a novel, automated computer vision pipeline designed to standardize mammography data from diverse clinical registries. By

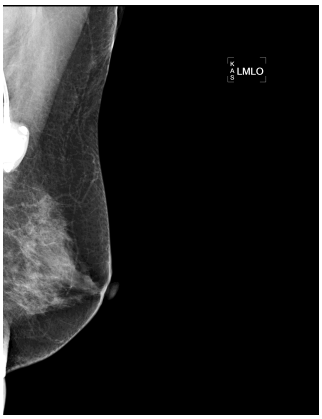

Figure 3. Failure case analysis. An example of a False Negative where the artifact was missed by the detection heuristics.

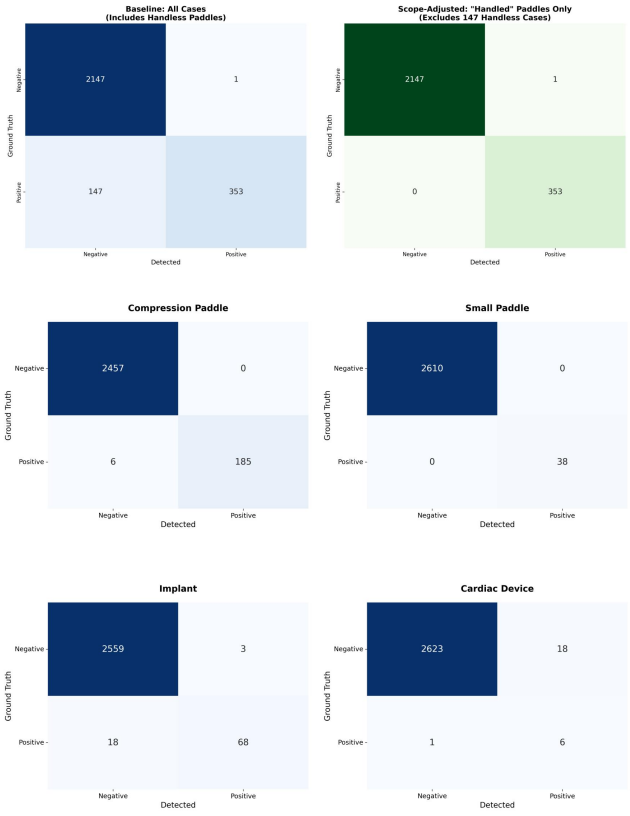

Figure 4. Confusion Matrices for the five anomaly categories. Note the class imbalance, with clean (Negative) images significantly outnumbering anomalies (Positive). The 'Small Paddle' detector achieved perfect recall (0 False Negatives).

developing specific detection modules for acquisition artifacts, we successfully curated a high-quality validation set from the Hawaii and Pacific Islands Mammography Registry.

Our results demonstrate that rule-based computer vision offers a transparent and explainable method for cleaning medical data, achieving near-perfect detection for standard compression artifacts. Crucially, this tool unlocks the potential of the remaining unlabeled dataset ($N \approx 9,000$), providing the necessary infrastructure to train and validate AI models specifically for Asian, Native Hawaiian, and Pacific Islander women. Addressing the data quality gap is the first step toward mitigating algorithmic bias and ensuring equitable healthcare outcomes [6].

Future Work: While our rule-based approach excels at detecting geometric artifacts, it struggles with subtle or irregular anomalies, such as spot compression paddles without visible handles. To address this, we are currently conducting a large-scale manual annotation campaign to expand our ground truth. We plan to use this labeled data to train a supervised deep learning model capable of detecting these complex, non-geometric artifacts that resist rule-based definition.

Author Contributions

Kanta Saito, Wilson Hyunh, and Elijah Saloma designed the computational pipeline, implemented the computer vision algorithms, and performed the error analysis.

Acknowledgments

This research was supported by the Hawaii and Pacific Islands Mammography Registry (HIPIMR). We acknowledge the University of Hawaii Cancer Center for providing data access and the University of Hawaii Koa High Performance Computing Cluster for computational resources.

We would like to express our gratitude to **Dr. Peter Sadowski** and **Arianna Bunnell** for their mentorship, valuable discussions, and guidance throughout this project. We also thank **Dr. John Shepherd** for the images and help.

References

[1] Gary Bradski. The opencv library. *Dr. Dobb's Journal of Software Tools*, 25(11):120–123, 2000. 3

[2] 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

[3] 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

[4] Epimack Michael et al. Breast cancer segmentation methods: Current status and future potentials. *BioMed Research International*, 2021:9962109, 2021. 1

[5] Vaijayanthi Nagarajan et al. Feature extraction based on empirical mode decomposition for automatic mass classifica-

tion of mammogram images. *Medicine in Novel Technology and Devices*, 1:100004, 2019. 1

[6] 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

[7] 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

[8] 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

[9] 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

[10] World Health Organization. Breast cancer. https://www.who.int/news-room/fact-sheets/detail/breast-cancer, 2025. Accessed: 2025-10-15. 1