EARLY BREAST CANCER DIAGNOSIS VIA BREAST ULTRASOUND AND DEEP LEARNING

A THESIS SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERISTY OF HAWAIʻI AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER SCIENCE

MAY 2023

By

Arianna Bunnell

Thesis Committee:

Peter Sadowski, Chairperson

John Shepherd

Peter Washington

Keywords: deep learning, breast cancer, ultrasound, diagnosis, cancer detection

ABSTRACT

Low- and middle-income countries, such as the U.S.-Affiliated Pacific islands, suffer from much higher advanced stage breast cancer (Stages III and IV) rates than high-income countries, especially where mammography services do not exist or have low accessibility. Examples include Palau (77% of breast cancer cases diagnosed at an advanced stage), American Samoa (72%), and the Federated States of Micronesia (82%). Portable, handheld, AI-enabled breast ultrasound devices operated by a local healthcare worker could greatly reduce advanced stage cancer rates in the U.S.-Affiliated Pacific Islands by making screenings drastically more accessible. In this work, we have explored AI models for both breast lesion detection and breast density estimation from clinical breast ultrasound. Breast density assessment and lesion detection and diagnosis were trained and evaluated on task-specific datasets collected from clinical breast imaging centers across Hawaiʻi, available through the Hawaiʻi Pacific Island Mammography Registry.

The results of the breast lesion detection task show that diagnosis of breast lesions is possible on ultrasound with concurrent classification of lesion descriptors for explainability, achieving 0.39 average precision. Precise delineation and classification of breast lesions is possible with AI applied to breast ultrasound. We expect performance to increase as more data become available. The typical performance across the breast lesion detection literature for non-explainable methods is 0.7 mean average precision. The breast density model is the first application of deep learning to predicting the BI-RADS mammographic breast density category from clinical breast ultrasound (inter-modality) and achieves 0.69 mean one-vs.-rest AUROC on a held-out test set. There is signal detectable by AI which relates mammographic breast density to breast ultrasound images. Methods for intra-modality classification of mammographic breast density with deep learning achieve approximately 0.93 mean one vs. rest AUROC on an internal test set.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Advanced stage breast cancer (Stages III and IV) rates are higher in low- and middle-income countries (LMICs) than in high-income countries. In LMICs, advanced stage breast cancer represents 60-80% of new breast cancer cases as compared to 25-40% of new breast cancer cases in the United States [3-5]. We take the U.S.-Affiliated Pacific Islands (USAPI) as a particular example. Examples of elevated advanced stage breast cancer rates in the USAPI include Palau (77% advanced stage breast cancer rate), American Samoa (72%), and the Federated States of Micronesia (82%) [6]. Excepting Guam and the Commonwealth of the Northern Marianas Islands, all territories included in the USAPI are classified as LMICs by the World Bank [7]. Advanced stage breast cancer has higher mortality than cancer discovered in the earlier stages. The CDC reports a 31% 5-year relative survival rate for breast cancer diagnosed at the distant stage (Stage IV), as compared to 85% and 98% for breast cancer in the regional and localized stages (Stages I-III), respectively [8]. Discovering breast cancer at earlier stages leads to higher survival for patients.

The USAPI (and more generally, geographically remote LMICs) suffer from low or nonexistent access to mammography, lack of radiologists and radiology technicians, and an undue burden of travel between neighboring islands. Population-level breast cancer screening is functionally nonexistent throughout the USAPI. An effective breast cancer screening paradigm for the USAPI would address the lack of mammography, burden of travel, and personnel shortage. Reducing advanced stage breast cancer rates involves implementing early diagnosis outreach efforts to catch breast cancer earlier, when there is a higher chance of survival and cancer is more responsive to treatment.

Portable breast ultrasound (BUS) is a breast cancer screening and diagnosis technology well-suited to addressing barriers to care in LMICs. Portable BUS systems allow for breast imaging to be brought directly to the patient, removing the need for patients to travel great distances to receive imaging. BUS is a viable replacement for mammography, maintaining a sensitivity and specificity (95% confidence interval) of 75% (64% to 83%) and 87% (74% to 94%) in breast cancer detection, respectively compared to 56% (45% to 66%) and 94% (86% to 98%) for mammography [9]. Artificial intelligence (AI) for BUS is a developing research area; some algorithms have shown radiologist-level performance [10-12]. AI in BUS can ameliorate constraints on radiologists by providing breast cancer diagnoses from imaging directly, as well as addressing the inflated false-positive rate of BUS for breast cancer detection [10].

In this work, we developed proof-of-concept AI algorithms for both breast lesion detection/description and breast density estimation on clinical BUS. Our overall hypothesis is that introducing portable handheld ultrasound systems coupled with similar algorithms and operated by a trained healthcare worker will reduce advanced stage cancer rates in low-resource areas where mammography is not available. For this to be true, the proposed early diagnosis approach will have to have a similar sensitivity and specificity for breast cancer detection as mammography read by a radiologist.

2 BACKGROUND

2.1 Breast Cancer Screening in Low-Resource Environments

Breast cancer screening in low-resource environments, such as in low- and middle-income countries (i.e., the US-Affiliated Pacific Islands) presents unique challenges to the conventional approach to screening in higher-resource environments. Lack of health literacy, geographical barriers, cancer stigma, lack of healthcare infrastructure, and cost of examination have all been cited as possible reasons why screening is not highly utilized in resource-constrained settings [13]. Despite mammography being the standard of care for breast cancer screening and early diagnosis in the United States, it may not be well-suited for resource-constrained environments. We propose that AI-assisted portable breast ultrasound (BUS) could provide the best balance of access, cost-effectiveness, and practicality for breast cancer screening and early diagnosis in low- and middle-income countries.

2.1.1 Mammography

Mammography is the currently recommended mode for annual breast cancer screening for early cancer detection of average- to high-risk women by the American College of Radiology (jointly recommend digital breast tomosynthesis) [14], the European Commission Initiative for Breast Cancer Screening and Diagnosis [15], and the World Health Organization (recommend in all but limited-resource settings with weak health systems) [16]. There is a substantial body of literature supporting mammography's efficacy as an early detection tool in reducing deaths from breast cancer [17-19]. Some guidelines recommend supplemental imaging for women with "extremely" dense breasts due to their increased risk of breast cancer coupled with the known decrease in sensitivity of mammography on higher density breasts [20].

Mammography consists of a low-dose X-ray of compressed breast tissue which allows radiologists to examine the internal structure of the breast and examine it for abnormalities. The low-dose X-ray beams are directed through the breast and received by an X-ray detector which converts the transmitted X-rays into an electrical signal. Cancerous breast tissue is typically denser than healthy breast tissue, causing it to attenuate the X-ray beam more than surrounding healthy tissue. Dense areas in the breast appear as bright spots on the generated mammography image. The cost of a digital 2D mammography machine is typically between $50,000-$100,000 [21]. Mammography systems are designed to be used in a single exam room, with minimal relocation, requiring the patient to travel to the clinic with a mammography machine (possibly on another island, when considering the USAPI) to get their exam.

2.1.2 Breast Ultrasound

The high up-front and continuing maintenance costs of mammography machines as well as their size and requirement of dedicated exam rooms makes mammography inappropriate for many low-resource settings. If people face an undue burden of travel, mammography cannot be moved to them. If there is a lack of resources for the healthcare system, the high cost of initial purchase and regular maintenance may place mammography out of reach. Lastly, a mammography machine is a single-use tool; it can only feasibly be used for breast examinations. Mammography's sensitivity is known to decrease dramatically in women with dense breast tissue, resulting in a risk of missing mammographically occult malignancies [20, 22, 23]. Asian, Pacific Islander, and Native Hawaiian women have been found to have denser breast tissue on average than White women [24-26].

Handheld breast ultrasound (HHBUS) seems to address many of the limitations of mammography, especially in a low-resource setting. However, HHBUS still requires a trained sonographer or radiologist to perform exams. Ultrasonography systems can be used for imaging of many different anatomical structures and are comparatively low-cost. In a BUS system, high-frequency sound waves are passed through the breast tissue via a handheld transducer. Dense, possibly cancerous tissue reflects more sound waves than healthy tissue or benign, fluid-filled cysts. Unlike in mammography, dense breast tissue does not typically obstruct the view of a lesion. However, BUS is known to suffer from a high false-positive rate, leading to unnecessary biopsies being performed [10]. The cost of a BUS machine is typically between $20,000-$50,000 [27].

2.1.2.1 Portable Breast Ultrasound

Portable BUS addresses the final barrier to early breast cancer detection in limited-resource environments: the undue burden of travel. Portable BUS systems are highly compact, allowing a healthcare provider to travel to the patient with the equipment. Portable BUS systems range in design with different manufacturers offering purpose-built laptop-size hardware, transducers coupled with mobile apps, and transducers with built-in screens. The cost of a portable BUS system is typically between $5,000-$15,000 [28].

2.2 Breast Lesion Detection in Ultrasound

Lesion detection is the most fundamental problem in using imaging in diagnosing breast cancer. If a suspicious lesion is found, a breast biopsy will be scheduled and location information (i.e., breast laterality and quadrant) is essential for procedural planning. Radiologists use irregularities in tissue structure to recognize breast lesions from BUS imaging. Lesions are typically marked and measured by the radiologist or radiology technician. We propose that AI-empowered breast lesion detection, coupled with comprehensive lesion descriptions in accordance with the ACR BI-RADS masses lexicon, can be used for early detection of breast cancer in low-resource areas.

2.2.1 ACR BI-RADS Masses Lexicon

Table 1: ACR BI-RADS Masses lexicon for Ultrasound. Lesion attribute categories considered indicative of malignancy are highlighted in italics.

Lesion Attribute	**Categories**
Shape	Oval
	Round
	Irregular
Orientation	Parallel
	Not parallel
Margin	Circumscribed
	Not circumscribed – Indistinct – Angular – Microlobulated – Spiculated
Echo Pattern	Anechoic
	Hyperechoic
	Complex cystic and solid
	Hypoechoic
	Isoechoic
	Heterogeneous
Posterior Features	No posterior features
	Enhancement
	Shadowing
	Combined pattern

The American College of Radiology (ACR) publishes the Breast Imaging Reporting & Data System (BI-RADS) Atlas to provide radiologists with breast cancer diagnosis criteria, reporting guidelines, and risk assessment guidance from all types of imaging [1]. The ultrasound section of the BI-RADS Atlas contains a list of criteria specific to assessing breast health and recognizing breast cancer from BUS. The ACR BI-RADS Masses lexicon has five characteristics to describe lesions in BUS: shape, orientation, margin, echo pattern, and posterior features. *Table 1* contains a comprehensive breakdown of all sub-categories of each of the characteristics in the Masses lexicon. Certain characteristics in the Masses lexicon are more indicative of malignancy than others. Thus, a lesion's description according to the Masses lexicon can be highly suggestive of malignancy status and is a large component of the final decision on whether to biopsy the lesion.

Each of the characteristics in the Masses lexicon is a description of a specific feature, as well as an indicator of malignancy status. Shape describes the structure of the lesion; irregular being more indicative of malignancy. Orientation describes the position of the lesion relative to the skin boundary on the breast; not parallel being more indicative of malignancy. Margin refers to the integrity of the border of the lesion, whether it is well defined (circumscribed) or blended into the surrounding tissue in some way (not circumscribed). All non-circumscribed subcategories are more indicative of malignancy. Echo pattern describes the echogenicity of the lesion in comparison to the surrounding breast tissue; all categories except anechoic being possible indicators of lesion malignancy. Lastly, the presence/absence of imaging artifacts below the lesion are the posterior features. The presence of any posterior features suggests a difference in ultrasound wave speed through the tissue and can be indicative of malignancy.

2.2.2 AI for Lesion Detection

AI for lesion detection from BUS is an exploding research area. The popularity of BUS as a supplementary breast cancer screening technology in China (all women 45-69 screened with mammography and BUS [29]) has spurred substantial new research in AI-powered systems for breast lesion detection, with 98 new papers indexed by PubMed since January 2022 alone. Lesion detection can be considered as a single task (as in modern object detection frameworks) or as a combination of lesion localization and classification. Classification-alone AI solutions rely on the presence of a radiologist to either: a) preselect region(s) of interest from a BUS scan image or b) receive a malignancy classification for an entire image/exam and perform post-hoc lesion localization. Segmentation-alone AI solutions also depend on radiologist presence for diagnosis and biopsy recommendation decision based on lesion characteristic not captured by lesion delineation alone. Automation of both lesion localization and classification provides the most direct and cost-effective solution for lesion detection. We provide a brief review of breast lesion detection and combination classification/segmentation AI algorithms on B-mode (unenhanced) breast ultrasound only.

There is a sizable amount of literature in using traditional machine learning (non-deep learning methods, ML) combined with computer vision features to classify BUS scans by malignancy status. There is comparably less literature for traditional ML to classify and segment lesions from normal breast tissue. Ding et al. use texture features derived from gray-level co-occurrence matrices to determine a rough region of interest (ROI) for the lesion, then use generalized multiple-instance learning to classify subregions of the ROI according to malignancy status. This two-step approach implicitly excludes isoechoic lesions, which are unlikely to differ in texture from the surrounding breast tissue [30]. Zhou et al. use adaptive thresholding and disk expansion [31] for lesion boundary detection and computer vision features defined on half- and full-lesion contours [32]. Zhou et al.'s method is likely to overlook features of lesions with posterior enhancement or combined pattern by throwing out the bottom half of the lesion. Most recently, Masjidi et al. perform contourlet transformations for both localization and feature extraction before using a decision tree for their final classification [33].

Deep learning (DL) for breast ultrasound is an accelerating research area. We further categorize DL approaches into location-explicit and location-implicit methods. Location-explicit methods train for localization performance; this category includes both detection and combination classification/segmentation methods. Location-implicit methods train only for classification performance and use saliency maps for lesion localization. We highlight a few notable contributions but do not attempt provide a comprehensive review of the state of the field.

We first cover location-implicit methods. Shen et al. present the most robust breast cancer classification model to date, developed on 5 million scans from 150,000 patients. Shen et al.'s method considers all imaging for a single breast when computing their final decision and uses saliency maps for rough lesion localization. This method provides a well-informed breast cancer classification due to consideration of all lesion views and the surrounding breast tissue and performs with Area Under the Receiver Operating Characteristic (AUROC) of 0.976 on an external test set [10]. Xing et al. integrate radiologist-assigned BI-RADS malignancy scores (ACR BI-RADS scale, ranges from 1 to 5 in order of increasing likelihood of lesion malignancy) using attention with raw image data when training their model. A lesion must be scored as at least BI-RADS 4A (>2% chance of malignancy) or classified as malignant by two separate network branches for the image to be flagged as cancerous. This approach displays comparatively low sensitivity and the authors provide no assessment of saliency map quality [34]. Tanaka et al. consider mass-level cancer classification through a dual-ensemble approach wherein they separately classify views of the same lesion, using three varying crops of each view. Using a geographically diverse dataset from seventeen imaging centers across Japan, Tanaka et al. achieve an AUROC of 0.951. This method considers multiple views of the same lesion, but applies the same network to every view, possibly exacerbating weaknesses of the model through ninefold application per breast mass [35].

Location-explicit methods have been applied to both still BUS image and video recordings. Lin et al. and Huang et al. both present methods which work directly on pre-recorded BUS video. Lin et al. use attention modules to fuse features from sets of three shuffled and consecutive frames, making their method ill-suited for real-time lesion detection [36]. Huang et al. construct a workflow to present cropped videos containing only video frames responsible for cancer classification to the radiologist, again resulting in a method which can only act on complete exam videos. Keyframes are selected using a reinforcement learning technique wherein the model is rewarded if the frame is highly indicative of malignancy or contains an annotated lesion [37].

Location-explicit methods working on still BUS images are more easily constructed from standard clinical BUS, as complete video segments are rarely stored past exam time. Yun et al. and Gao et al. develop semi-supervised methods using a combination of BUS scans annotated with lesion location (strongly-labeled) and scans with only biopsy results (weakly-labeled). Annotation cost for lesion location is high, making semi-supervised methods attractive for location-explicit methods. Yun et al. use a pretrained VGG-16 backbone to generate bounding boxes, achieving 66% correct localization and classification over their internal test set [38]. Gao et al. do systematic comparisons of semi- and fully-supervised labels using a mean teacher strategy. Gao et al. provide no standard metrics for localization performance, limiting reliability of their reported results [39]. Dai et al. use strong supervision and further enhance their input data with a super-resolution generative adversarial network (SRGAN). An ensemble of finetunes YOLOv4 and CenterNet are used in linear combination to decide the final classification. The development of SRGAN necessitates significant training time and effort, while supplying marginal gains in overall performance [40]. Lastly, Meng et al. enhance the YOLOv3 architecture by integrating channel- and lesion-attention modules to emphasize the importance of whole scan-based features (i.e., tissue texture) in relation to lesion-specific features. Meng et al. achieve an impressive mean average precision of 0.84 on the homogeneous population contained in their internal testing set, collected from various BUS imaging centers in China [41].

Commercial solutions exist in both the location-explicit and location-implicit categories. TaiHao Medical holds a current FDA approval for BU-CAD, their location-explicit AI-powered lesion detection BUS software. TaiHao Medical has not publicly released details of BU-CAD’s architecture. BU-CAD provides radiologists with lesion contour mapping, malignancy scores, and BI-RADS descriptors [42]. Koios Medical holds a current FDA approval for Koios DS, their location-implicit AI-powered lesion classification US software for both breast and thyroid lesions. Koios DS requires the lesion be localized by the radiologist before malignancy, shape, and orientation classification are performed, making it poorly suited for resource-limited scenarios [43]. Both BUS-CAD and Koios DS are approved only as reading aids for trained breast radiologists.

Despite the wealth of research into AI-assisted systems for both breast lesion localization and classification, there are significant gaps in the literature. BUS data requires specialized training for acquisition and is noisy and unstandardized. These limitations have led to an over-reliance of academic work on several small public datasets for both training and validation, limiting the real-world generalizability of academic work in this area [44, 45]. When considering application of a system in a resource-constrained environment such as the USAPI, there are few location-explicit methods which consider model performance without radiologist intervention or “error-catching.” To the best of our knowledge, the literature contains no works which provide complete demographic statistics; including breast density, age, race, BMI, and ethnic background. These notable gaps in previous work motivate the development of our own algorithm for breast lesion localization and classification.

2.3 Mammographic Breast Density Classification in Ultrasound

Computing personal breast cancer risk is a multi-faceted problem. The Breast Cancer Surveillance Consortium's Risk Calculator and the Tyrer-Cuzick Risk Assessment Calculator are two well-known tools developed to aid physicians in accurately assessing patients' breast cancer risk. The Breast Cancer Surveillance Consortium's Risk Calculator was developed and validated in over 1.1 million women in the U.S. and relies on a woman's age, race/ethnicity, family history in first-degree relatives (yes/no), existence and type of prior biopsies, and breast density to compute five- and 10-year risk scores relative to a woman's racial/ethnic and age group [46]. The Tyrer-Cuzick Risk Assessment Calculator increases the granularity of these variables (i.e., taking in family history as counts of specific relatives who developed breast cancer) as well as including the following: menopause status, birth history, history of hormone replacement therapy, BRCA mutation status, and ovarian cancer diagnosis [47].

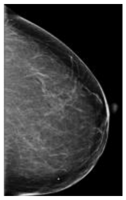
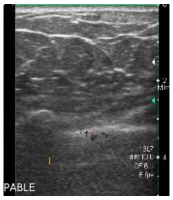
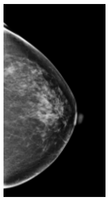
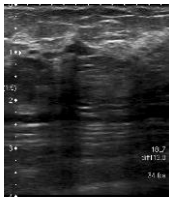
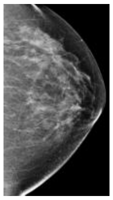
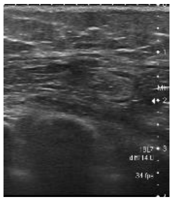
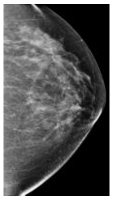
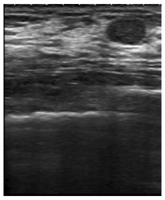
Mammographic density is independent associated with breast cancer risk and represents a significant variable in overall breast cancer risk assessment, with higher breast density being associated with a higher risk of breast cancer [48]. Mammographic density can be reported as percentage/volumetric density, representing the amount of fibrous tissue in the breast relative to fatty tissue, or as an ordinal category (A, B, C, and D, in increasing order of density). A and B are broadly considered low density and C and D are considered high density. The categories for mammographic breast density are set by the ACR BI-RADS breast composition guidelines for mammography and represent increasing proportions of the breast composed of fibroglandular tissue, as visually assessed by the radiologist. Historically, the categories were assigned to increasing quartiles of the breast composed of fibroglandular tissue, but this definition has been removed. *Table 2* includes example mammograms and BUS scans for women in each of the four defined ACR mammographic density categories. Assessment of mammographic breast density is crucial for providing accurate risk estimates and allocating screening to women at high risk, particularly in areas where resources are limited. We propose that AI-powered estimation of mammographic breast density from BUS can be used for breast cancer risk assessment in low-resource areas.

Assessing mammographic breast density from BUS is not well-explored in the literature. The paradigm of needing to get a measure defined on mammography from HHBUS seems only applicable in resource-limited settings where mammography is not available as trans-modal estimation of risk factors seems unnecessary when mammography is readily accessed. One small study of 41 women found that radiologist assessments of mammographic breast density on HHBUS and mammography demonstrate substantial intermodal (κ = 0.65) and interobserver (κ = 0.63) agreement [49]. Proxies for mammographic breast density based on acoustic speed of automated BUS (ABUS) through the breast tissue have been proposed and show strong correlation ($\rho > 0.8$) with percentage-based density measures [50, 51].

2.3.1 AI for Breast Density Classification

The volume of AI developed for breast density classification varies greatly on imaging modality. There are many commercial and academic strategies for determining breast density from mammography images [52, 53], but few exist for HHBUS. Jud et al. create linear regression models with features representing equally-binned gray-level histograms for B-mode BUS images. They achieve a 0.67 coefficient of variation with percentage of fibroglandular tissue from subjects' mammograms [54]. To the best of our knowledge, no literature exists on using DL for mammographic density assessment on BUS.

Table 2: ACR BI-RADS lexicon for describing breast composition from mammography [1].

Category	Description	Example Mammogram	Example Ultrasound
A	The breasts are almost entirely fatty		
B	There are scattered areas of fibroglandular density		
C	The breasts are heterogeneously dense, which may obscure small masses		
D	The breasts are extremely dense, which lowers the sensitivity of mammography		

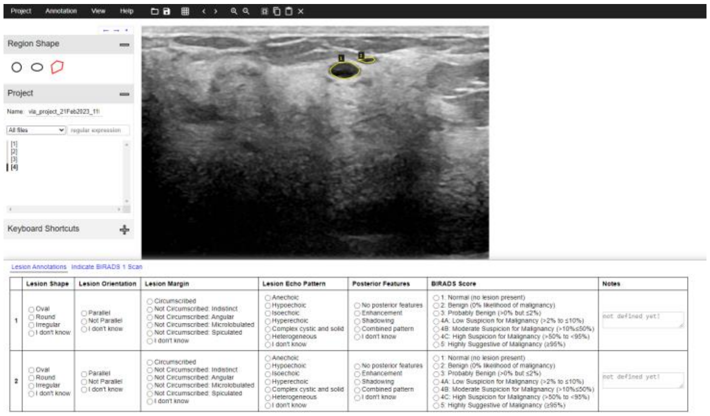
3 DATA

The data used in this study are sourced from the Hawai‘i Pacific Island Mammography Registry (HIPIMR). The HIPIMR is a prospective cohort of women that have participated in breast cancer screening at affiliated registry sites located within the catchment area of the University of Hawai‘i Cancer Center. The HIPIMR collects breast imaging and breast health information for women participating in screening from 2009 to the present. Personal health information remains with the images to allow for annual matching to the Hawai‘i Tumor Registry (HTR) to identify cancer cases.

HIPIMR data consist of imaging, imaging metadata, selected clinical variables, selected patient characteristics, and biopsy-confirmed cancer status. Presence and granularity of clinical variables and patient characteristics depend on the source registry site and personnel collecting information. More complex imaging variables, such as ACR BI-RADS characterization or lesion location, are not included in HIPIMR data and must be determined on historical imaging by consulting radiologists. In this section, we outline data collection and cleaning procedures for HIPIMR and consultant-sourced data.

3.1 Lesion Detection

There were two mutually-exclusive datasets collected for the lesion detection task. The strongly-supervised dataset was collected from a single clinical partner and additionally annotated by the study radiologist. The weakly-annotated dataset was collected from a single clinical partner and *not* additionally annotated. The strongly annotated dataset was pulled from the HIPIMR in August 2021, while the weakly annotated dataset was pulled from the HIPIMR in October 2022. The weakly annotated dataset was used to pretrain the backbone of our lesion detection model, while the strongly supervised dataset was used to train the complete model.



Name: via_project_21Feb2023_11

All files

regular expression

[1]
[2]
[3]
[4]

	Lesion Shape	Lesion Orientation	Lesion Margin	Lesion Echo Pattern	Posterior Features	BIRADS Score	Notes
1	○ Oval ○ Round ○ Irregular ○ I don't know	○ Parallel ○ Not Parallel ○ I don't know	○ Circumscribed ○ Not Circumscribed: Indistinct ○ Not Circumscribed: Angular ○ Not Circumscribed: Microlobulated ○ Not Circumscribed: Spiculated ○ I don't know	○ Anechoic ○ Hypoechoic ○ Isoechoic ○ Hyperechoic ○ Complex cystic and solid ○ Heterogeneous ○ I don't know	○ No posterior features ○ Enhancement ○ Shadowing ○ Combined pattern ○ I don't know	○ 1: Normal (no lesion present) ○ 2: Benign (0% likelihood of malignancy) ○ 3: Probably Benign (>0% but ≤2%) ○ 4A: Low Suspicion for Malignancy (>2% to ≤10%) ○ 4B: Moderate Suspicion for Malignancy (>10%≤50%) ○ 4C: High Suspicion for Malignancy (>50% to <95%) ○ 5: Highly Suggestive of Malignancy (≥95%)	not defined yet!
2	○ Oval ○ Round ○ Irregular ○ I don't know	○ Parallel ○ Not Parallel ○ I don't know	○ Circumscribed ○ Not Circumscribed: Indistinct ○ Not Circumscribed: Angular ○ Not Circumscribed: Microlobulated ○ Not Circumscribed: Spiculated ○ I don't know	○ Anechoic ○ Hypoechoic ○ Isoechoic ○ Hyperechoic ○ Complex cystic and solid ○ Heterogeneous ○ I don't know	○ No posterior features ○ Enhancement ○ Shadowing ○ Combined pattern ○ I don't know	○ 1: Normal (no lesion present) ○ 2: Benign (0% likelihood of malignancy) ○ 3: Probably Benign (>0% but ≤2%) ○ 4A: Low Suspicion for Malignancy (>2% to ≤10%) ○ 4B: Moderate Suspicion for Malignancy (>10%≤50%) ○ 4C: High Suspicion for Malignancy (>50% to <95%) ○ 5: Highly Suggestive of Malignancy (≥95%)	not defined yet!

Figure 1: Screen capture of the lesion annotation tool. For each scan, if lesions are present, the consulting radiologist delineates visible lesions, provides the ACR BI-RADS Mass lexicon and final classifications, and writes free-text notes if necessary.

3.1.1 Strongly Supervised

Breast lesion location and the ACR BI-RADS Mass lexicon characteristics were annotated by the consulting radiologist using a modified version of the VIA Annotation Software [55]. The modified tool was used to precisely delineate lesion boundary, provide classification of the Mass lexicon, and estimate lesion cancer status using the ACR BI-RADS categories for lesion malignancy. *Figure 1* displays a screen capture of an example lesion delineation using this tool. The lesion malignancy ratings were used for internal comparison with both the developed AI tool and biopsy records from the HTR.

Cancer status was collected from biopsy results stored in the HTR. If a record of cancer existed, all lesions for that subject were considered cases. Alternatively, if no record existed for the subject, then they were considered a control and all lesions were labeled as benign. This method of labeling retains the purest labels for a woman's overall cancer status, as biopsy is the only way to definitively diagnose breast cancer. However, there is no guarantee that either a) every lesion in a woman's breast is cancerous or b) the lesion in the image was in fact, the biopsied lesion.

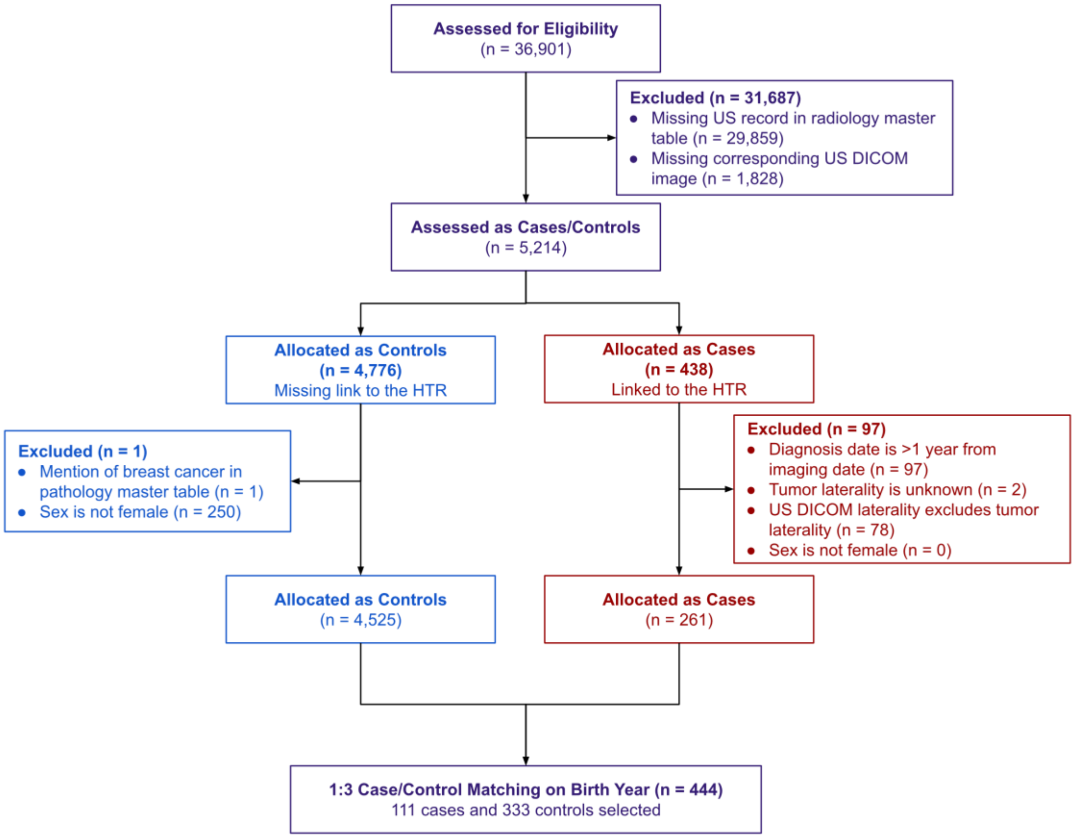

Assessed for Eligibility
(n = 36,901)
Excluded (n = 31,687)
• Missing US record in radiology master table (n = 29,859)
• Missing corresponding US DICOM image (n = 1,828)
Assessed as Cases/Controls
(n = 5,214)
Allocated as Controls
(n = 4,776)
Missing link to the HTR
Allocated as Cases
(n = 438)
Linked to the HTR
Excluded (n = 1)
• Mention of breast cancer in pathology master table (n = 1)
• Sex is not female (n = 250)
Excluded (n = 97)
• Diagnosis date is >1 year from imaging date (n = 97)
• Tumor laterality is unknown (n = 2)
• US DICOM laterality excludes tumor laterality (n = 78)
• Sex is not female (n = 0)
Allocated as Controls
(n = 4,525)
Allocated as Cases
(n = 261)
1:3 Case/Control Matching on Birth Year (n = 444)
111 cases and 333 controls selected

Figure 2: CONSORT-style chart illustrating selection criteria for patients for the strongly-annotated lesion detection dataset.

All patients with a record of BUS imaging in the HIPIMR were considered as possible cancer cases (n = 7,042). Exclusion criteria for cancer cases were as follows: 1) Subject missing US DICOM image (n = 5,214); 2) Subject missing a positive biopsy record in the HTR (n = 438); 3) Breast cancer diagnosis date more than a year from most recent BUS imaging date (n = 341); 4) HTR has no record of breast tumor laterality (n = 339); 5) US DICOM recorded laterality excludes HTR tumor laterality (n = 261); 5) Patient is not female (n = 261). 111 cases were matched to 333 controls by birth year for labelling by the study radiologist. Exclusion criteria for controls were as follows: 1) Subject missing US DICOM image (n = 5,214); 2) Subject has a positive biopsy record in the HTR (n = 4,776); 3) Subject has a mention of breast cancer in the HTR pathology master table (n=4,775); 5) Patient is not female (n = 4,525). Controls were semi-randomly selected from this pool. *Figure 2* provides an illustrative CONSORT-style chart delineating the patient selection process. Patients were randomly split into training (70%), validation (20%), and testing (10%) sets.

The selected 444 women had a total of 4,812 scans. After excluding scans with no assigned case group 4,759 scans were annotated by the study radiologist with a total of 6,266 lesions. Lesions (and their parent scans, if all lesions were incomplete) where the radiologist labeled any of the BI-RADS mass characteristics as "I don't know" were excluded, leaving 6,252 lesions remaining (n = 4,751). Previously detected split scans (see Section 3.3.3.1) were split and assigned their respective lesions (n = 5,683). Scans were then subject to the following exclusion criteria: 1) Scans with a missing case assignment (n = 5,619); 2) Scans with a missing laterality through the HTR (cases) or DICOM record (controls) (n = 4,630); 3) Invalid or elastography scans as indicated by free-text notes left by the study radiologist. (n = 4,446). After all filtering, 282 controls and 111 cases remained (n = 4,446). *Table 3* displays a breakdown of image counts by category and data split.

3.1.1.1 Weakly Supervised

"Weakly-labeled" refers to scans which only have cancer/no cancer labels based on their matching to a biopsy record in the HTR. No additional lesion localization or BI-RADS mass lexicon annotation were done on these scans. The weakly labeled scans were used to perform image-level classification (cancer vs. no cancer) DL model, which was then fine-tuned on the lesion detection task with the strongly supervised data.

All patients with a record of BUS imaging in the HIPIMR were considered as possible cancer cases (n = 36,052). Exclusion criteria for cancer cases were as follows: 1) Subject missing US DICOM image (n = 34,735); 2) Subject missing a positive biopsy record in the HTR (n = 707); 3) Subject is missing laterality in the HTR (n = 705); 4) Subject's record in the HTR not coded as "invasive" cancer (n = 607); 5) Subject's diagnosis is more than a year from their US imaging date (n = 505); 6) US DICOM recorded laterality excludes HTR tumor laterality (n = 501). 501 cases were matched to 1,503 controls by birth year and US scan manufacturer for labelling by the study radiologist. Exclusion criteria for controls were as follows: 1) Subject missing US DICOM image (n = 34,735); 2) Subject has a positive biopsy record in the HTR (n = 34,028); 3) Subject matched with a tumor of any behavior in the HTR (n = 6,721); 5) Controls were semi-randomly selected from this pool. *Figure 3* provides an illustrative CONSORT-style chart delineating the patient selection process. Patients were randomly split into training (70%) and validation (30%) sets.

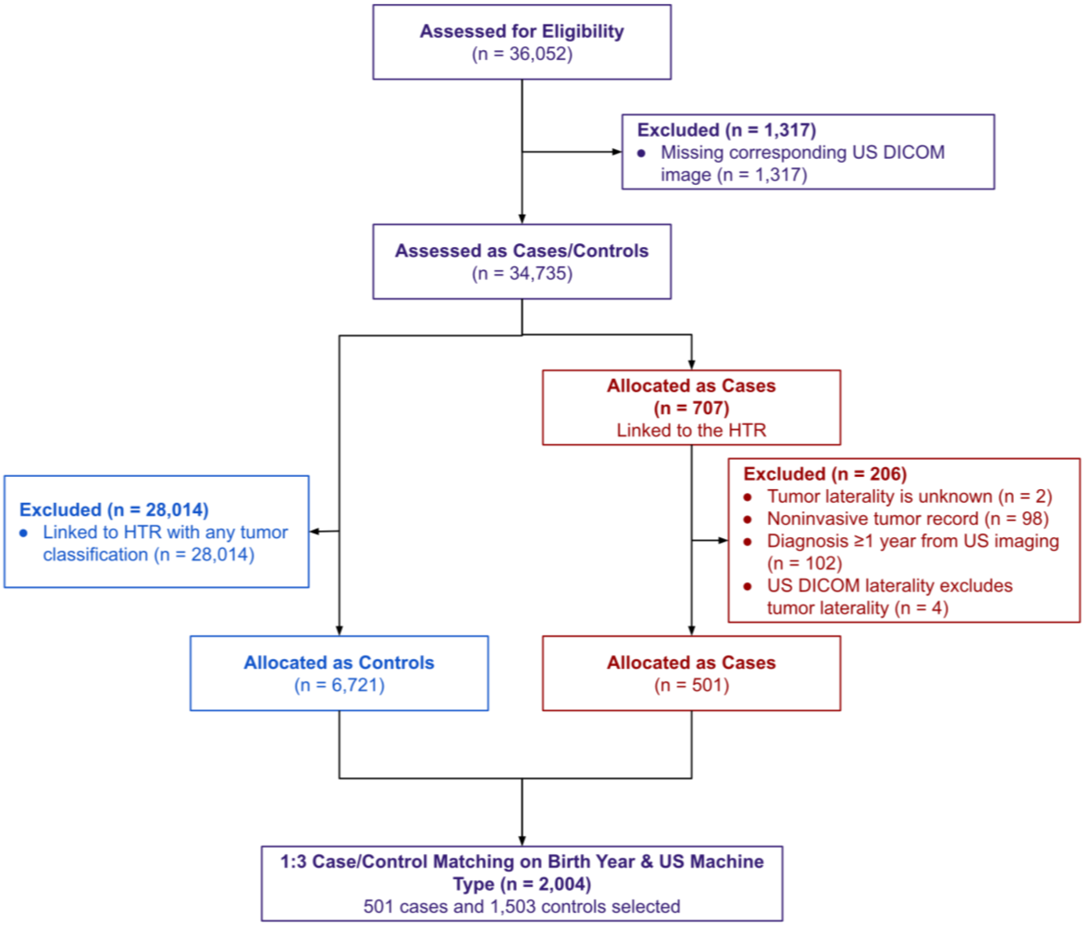

Assessed for Eligibility
(n = 36,052)
Excluded (n = 1,317)
• Missing corresponding US DICOM image (n = 1,317)
Assessed as Cases/Controls
(n = 34,735)
Allocated as Cases
(n = 707)
Linked to the HTR
Excluded (n = 28,014)
• Linked to HTR with any tumor classification (n = 28,014)
Excluded (n = 206)
• Tumor laterality is unknown (n = 2)
• Noninvasive tumor record (n = 98)
• Diagnosis ≥1 year from US imaging (n = 102)
• US DICOM laterality excludes tumor laterality (n = 4)
Allocated as Controls
(n = 6,721)
Allocated as Cases
(n = 501)
1:3 Case/Control Matching on Birth Year & US Machine Type (n = 2,004)
501 cases and 1,503 controls selected

Figure 3: CONSORT-style chart illustrating the selection criteria for patients in the weakly-labeled lesion detection dataset.

Table 3: Final image counts and averages per woman for both the weakly annotated and strongly annotated lesion detection datasets.

	Weakly Labeled			Strongly Labeled			
Characteristic, Unit	**Overall**	**Train**	**Valid**	**Overall**	**Train**	**Valid**	**Test**
Women, N	1,755	1,223	532	393	272	76	45
Women with benign findings, N	1,347	937	410	282	195	54	33
Women with malignant findings, N	408	286	122	111	77	22	12
Images, N	33,475	23,437	10,038	4,446	3,180	819	447
Images with benign findings, N	23,2376	16,453	6,823	1,661	1,206	523	292
Images with malignant findings, N	10,199	6,984	3,215	2,785	1,974	296	155
Average no. of images per woman, N	19.58	19.69	19.35	11.31	11.69	10.78	9.93

The selected 2,004 women had a total of 39,563 scans. Due to the nature of patient matching in the HTR, the export of scans for the weakly annotated dataset was not a mutually exclusive set with the set of strongly annotated images. To remedy this, we matched images based on their deterministic hashed file path, then excluded patients with strong annotations from the weakly annotated dataset. After this matching and exclusion, 1,401 controls and 410 cases remained (n = 37,662). Scans were subject to the following exclusion criteria: 1) Scans with scan area composed of more than 75% black pixels (n = 37,308); 2) Elastography scans (n = 37,217); 3) Scans with Color Doppler highlighting indicated in the image metadata (see Section 3.3.2) (n = 34,281). Detected split scans (see Section 3.3.3.1) were split and added as separate images (n = 39,362). Split scans were then excluded based on presence of Color Doppler highlighting indicated in the image itself (see Section 3.3.2) (n = 33,475). After all filtering, 1,347 controls and 408 cases remained (n = 33,475). *Table 3* displays a breakdown of image counts by category and data split.

3.2 Breast Density Classification

Mammographic breast density labels were sourced from both clinical records and the convolutional network developed by Wu et al. for classifying BI-RADS breast density category from standard four-view screening mammograms [56]. Wu et al. report an overall Area Under the Receiver Operating Characteristic curve (AUROC) of 0.916 on their internal, held-out test dataset. We compute the AI labels for two reasons: first, radiologists' clinical density assessment has been shown to be less reliable than computerized estimates [57-59]. Second, breast density labels from the Wu et al. classifier consist of a categorical pseudo-probability for each scan, across all four density categories. This additional information increases label richness and eases the AI learning process.

All patients with a record of BUS imaging in the HIPIMR were considered as possible cancer cases (n = 36,710). Note that the breast density dataset was extracted from the HIPIMR after the addition of a new clinical partner, greatly increasing the potential data pool. Exclusion criteria for all women were as follows: 1) Subject missing US DICOM image (n = 34,735); 2) Subject's US record is non-negative (BI-RADS >2) (n = 13,263); 3) Subject missing BI-RADS mammographic density within a year (n = 11,845); 4) Subject is missing standard four-view mammogram from within a year of their US record (n = 8,795). Cancer cases were matched to controls 10:1 on birth year and US machine manufacturer. Cases were subject to the following additional exclusion criteria: 1) Subject missing a positive biopsy record in the HTR (n = 1,062); 2) Subject's date of breast cancer diagnosis was on or after their US imaging (n = 383). Indicator variables were created for US less than a year before diagnosis (n = 232), US more than 10 years before diagnosis (n = 2), and non-contralateral US (n = 45). Controls were subject to the

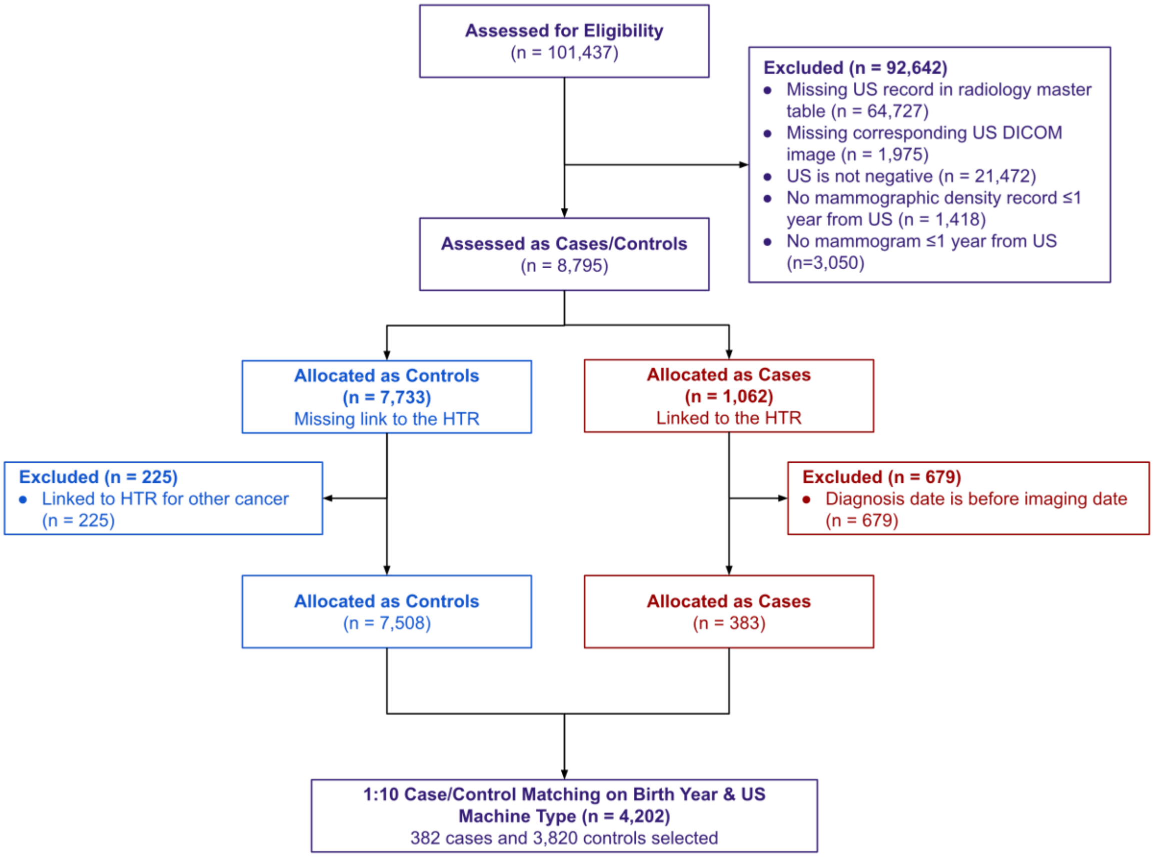

Assessed for Eligibility
(n = 101,437)
Excluded (n = 92,642)
Missing US record in radiology master table (n = 64,727)
Missing corresponding US DICOM image (n = 1,975)
US is not negative (n = 21,472)
No mammographic density record ≤1 year from US (n = 1,418)
No mammogram ≤1 year from US (n=3,050)
Assessed as Cases/Controls
(n = 8,795)
Allocated as Controls
(n = 7,733)
Missing link to the HTR
Allocated as Cases
(n = 1,062)
Linked to the HTR
Excluded (n = 225)
Linked to HTR for other cancer (n = 225)
Excluded (n = 679)
Diagnosis date is before imaging date (n = 679)
Allocated as Controls
(n = 7,508)
Allocated as Cases
(n = 383)
1:10 Case/Control Matching on Birth Year & US Machine Type (n = 4,202)
382 cases and 3,820 controls selected

Figure 4: CONSORT-style chart illustrating the selection criteria for patients in the ultrasound density dataset.

Table 4: Final image counts and averages per woman for the breast density classification task.

Characteristic, Unit	Overall	Train	Validation		Test
			A	B	
Women, N	4,100	2,452	819	814	829
Women with benign findings, N	3,722	2,226	742	737	754
Women with malignant findings, N	378	226	77	77	75
Images, N	104,965	63,467	22,185	18,337	19,313
Images with benign findings, N	93,692	56,406	19,776	16,474	17,510
Images with malignant findings, N	11,273	7,061	2,409	1,863	1,803
Average no. of images per woman, N	25.60	25.88	27.09	22.53	23.30

following additional exclusion criteria: 1) Subject has a positive biopsy record in the HTR (n = 7,508). Controls were semi-randomly selected from this pool. *Figure 4* illustrates the complete patient selection process in a CONSORT-style flowchart.

Case-control sets were randomly split into training (60%), validation (20%) and testing (20%) sets, stratified by breast density. Breast density strata were assigned by the most extreme AI-predicted density value in each group. A case-control set containing only women in BI-RADS categories B and C would be in the "middle" stratum, a case control-set containing at least one woman with BI-RADS category D breasts would be in the "dense" stratum, and a case-control set containing at least one woman with BI-RADS category A breasts, and no women with BI-RADS category D breasts, would be placed in the "fatty" stratum. Two copies were made of the validation set; validation set A and validation set B.

The selected 4,202 women had a total of 144,456 scans. The US scans were subject to the following exclusion criteria: 1) Invalid and duplicate scans were excluded (n = 114,210); 2) Scans with Color Doppler highlighting (see Section 3.3.2) (n = 111,863); 3) Split scans were detected and removed (n = 111,214); 4) Scans whose scan area was less than 224×224 pixels after cropping (see Section 3.3.3) (n = 109,178). The scan filtering process for the training set and validation set A concluded here. The testing and validation set B underwent the following additional cleaning steps: 1) Scans were cropped to exclude any text information in the scan area; 2) Scans with detected lesion highlighting were dropped from the dataset; 3) Scans whose scan area was less than 220×220 pixels after cropping were excluded. *Table 4* displays a breakdown of image counts by category and data split.

3.3 Breast Ultrasound Scan Preprocessing

Breast imaging collected through the HIPIMR originate from various clinical environments each housing a collection of hardware, software, and technician combinations. This natural data collection process results in highly irregular data, presenting in many different formats. We identified six factors which also influence the presentation of the scan, which cannot be reliably identified from software metadata or technician notes: resolution, shape, number of views, method, marker presence, and text presence. *Table 5* provides examples and a complete breakdown of all six factors contributing to image heterogeneity.

Scan preprocessing varied slightly across the density assessment and lesion detection tasks. For lesion detection, neither lesion markers nor overlaid text were identified or removed from BUS scans. For the sake of simplicity, dual-view scans were simply excluded from the dataset rather than being split into two training examples in the density classification task. Exclusions were minimized for lesion detection scan preprocessing to preserve as much of our limited dataset as possible. Exclusions were more comprehensive with the ultrasound density classification task.

3.3.1 PHI Removal

Protected Health Information (PHI) was removed from the breast ultrasound scans using several different methods. Breast ultrasound scans typically contain identifying information about the patient, as well as the time/location of scan collection. This information is stored on the scan image itself, as well as in the DICOM header. Thus, an independent de-identifying approach had to be used for each possible location of PHI. To remove PHI from the image metadata, the scans were exported from the HIPIMR using a PNG image format, detaching the image information from the information included in the DICOM header. This separation of image from DICOM information removed a large portion of the PHI (including collection location/time, patient identifying information, and specific patient demographic information). To remove

Table 5: Description of factors contributing to scan heterogeneity which cannot be otherwise reliably identified through metadata or technician notes. Each factor is assessed through examination of the BUS image itself. Scan shape, number of views, and method jointly determine preprocessing and inclusion/exclusion of the image.

Factor	Description	Possible Values
Scan Resolution	Scan resolution in pixels. Depends on hardware/software combination.	322 scan resolutions ranging from 278 × 432 to 956 × 668 pixels.
Scan Shape	Shape of scan area containing breast tissue. Depends on probe type (linear/convex) and software vendor.	Rectangular, trapezoidal, or convex.
Number of Views	Number of scan areas in a single image. Depends on technician preference and software capabilities.	Single scan area (single-view), or two separate scan areas (dual-view).
Method	Specific BUS method used. Depends on radiologist preference. Blood flow (Doppler) and tissue stiffness (elastography) can be indicative of malignancy.	B-mode (unenhanced), Doppler (contain bounding box and blood flow highlighting), or elastography (contain system overlays).
Marker Presence	Presence/absence of markers highlighting lesion location/size, added by technician. Depends on combination of software and technician preference.	At least four different annotation styles varying by color and annotation symbol
Text Presence	Presence/absence of text overlaid on scan area, added by technician. Depends on combination of software and technician preference.	Can describe probe position, laterality, or additional features of the scan (presence of an implant, nipple location, etc.)

PHI from the scan image itself, the images were cropped based on the RegionLocation coordinate values in the DICOM header. If there were more than one set of coordinates present, the first set was used. These coordinates specify which areas of the image contain the scan, excluding the header information.

3.3.2 Method

Breast ultrasound scans can be captured either unenhanced (standard US B-mode with no overlays or special features), with Color Doppler blood flow highlighting (colorful graphic overlays showing blood flow velocity in breast tissue) and strain elastography (also measures tissue stiffness/elasticity in response to pressure on breast tissue). Doppler and elastography scans were programmatically identified and excluded from both the lesion detection and ultrasound density analyses for the sake of consistency in the input data. Furthermore, enhancements to BUS represent additional cost and may not be accessible for low-resource environments.

Color Doppler and elastography scans were excluded from both the ultrasound density and lesion detection tasks and were jointly identified through HSV color masks for red, orange, green, and blue tones in the image as well as a mask highlighting Color Doppler scan artifacts (a white square outlining where blood flow highlighting will be shown). A subset of the Color Doppler scans can be identified using the PulseRepetitionFrequency DICOM tag, but the remaining scans need to be identified using the image-based approach.

3.3.3 Number of Views

In an effort to standardize the number of lesions of interest per scan image, splitting dual-view scans into two single-view images is desirable. Splitting dual-view scans also facilitates a standardized center cropping procedure for all examples. Without splitting, center crops of dual-view scans would contain the center dividing line, a highly unnatural shape. Because we were unable to identify a feature in the DICOM header metadata which identified dual-view scans, a programmatic approach was needed to both identify and split dual-view scans. Identification and splitting of dual-view scans were approached separately. Identification of dual-view scans was applied to both the ultrasound density and lesion detection tasks, while splitting was applied to lesion detection *only*. Dual-view identification algorithms differed between lesion detection and density classification due to the relative proportion of both SIEMENS brand and elastography scans in the datasets for each task, with comparatively less in the density task.

3.3.3.1 Lesion Detection

To identify a dual-view scan, HSV color masks were first applied to certain green and teal color ranges in the scan area. Elastography scans contain a green guiding line which is frequently near the center of the image which displays very similarly to the dividing line in dual-view scans when analyzed with edge detection. If a green guiding line was present, the scan was identified as single-view, as elastography imaging is only present in our dataset with single-view scans. Single-view scan from SIEMENS brand machines in our dataset had a certain number of characters of the teal SIEMENS logo cropped when extracted from the HIPIMR, distinct from the number of characters cropped from dual-view scans. The HSV color ranges of both the guiding lines and logo were determined heuristically.

After elastography and SIEMENS single-view images were identified using color ranges, an additional filter was applied on the image dimensions. If the scan width was less than 75% of its height, the scan was determined to be single-view. This proportion was determined heuristically from the dataset, as all dual-view scans were at least as wide as they were tall. The remaining scans were converted to grayscale and Canny edge detection was used to identify edges in the image. Canny edge detection involves a series of steps applied to a grayscale image: noise reduction, determining pixel intensity gradients, local-maximum suppression, and pixel connectivity thresholding. We will refer to pixels which the Canny algorithm identified as edges in the scan as "edge-pixels." If a scan's midline contained more than 100 edge-pixels, and at least ten more edge pixels than displaced midlines (10 pixels to either side of the midline), the scan was determined to be dual-view. Dual-view scans were then cropped along the identified midline to produce two separate data points. The metadata from the original scan, including case-control group assignment, patient identifier, and label, were duplicated into the new records to maintain integrity between our data partitions. The original scan was removed from the dataset.

3.3.3.2 Breast Density Classification

To identify a dual-view scan, first two filters were applied on the image dimensions. If the scan width was less than its width the scan was determined to be single-view. This proportion was determined heuristically from the dataset, as all dual-view scans were at least as wide as they were tall. The second filter identified scans with a height of more than 600 pixels and a width of less than 900 pixels as being dual-view. These pixel values were determined through examination of the SIEMENS brand dual- and single-view scans present in the dataset. The remaining scans were converted to grayscale and Canny edge detection was used to identify edges in the image. If a scan's midline contained more than 100 edge-pixels, the scan was determined to be dual-view. Dual-view scans were dropped from the dataset.

3.3.4 Scan Shape

The scan area cropping approach used was largely influenced by [59] and [9] and was performed after dual-view images were identified and optionally split. Breast ultrasound scans typically contain background areas containing textual metadata about the scan. These background areas are filled with black pixels and annotated with light-colored text. Removing the background increases the ratio of useful pixels in the image. The method for removing the black background from the scan involved two steps, consistent with the approach described in [59]. First, cropping of the background pixels was done via scan area identification through binary erosion and dilation. Second, additional cropping was done dependent on the scan shape (one of rectangular, trapezoidal, and convex). Cropping via erosion and dilation was applied to both the ultrasound density and lesion detection tasks the same way, while scan shape-dependent cropping was applied differently to each task.

The erosion and dilation procedure began with converting the image to a binary mask based on pixel color. All pixels with RGB values distinct from the mode pixel value were saved in a binary image. Examples of binary masks can be found in *Figure 5*(b). Binary dilation and erosion were then applied to the binary mask for n iterations, depending on scan manufacturer. The number of iterations and the size

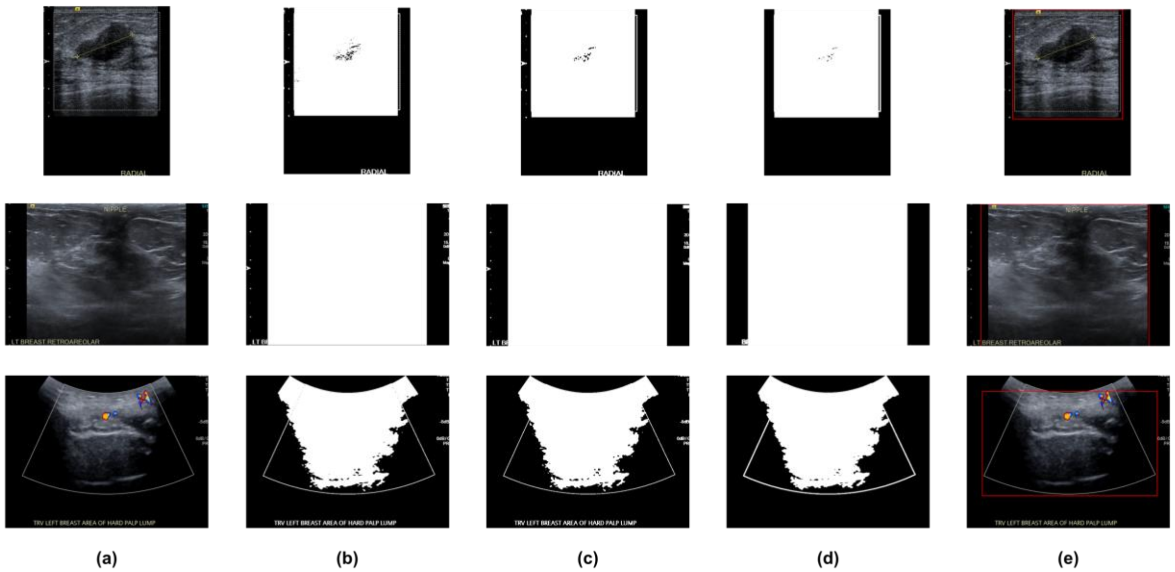


Figure 5: Examples of the erosion/dilation scan cropping procedure applied to the scans in both the breast density and lesion detection tasks. **(a)** Images as they were extracted from the HIPIMR. **(b)** Binary masks of the images, based on thresholding mode-valued pixels. **(c)** Binary masks of the scans after five iterations of dilation and erosion were applied. **(d)** The largest connected nonzero component (LCC) of the scan that was identified, after an additional round of dilation had been applied. **(e)** The red bounding box highlights the final dimensions each scan was cropped to.

of the kernel were determined heuristically. In essence, dilation of a binary image enlarges the nonzero image components, filling gaps and smoothing out intrusions into the component. Erosion of a binary image shrinks the nonzero image components, smoothing extrusions from the nonzero components. *Figure 5*(c) shows examples of binary masks post-erosion and dilation procedures.

After erosion and dilation were performed on the binary mask of the ultrasound scan, the largest connected nonzero component (LCC) in the image was identified, and a single dilation iteration was performed to smooth zero-valued intrusions into the LCC. Examples of the identified LCC can be seen in *Figure 5*(d), post-dilation. A bounding rectangle was constructed around the LCC, containing the entire component. For scans containing breast implants, high levels of acoustic shadowing, or very large benign lesions, the binary mask did not retain the entire scan area. Examples of the bounding rectangles are highlighted in red in *Figure 5*(e).

3.3.4.1 Lesion Detection

For images with a rectangular scan area, this was the end of the cropping procedure in the lesion detection task. Rectangular scans were cropped to the dimensions of their bounding rectangle found via the erosion/dilation procedure. Scans with a convex scan area (as shown in the third row of *Figure 5*) or a trapezoidal scan area, still need to be identified and cropped further. Note that, differing from [59], convex scans were only additionally cropped along the vertical axis, and trapezoidal scans were only additionally cropped along the horizontal axis.

To detect and vertically crop a scan with a convex scan area, our method was a subset of the scan area cropping method in [60]. Imagine a Euclidean coordinate system overlaid on our binary mask, wherein the top, left pixel corresponds to (0,0). The y-coordinate of the first nonzero pixel along the vertical midline of the bounding rectangle is defined as y_a. The y-coordinate of the upper edge of the bounding rectangle is defined as y_{top}. If $y_a > y_{top} + 20$, we determined we had detected a convex scan and y_a was defined as the new top of the bounding rectangle. Similarly, the y-coordinate of the row with the highest proportion of nonzero pixels is defined as y_b, and the y-coordinate of the bottom of the bounding rectangle is defined as y_{bottom}. *Figure 6*(a) provides an illustrative example of the reference coordinates. If $y_b > y_{bottom} - \frac{1}{2}h$, where h represents the height of the bounding rectangle, y_b was defined as the new bottom of the bounding rectangle.

To detect and horizontally crop a scan with a trapezoidal scan area, we took a similar approach. Imagine the same Euclidean coordinate system overlaid on our binary mask. The leftmost x-coordinate of of the bounding rectangle is defined as x_{left}, the rightmost x-coordinate of the bounding rectangle is defined as x_{right}, the y-coordinate of the horizontal midline of the bounding rectangle is defined as y_{mid}, and y_{top} is defined as before. *Figure 6*(b) provides an illustrative example of the reference coordinates. If the number of mode-valued pixels along y_{mid} was greater than two times the number of zero-valued pixels along y_{top}, a trapezoidal scan was detected and x_{left} and x_{right} were both moved in by half the number of zero-valued pixels.

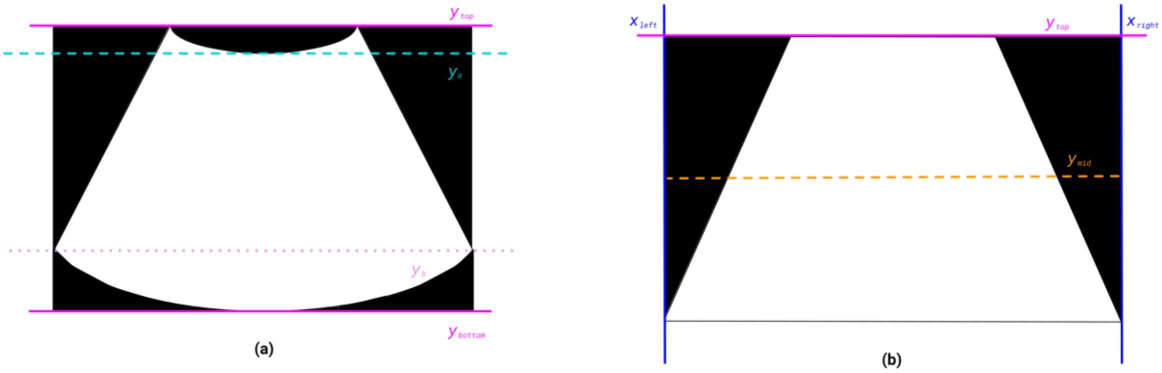


Figure 6: Illustration of additional cropping done on the lesion detection task based on programmatically-detected scan shape. **(a)** The procedure for convex scans was as follows: y-coordinate of the first nonzero pixel along the vertical midline of the bounding rectangle identified by the dilation and erosion procedure is defined as y_a. If $y_a > y_{top} + 20$, $y_{top} \coloneqq y_a$. The y-coordinate of the row with the highest proportion of nonzero pixels is defined as y_b. If $y_b > y_{bottom} - \frac{1}{2}h$, where h represents the height of the bounding rectangle, $y_{bottom} \coloneqq y_b$. **(b)** The procedure for trapezoidal scans was as follows: The y-coordinate of the horizontal midline of the bounding rectangle is defined as y_{mid}. If the number of mode-valued pixels along y_{mid} was greater than 2 times the number of zero-valued pixels along y_{top}, a trapezoidal scan was detected and x_{left} and x_{right} were both moved in by half the number of zero-valued pixels along y_{top}.

3.3.4.2 Breast Density Classification

All images were passed through an additional scan-shape determining cropping step in the breast density classification task. Imagine a Euclidean coordinate system overlaid on our binary mask, wherein the top, left pixel corresponds to (0,0). The y-coordinate of the upper edge of the bounding rectangle is defined as y_{top}. The y-coordinate of the bottom edge of the bounding rectangle is defined as y_{bottom}. We then determine the first and last x-coordinates where there is a white pixel (in our non-masked image, corresponds to the beginning of the scan area along the x-axis) in three horizontal slices of the image: $\left[y_{top}, \frac{1}{3}h + y_{top}\right), \left[\frac{1}{3}h + y_{top}, \frac{2}{3}h + y_{top}\right)$, and $\left[\frac{2}{3}h + y_{top}, y_{bottom}\right]$ where $h = y_{bottom} - y_{top}$. This process results in three sets of two x-values; $(x_L^1, x_R^1), (x_L^2, x_R^2), (x_L^3, x_R^3)$. We take the median values of $x_L^{\cdot}$ and $x_R^{\cdot}$ and crop the image width-wise with these median x-coordinates.

Height cropping was done using a similar process. Explicitly, the x-coordinate of the leftmost edge of the bounding rectangle is defined as x_{left}. The x-coordinate of the rightmost edge of the bounding rectangle is defined as x_{right}. We then determine the first and last y-coordinates where there is a white pixel (in our non-masked image, corresponds to the beginning of the scan area along the y-axis) in three vertical slices of the image and where $w = x_{right} - x_{left}$. This process results in three sets of two y-values; $(y_T^1, y_B^1), (y_T^2, y_B^2), (y_T^3, y_B^3)$. We take the median values of $y_T^{\cdot}$ and $y_B^{\cdot}$ and crop the image height-wise with these median y-coordinates. *Figure 7* provides an illustrative example of the reference coordinates.

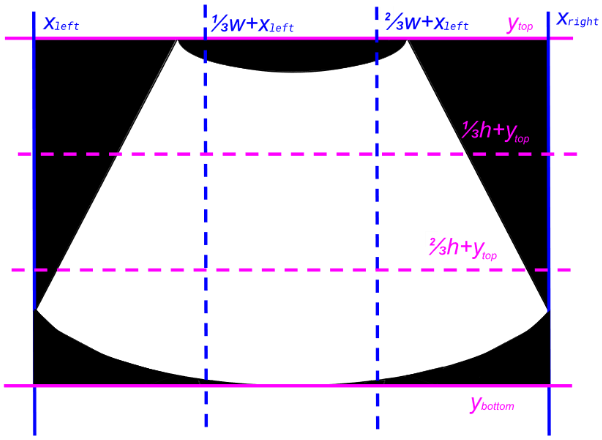


Figure 7: Illustration of additional cropping done on the breast density classification task. Pink lines represent the borders of our horizontal image slices used to determine $x_L^{\cdot}$ and $x_R^{\cdot}$. Blue lines represent the borders of our vertical image slices used to determine $y_T^{\cdot}$ and $y_B^{\cdot}$.

3.3.5 Marker Presence

Marker presence was only evaluated for the breast density classification task. BUS scans with markers were flagged and optionally (depending on data split) excluded based on the lesion marker presence. Scan artifacts such as lesion markers can provide signals of scan malignancy and cancer risk that are undesirable for an AI system designed for unenhanced scans. Scans with markers were not excluded from the lesion detection class due to the limited amount of data available. Scans with lesion markers were not cropped and re-introduced into the dataset simply because lesion markers tend to be placed near the center of the scan area, leaving little surrounding breast tissue to be allocated to a new image. Lesion markers were identified using HSV color masks for green, yellow, white (only included for BUS images collected on ATL and Philips Medical Systems equipment), and blue tones in the image. The resulting binary mask was center cropped to exclude software artifacts and passed through a single round of binary dilation. After dilation, the process for assessing lesion marker presence varied by BUS machine manufacturer.

For scans collected on Philips Medical Systems or ATL-branded equipment, contours were extracted from the image. Because the gray and white HSV color masks frequently picked up areas of high fibroglandular tissue density in addition to the lesion markers, we applied a size- and shape-based analysis to determine if a contour represented a lesion marker or not. If the approximate polynomial shape of the contour had between 14 and 17 vertices (correspond to the cross and "X" shapes of lesion markers) and the height and width of the contour was between 10 and 20 pixels, then a lesion marker was detected in the scan. The number of vertices and number of pixels were determined empirically. If more than two markers were detected in the scan it was excluded.

For scans collected with all other brands of BUS machine, only a size-based analysis was performed. If the height and width of the contour were each greater than five pixels, then a lesion marker was detected in the scan. The number of pixels was determined empirically. If at least one marker was detected in the scan it was excluded.

3.3.6 Text Presence

Text presence was only evaluated for the breast density classification task. BUS scans with text were flagged and optionally (depending on data split) cropped to exclude text. Scans with text in the image such that a 200×200 pixel crop could not be constructed from the remaining scan area were dropped. Text overlaying the scan area was identified using HSV color masks for yellow, white (only included for BUS images collected on Philips Medical Systems equipment), and gray tones (only included for BUS images collected on ATL equipment) in the image. The resulting binary mask was passed through four iterations of binary erosion and dilation with a (2,2) kernel to reduce noise, then a final dilation with a (9,9) kernel to ensure cropping excluded the entire annotation area. *Figure 8* provides an illustrative example of text presence cropping done on three scans.

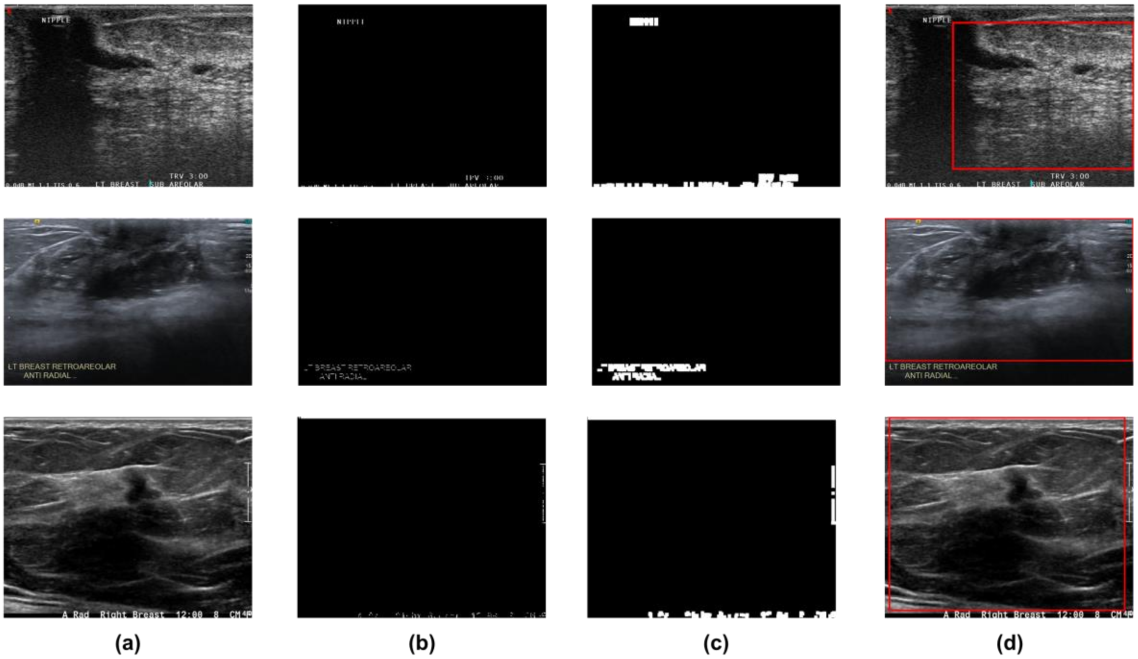


Figure 8: Examples of the erosion/dilation text cropping procedure applied to the scans in the breast density task. **(a)** Images after applying scan area cropping. **(b)** Binary masks of the images, based on thresholding mode-valued pixels. **(c)** Binary masks of the scans after four iterations of dilation and erosion, and a final (9,9) dilation were applied. **(d)** The red bounding box highlights the final dimensions each scan was cropped to.

4 MODELS

In this section we describe different model architecture choices for the breast density classification and lesion detection and BI-RADS mass lexicon description tasks. Convolutional neural networks (CNN) are the standard for deep learning development with images and have shown to be effective for medical imaging as well as natural images. CNNs encode spatial dependency between pixels in an image, making them well-suited for problems involving tissue texture, such as breast density classification and localization of breast lesions.

4.1 Lesion Detection

Object instance detection is a classical computer vision problem wherein we want to both localize objects with a segmentation mask and classify them into object types. For lesion detection, this means precisely outlining the lesion boundary, identifying the lesion as benign or malignant, and classifying the lesion according to the BI-RADS masses lexicon. All model building and training were done using PyTorch [61] and MetaAI's Detectron2 [62].

4.1.1 Architecture

Our lesion detection model is a Mask-RCNN implemented through MetaAI's Detectron2 object detection framework [62, 63]. Mask-RCNN is a state-of-the-art large object detection model extending the Faster-RCNN and RCNN architectures to provide instance segmentation masks in addition to bounding boxes for detected objects. Mask-RCNN involves three separate sub-networks in its original form: the *feature proposal network*, the *region proposal network*, and the *ROI head*. We extend the ROI heads section of the architecture to include additional sub-networks for the BI-RADS mass lexicon classification. A short description of each sub-network of our Mask-RCNN architecture is provided below.

The feature proposal network (or the backbone network) is a large convolutional network. Mask-RCNN was originally developed on ResNet frameworks but can be extended to use any reasonably large convolutional network for the backbone. We use a ResNet-101. Embeddings are extracted from the backbone network and fed into the next step of the network, the *region proposal network.* Backbone networks are further defined by at what stage in the backbone architecture features are extracted. The *ResNet101-FPN* feature proposal network we use is feature pyramid network-style backbone [64]. Features are from four intermediate layers in the network as well as the final residual block to be fed into the next step, the region proposal network. The feature maps extracted from different layers in a feature pyramid network are up-sampled to match the input size of the network. Features from later backbone layers detect larger objects/coarser patterns, while earlier layers focus on smaller objects/finer patterns due to the relative size of the receptive fields.

The region proposal network (RPN) is where proposals for object bounding boxes are suggested to the final stage, the ROI head. Embeddings extracted from the feature proposal network are fed into an initial set of convolutional layers in the RPN. For each feature map, the convolutional layers predict anchor and “objectness” scores for each pixel in the input image. Possible different sizes of bounding boxes, called anchors, are placed at each point on the feature maps and aligned to ground-truth bounding boxes during training. Boxes are then re-sampled to match a predefined foreground/background proportion and decrease the overall number of redundant boxes proposed. Finally, the best set of *n* (typically 1,000) bounding boxes are chosen based on their predicted “objectness” scores and passed to the final stage; the ROI head.

Input to the ROI head consists of the proposed bounding boxes from the RPN, as well as all the feature maps from the feature proposal network. Regions of Interest (ROIs) are chosen from the feature maps according to the proposed bounding boxes and rescaled/aligned to the input image (ROI pooling). The ROIs are then fed jointly into the mask and box heads in a traditional Mask-RCNN. We further add on heads for each of the BI-RADS mass categories, for a total of seven output heads. All heads follow the same general workflow: ROI pooling, convolutional layers, fully-connected layers, and non-maximum suppression to a target number of detections (we used MetaAI’s default value of $n = 100$ in training and $n = 4$ in testing).

4.1.2 Training

Model training was undertaken in two stages: feature proposal network pretraining and Mask-RCNN training. The feature proposal network was pretrained starting from an ImageNet-initialized ResNet-101 backbone, and training on the weakly-labeled BUS dataset for image-level cancer vs. no cancer classification. The 3-channel images were heavily augmented during training with random cropping, random ColorJitter, random equalization, and random horizontal/vertical flipping to discourage overfitting on the pretraining data. All models were trained with stochastic gradient descent with a cyclical learning rate scheduler to vary the learning rate throughout training. The minimum and maximum learning rates were tuned between [0.0001, 0.01] and [0.01, 0.5], respectively and the batch size was tuned between [32, 128]. The input features to the model were also tuned between 3-channel grayscale input and RGB input. All models were trained with early stopping with a patience of 5 epochs for a decrease in the validation binary cross-entropy loss, with a maximum of 500 epochs. Different model configurations were compared and the best was chosen based on validation loss. The chosen feature proposal network backbone (minimum and maximum learning rates were 0.01 and 0.1, respectively; batch size of 128; RGB input) was then passed into the training pipeline for the Mask-RCNN as initial weights to provide a warm start for training the model. The pretrained backbone had an AUROC of 0.68 on the validation dataset.

Table 6: Architecture structure for the ROI heads of the mask, box, and BI-RADS Mass characteristic heads. Note that the dimensions of the mass characteristic heads were tuned in tandem to decrease the search space.

	Mask	**Box**	**BI-RADS Mass Characteristics**
# FC layers	0	3	2
FC dimensions	N/A	1024	512, 256
# Conv layers	4	6	4
Conv dimensions	256	256	128
Pooler resolution	14	7	7

This pretrained model was then used to initialize the feature extraction backbone for the lesion detection model. The strongly-labeled BUS dataset was used for training, validation, and testing of this stage of model design. We tuned our model structure and training loop on top of the Detectron2 default architecture and hyperparameter values for Mask-RCNN [65]. *Table 6* describes the default architectures (Mask head) and final tuned architectures (Box and BI-RADS mass characteristic heads). *Table 7* describes the hyperparameter search space. All search spaces were explored by hand and a total of nine models were trained. Mask heads were also trained on non-optimal backbone networks in the same search space, but none outperformed the chosen backbone network.

Table 7: Search space of the hand-tuning of the hyperparameters and ROIHead branches architecture. Note that the dimensions of the mass characteristic heads were tuned in tandem to decrease the search space.

Parameter	**Search Space**	**Selected Value**
Backbone Frozen Stage	• Stem only • 1 residual block	1 residual block
BI-RADS Heads FC layer 1	{256, 512}	512
BI-RADS Heads FC layer 2	{256, 512}	256
Learning rate warmup iterations	{500, 1000, 2000}	1000
Box Head # conv. layers	{4, 5, 6}	6
Box Head # FC layers	{2, 3, 4}	3
Base learning rate	[0.001, 0.1]	0.02

The box head was trained class-agnostic due to no marked difference in lesion size between benign and malignant lesions in our dataset and the mask head was trained class-aware due to the marked differences in shape between benign and malignant lesions. All BI-RADS characteristics were trained class-agnostic, but class-aware training for the BI-RADS heads, as well as custom head architectures for each, is an area of future work.

4.2 Breast Density Classification

In classification in computer vision, we typically want to categorize our images into n nominal categories according to their features. In breast density classification, the problem is originally defined as an ordinal classification. However, we approach it as nominal. Our pseudo-probabilistic labels from the deep learning algorithm in [56] were found to display non-ordinal distributions (bi-modal) in only 1.3% of all scans. We chose not to enforce ordinality through the loss due to ordinality being implied in the image labels. All model building and training were done using PyTorch [61], Optuna [2], and Kornia [66].

4.2.1 Architectures

Our breast density classification model is a fully convolutional network, with the last layer concatenating all features and computing the final four-class multinomial pseudo-probability distribution. We chose this architecture based on the structure of the problem; breast density is a textural feature, defined over the entire breast. Using a fully convolutional structure (i.e., no pooling) allows for all information about the texture in the image to be retained passing between layers. We also experimented with adding residual blocks, batch normalization, and a U-Net style architecture. However, the fully convolutional network consistently outperformed the other architecture choices.

4.2.2 Training

Model training was undertaken in a single stage for this section of the project. The single-channel breast density images were augmented during training with random cropping, random to discourage overfitting. Median filtering, histogram equalization, and image normalization were also experimented with, but were found to have no effect on model validation performance. Based on initial experiments [67], we determined that brightness augmentation of images was helpful to the network. Breast density is defined on the visually assessed proportion of fibroglandular tissue, so brightness augmentation may make it easier for the network to pick out the brighter portions which are contributing to the final judgment. Three brightness augmentations were tested and added to the image as additional input channels: $p + \overline{p}$ (up), $p - \overline{p}$ (down), $p \pm \overline{p}$ (both), and no augmentation (none) where $\overline{p}$ represents the mean gray-level value in the scan. Brightness augmentation was tuned with Optuna.

All models were trained with either (1) a cyclical learning rate scheduler to vary the learning rate throughout training (with stochastic gradient descent optimization algorithm), or (2) with a decaying learning rate based on encountering plateaus in validation loss (with Adam optimization algorithm). The learning rate plateau strategy was found to be more effective in decreasing the validation loss throughout training. All remaining hyperparameters were tuned through Optuna. A 50-sample optimization search was performed with a Tree-structured Parzen Estimator sampler with 16 warmup, randomly-sampled trials. *Table 8* details the search space for the Optuna runs.

Table 8: Search space of the tuning of the hyperparameters and architecture of the fully convolutional network for breast density classification. 50 samples were drawn from the search space using the Optuna TPESampler with 16 warmup iterations [2].

Parameter	Search Space	Selected Value(s)
# Convolutional layers	{2, 3, 4, 5, 6}	3
# Convolutional filters	[16, 128]	32, 96, 64
Kernel Size	{3, 5, 7, 9}	7, 5, 3
Learning rate (LR)	$[1 \cdot 10^{-7}, 1 \cdot 10^{-3}]$	0.001
Dropout	[0, 0.7]	0.3
Brightness augmentation	{up, down, both, none}	down
LR Scheduler/ Optimization algorithm	{cyclical/SGD, plateau/Adam}	plateau/Adam

Due to breast density estimation from BUS being a relatively unexplored research area, we also chose to train two simpler models for comparison, derived from the features developed by Jud et al. [54]. Jud et al. define 33 evenly-spaced gray-level bins which are used as input to a linear regression model to precisely define proportion of fibroglandular tissue in the breast. We take Jud et al.'s gray-level features and feed them into a multiclass logistic regression and multi-layer perceptron model to provide a baseline of performance for our model. The multi-layer perceptron was trained with two hidden layers (sizes 512 and 256, respectively), a learning rate of 0.001, and the Adam optimizer.

5 RESULTS

5.1 Lesion Detection

Table 9: Average precision at intersection over union 0.5 values for lesion cancer status and the BI-RADS mass lexicon (shape, orientation, margin, echo pattern, posterior features) for both bounding box detections and segmentation masks.

Target	**Bounding Box AP@50**	**Segmentation AP@50**
Cancer	38.5	39.2
Shape	13.3	14.2
Orientation	17.6	18.2
Margin	7.9	8.4
Echo Pattern	11.6	12.2
Posterior Features	11.3	11.8

The lesion detection model was evaluated using average precision (AP) at intersection over union (IoU) 0.5. (AP@50). IoU is defined as the area of intersection between the ground truth object and the predicted object over the union of their areas. An IoU of 0 would mean that the prediction does not overlap the ground truth whatsoever and an IoU of 1 would indicate perfect correspondence of ground truth and prediction. We can threshold the IoU value at α for computing the average precision. A detection with IoU ≥ α is considered a true positive, 0 < IoU < α is considered a false positive, and IoU = 0 is considered a false negative (ground truth detection missed completely). IoU can be defined for both mask and bounding box targets. AP@50 is defined as the area under the precision recall curve when we classify our detections according to the IoU threshold $\alpha = 0.5$. We compute the area under the precision recall curve for each sub-categorization separately, then take the mean to come to our final AP value. Each BI-RADS characteristic classification was evaluated independent of lesion cancer classification.

The best-performing Mask-RCNN model performed with average precision at intersection over union 0.5 (AP@50) = 38.5 (bounding box) and AP@50 = 39.2 (segmentation) for detection and classification of lesion cancer status. Results for each of the BI-RADS mass lexicon characteristics are enumerated in *Table 9.* Note for the echo pattern target, there were no images with a hyperechoic echo pattern in the testing set, so this categorization had no true-positives or false-negatives included in the AP calculation.

5.2 Breast Density Classification

Table 10: AUROC (95% confidence interval) of the BI-RADS mammographic breast density categories for each of the three methods on the held-out test set. (LogReg: logistic regression, MLP: gray-level multi-layer perceptron. 95% confidence intervals were computed using DeLong’s method.

	LogReg	**MLP**	**CNN**
A	0.53 (0.50, 0.57)	0.54 (0.50, 0.57)	0.71 (0.68, 0.74)
B	0.59 (0.58, 0.59)	0.64 (0.63, 0.64)	0.66 (0.65, 0.67)
C	0.57 (0.56, 0.57)	0.62 (0.61, 0.63)	0.65 (0.64, 0.65)
D	0.70 (0.68, 0.72)	0.74 (0.71, 0.76)	0.75 (0.73, 0.77)

The fully convolutional neural network approach (CNN) outperformed both the logistic regression (LogReg) and multi-layer perceptron (MLP) gray-level approaches in identifying all four breast density categories. *Table 10* denotes AUROC values and 95% confidence intervals. Performance was compared using AUROC values on the held-out test set. Due to our labels being four continuous values, we computed the AUROC values on the predicted class only. The four-tuples were condensed into a single value, representing the class for which they predicted the largest probability. The difference is the most striking for the low breast density (low breast cancer risk) women.

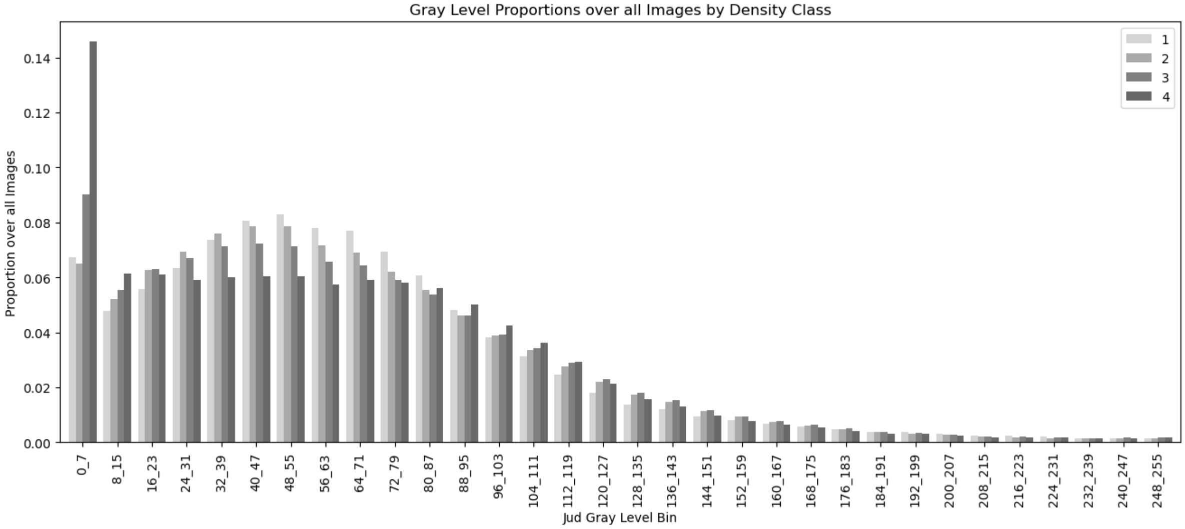
6 DISCUSSION

Our results for the breast density classification and breast lesion detection tasks show promising evidence that deep learning can be used in tandem with breast ultrasound to provide breast cancer risk analysis and diagnosis for women in low-resource areas, such as the USAPI.

In the lesion detection model, performance on the cancer classification task was much better than performance when classifying the other BI-RADS categories. One possible explanation is that cancer classification involves all of the BI-RADS mass lexicon into the final decision, as well as signal from the surrounding tissue, so the task of cancer classification is likely easier than determination of the BI-RADS mass lexicon. A possible area of future work would be implementing cross-talk between the output sub-networks for the mass characteristics and cancer status, due to the inherent dependency between them; i.e., a spiculated lesion is much more likely to be both cancerous and irregular than a lesion with a circumscribed margin. Including predictions of the BI-RADS mass lexicon is a form of explainable AI, wherein the system is providing related classifications which may help to support or discredit its final prediction of lesion cancer status. Another opportunity for future analysis would be the integration of more explicitly explainable techniques into the model structure, such as concept bottlenecks [68].

Our overall objective for the breast lesion detection task is to identify and provide meaningful interpretation of cancer status for malignant lesions, to identify breast cancer earlier. We define our

Figure 9: Bar chart showing the distribution of the pre-defined gray levels for each of the density categories (A:1, B:2, C:3, D:4) in the validation set. Good separability is observed in the lower gray-level categories.



benchmark for acceptable performance as achieving a similar average precision as the best current lesion detection models. The typical performance across the breast lesion detection literature for non-explainable methods is 0.7 mean average precision [40, 41, 69-72]. We believe achieving similar performance is possible given more training data to tune our model with, especially strongly annotated data.

In the breast density classification task, the base methods based on predefined gray-levels in the scan performed surprisingly well, given the simplicity of the models. *Figure 9* shows the distribution of gray-levels over the validation dataset. There is good separability in the lower gray-level bins, indicating that a method based on these bins may be able to identify scan density classes. More investigation is needed into any idiosyncratic differences between scans in each of the scan categories which may be contributing to the performance of the gray-level bin-based methods.

A possible area of future work is to include multiple BUS images for each woman into the determination of her breast density, rather than classifying each image separately. Breast density is a composite measure, typically derived from a mammogram with four different views of the breast tissue, two per breast. Thus, making use of all the scans per woman may provide a more comprehensive view of the breast tissue, leading to a more accurate final classification. This could potentially be a significant advantage for BUS over traditional mammograms, since BUS can quickly generate many views of the breast, and in theory, an intelligent system could help guide the exam administrator to capture better images. Our overall objective for the breast density classification task is to derive a measure of mammographic breast density from BUS which can be used to provide accurate breast cancer risking. For reference, our inter-modality mammographic breast density classification achieves 0.69 mean one vs. rest AUROC. Methods for intra-modality classification of mammographic breast density with deep learning achieve approximately 0.93 mean one vs. rest AUROC on an internal test set [56].

BIBLIOGRAPHY

1. CJ, D.O., et al., *ACR BI-RADS ® Atlas, Breast Imaging Reporting and Data System*. 2013, Reston, VA: American College of Radiology.
2. Akiba, T., et al. *Optuna: A Next-generation Hyperparameter Optimization Framework*. in *International Conference on Knowledge Discovery and Data Mining*. ACM.
3. National Cancer Institute, NIH, and HHS, *Cancer Trends Progress Report*. Bethesda, MD.
4. Yip, C.H., et al., *Improving Outcomes in Breast Cancer for Low and Middle Income Countries.* World Journal of Surgery, 2015. **39**(3): p. 686-692.
5. Heer, E., et al., *Global burden and trends in premenopausal and postmenopausal breast cancer: a population-based study.* The Lancet Global Health, 2020. **8**(8): p. e1027-e1037.
6. Buenconsejo-Lum, L., et al., *Cancer in the U.S. Affiliated Pacific Islands*. 2021, Pacific Regional Central Cancer Registry.
7. Serajuddin, U. and N. Hamadeh, *New World Bank country classifications by income level: 2020-2021.* World Bank Blogs, 2020. **1**.
8. Centers for Disease Control and Prevention, *U.S. Cancer Statistics Female Breast Cancer Stat Bite*, U.D.o.H.a.H. Services, Editor. 2022.
9. Sood, R., et al., *Ultrasound for Breast Cancer Detection Globally: A Systematic Review and Meta-Analysis.* Journal of Global Oncology, 2019(5): p. 1-17.
10. Shen, Y., et al., *Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams.* Nature Communications, 2021. **12**(1).
11. Zhang, G., et al., *SHA-MTL: soft and hard attention multi-task learning for automated breast cancer ultrasound image segmentation and classification.* International Journal of Computer Assisted Radiology and Surgery, 2021. **16**(10): p. 1719-1725.
12. Zhao, C., et al., *Enhancing performance of breast ultrasound in opportunistic screening women by a deep learning-based system: a multicenter prospective study.* Frontiers in Oncology, 2022. **12**: p. 91.
13. Brand, N.R., et al., *Delays and Barriers to Cancer Care in Low- and Middle-Income Countries: A Systematic Review.* The Oncologist, 2019. **24**(12): p. e1371-e1380.
14. Mainiero, M.B., et al., *ACR appropriateness criteria breast cancer screening.* Journal of the American College of Radiology, 2016. **13**(11): p. R45-R49.
15. Schunemann, H.J., et al., *Breast Cancer Screening and Diagnosis: A Synopsis of the European Breast Guidelines.* Annals of internal medicine, 2020. **172**(1): p. 46-+.
16. Organization, W.H., *WHO position paper on mammography screening*. 2014: World Health Organization.
17. Otto, S.J., et al., *Mammography screening and risk of breast cancer death: a population-based case-control study.* Cancer epidemiology, biomarkers & prevention, 2012. **21**(1): p. 66-73.
18. Tabar, L., et al., *Swedish Two-County Trial: Impact of Mammographic Screening on Breast Cancer Mortality during 3 Decades.* Radiology, 2011. **260**(3): p. 658-663.
19. Hellquist, B.N., 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, 2011. **117**(4): p. 714-722.
20. Weinstein, S.P., et al., *ACR Appropriateness Criteria® Supplemental Breast Cancer Screening Based on Breast Density*. Journal of the American College of Radiology, 2021. **18**(11): p. S456-S473.
21. Block Imaging. *Mammography Machine Price Guide*. 2023; Typical cost ranges for digital mammography systems. Installation and your first year of service coverage are included].

Available from: https://info.blockimaging.com/bid/95356/Digital-Mammography-Equipment-Price-Cost-Info.

22. Lee, C.I., L.E. Chen, and J.G. Elmore, *Risk-based Breast Cancer Screening.* Medical Clinics of North America, 2017. **101**(4): p. 725-741.
23. Rebolj, M., et al., *Addition of ultrasound to mammography in the case of dense breast tissue: systematic review and meta-analysis.* British Journal of Cancer, 2018. **118**(12): p. 1559-1570.
24. Habel, L.A., et al., *Mammographic density in a multiethnic cohort.* Menopause (New York, N.Y.), 2007. **14**(5): p. 891-899.
25. Nie, K., et al., *Age- and race-dependence of the fibroglandular breast density analyzed on 3D MRI.* Medical Physics, 2010. **37**(6Part1): p. 2770-2776.
26. del Carmen, M.G., E. Halpem, and D.B. Kopans, *Mammographic breast density and race.* AMERICAN JOURNAL OF ROENTGENOLOGY-NEW SERIES-, 2007. **188**(4): p. 1147.
27. Bimedis. *ULTRASOUND MACHINES FOR SALE (STATIONARY)*. 2023 [cited 2023 February 8]; Available from: https://bimedis.com/search/search-items/ultrasound-equipment-ultrasound-machines?mtid=0&state[]=1&val73=1&vid[]=73&val73=1&val964=13&vid[]=964&val964=13&buy=0&ps=1&ucur=1&page=3&nsk=&premiumads=1&proads=1&sadverts=1&nice_link=ultrasound-machines-for-pregnancy&object_id=289646&sar[]=-1&sar[]=224&mpage=15.
28. Bimedis. *ULTRASOUND MACHINES FOR SALE*. 2023 [cited 2023 February 8]; Available from: https://bimedis.com/search/search-items/ultrasound-equipment-ultrasound-machines?mtid=0&state[]=1&val73=1&vid[]=73&val73=1&val964=12&vid[]=964&val964=12&buy=0&ps=1&ucur=1&page=4&nsk=&premiumads=1&proads=1&sadverts=1&nice_link=ultrasound-machines-for-pregnancy&object_id=289646&sar[]=-1&sar[]=224&mpage=20.
29. Huang, Y., et al., *Interpretation of breast cancer screening guideline for Chinese women.* Cancer Biology & Medicine, 2019. **16**(4): p. 825-835.
30. Ding, J., et al., *Breast Ultrasound Image Classification Based on Multiple-Instance Learning.* Journal of digital imaging, 2012. **25**(5): p. 620-627.
31. Yeh, C.-K., et al., *A disk expansion segmentation method for ultrasonic breast lesions.* Pattern recognition, 2009. **42**(5): p. 596-606.
32. Zhou, Z., et al., *Classification of Benign and Malignant Breast Tumors in Ultrasound Images with Posterior Acoustic Shadowing Using Half-Contour Features.* Journal of Medical and Biological Engineering, 2015. **35**(2): p. 178-187.
33. Hajipour Khire Masjidi, B., et al., *CT-ML: Diagnosis of Breast Cancer Based on Ultrasound Images and Time-Dependent Feature Extraction Methods Using Contourlet Transformation and Machine Learning.* Computational intelligence and neuroscience, 2022. **2022**: p. 1493847-15.
34. Xing, J., et al., *Using BI-RADS Stratifications as Auxiliary Information for Breast Masses Classification in Ultrasound Images.* IEEE J Biomed Health Inform, 2021. **25**(6): p. 2058-2070.
35. Tanaka, H., et al., *Computer-aided diagnosis system for breast ultrasound images using deep learning.* Physics in medicine & biology, 2019. **64**(23): p. 235013-235013.
36. Lin, Z., et al., *A New Dataset and A Baseline Model for Breast Lesion Detection in*

Ultrasound Videos. 2022.

37. Huang, R., et al., *Extracting keyframes of breast ultrasound video using deep reinforcement learning.* Medical image analysis, 2022. **80**: p. 102490-102490.
38. Yun, J., J. Oh, and I. Yun, *Gradually Applying Weakly Supervised and Active Learning for Mass Detection in Breast Ultrasound Images.* Applied Sciences, 2020. **10**(13): p. 4519.
39. Gao, Y., et al., *Detection and recognition of ultrasound breast nodules based on semi-supervised deep learning: a powerful alternative strategy.* Quantitative Imaging in Medicine and Surgery, 2021. **11**(6): p. 2265-2278.

40. Dai, J., et al., *More Reliable AI Solution: Breast Ultrasound Diagnosis Using Multi-AI Combination.* ArXiv, 2021. **abs/2101.02639**.
41. Meng, H., et al., *DGANet: A Dual Global Attention Neural Network for Breast Lesion Detection in Ultrasound Images.* Ultrasound in medicine & biology, 2023. **49**(1): p. 31-44.
42. U.S. Food and Drug Administration Center for Devices and Radiological Health. *TaiHao Medical Inc BU-CAD Section 510(k) Premarket Notification*. 2021 [cited 2023 February 21]; Available from: https://www.accessdata.fda.gov/cdrh_docs/pdf21/K210670.pdf.
43. U.S. Food and Drug Administration Center for Devices and Radiological Health. *Koios Medical Inc Koios DS Section 510(k) Premarket Notification*. 2021 [cited 2023 February 21]; Available from: https://www.accessdata.fda.gov/cdrh_docs/pdf21/K212616.pdf.
44. Al-Dhabyani, W., et al., *Dataset of breast ultrasound images.* Data in brief, 2020. **28**: p. 104863.
45. Yap, M.H., et al., *Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks.* IEEE Journal of Biomedical and Health Informatics, 2018. **22**(4): p. 1218-1226.
46. Breast Cancer Surveillance Consortium. *Breast Cancer Surveillance Consortium Risk Calculator*. 2016 January 31, 2016 [cited 2023 February 21]; Available from: https://tools.bcsc-scc.org/BC5yearRisk/.
47. *Tyrer-Cuzick Risk Assessment Calculator*. 2021 [cited 2023 February 21]; Available from: https://ibis-risk-calculator.magview.com/.
48. Kerlikowske, K., et al., *Are Breast Density and Bone Mineral Density Independent Risk Factors for Breast Cancer?* JNCI : Journal of the National Cancer Institute, 2005. **97**(5): p. 368-374.
49. Kim, W.H., et al., *Ultrasonographic assessment of breast density.* Breast Cancer Research and Treatment, 2013. **138**(3): p. 851-859.
50. O'Flynn, E.A.M., et al., *Ultrasound Tomography Evaluation of Breast Density.* Investigative Radiology, 2017. **52**(6): p. 343-348.
51. Glide, C., N. Duric, and P. Littrup, *Novel approach to evaluating breast density utilizing ultrasound tomography.* Medical physics, 2007. **34**(2): p. 744-753.
52. Yoon, J.H. and E.-K. Kim, *Deep Learning-Based Artificial Intelligence for Mammography.* Korean Journal of Radiology, 2021. **22**(8): p. 1225.
53. *ACR Data Science Institute AI Central*. [cited 2022 July 27]; Available from: https://aicentral.acrdsi.org/.
54. Jud, S.M., et al., *Correlates of mammographic density in B-mode ultrasound and real time elastography.* European Journal of Cancer Prevention, 2012. **21**(4): p. 343-349.
55. Dutta, A. and A. Zisserman. *The VIA Annotation Software for Images, Audio and Video*. in *Proceedings of the 27th ACM International Conference on Multimedia*. 2019. ACM.
56. Wu, N., et al. *Breast Density Classification with Deep Convolutional Neural Networks*. in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018. IEEE.
57. Martin, K.E., et al., *Mammographic density measured with quantitative computer-aided method: Comparison with radiologists' estimates and BI-RADS categories.* Radiology, 2006. **240**(3): p. 656-665.
58. Østerås, B.H., et al., *Classification of fatty and dense breast parenchyma: comparison of automatic volumetric density measurement and radiologists' classification and their inter-observer variation.* Acta radiologica (1987), 2016. **57**(10): p. 1178-1185.
59. Damases, C.N.M.R., et al., *Mammographic Breast Density Assessment Using Automated Volumetric Software and Breast Imaging Reporting and Data System (BIRADS) Categorization by Expert Radiologists.* Academic radiology, 2016. **23**(1): p. 70-77.
60. Shamout, F.E., et al. *The NYU Breast Ultrasound Dataset v1.0*. 2021.

61. Paszke, A., et al., *PyTorch: An Imperative Style, High-Performance Deep Learning Library*, H. Wallach, et al., Editors. 2019.

62. Wu, Y., et al., *Detectron2*. https://github.com/facebookresearch/detectron2.

63. He, K., et al. *Mask R-CNN*. 2017. arXiv:1703.06870.

64. Lin, T.-Y., et al. *Feature Pyramid Networks for Object Detection*. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017. IEEE.

65. detectron2 contributors. *Yaml Config References*. 2020 [cited 2023 March 10]; Available from: https://detectron2.readthedocs.io/en/latest/modules/config.html.

66. Riba, E., et al., *Kornia: an Open Source Differentiable Computer Vision Library for PyTorch.* Winter Conference on Applications of Computer Vision, 2020.

67. Valdez, D., et al. *Can artificial intelligence derived ultrasound breast density provide comparable breast cancer risk estimates to density derived from mammograms*. in *San Antonio Breast Cancer Symposium*. 2022. San Antonio, TX.

68. Koh, P.W., et al., *Concept Bottleneck Models*. 2020.

69. Li, J., et al., *Development of a Deep Learning–Based Model for Diagnosing Breast Nodules With Ultrasound.* Journal of ultrasound in medicine, 2021. **40**(3): p. 513-520.

70. Lin, Z., et al. *A New Dataset and a Baseline Model for Breast Lesion Detection in Ultrasound Videos*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2022. Springer.

71. Zhang, X., et al., *Artificial Intelligence Medical Ultrasound Equipment: Application of Breast Lesions Detection.* Ultrasonic Imaging, 2020. **42**(4-5): p. 191-202.

72. Zhang, E., et al. *Boundary-aware Semi-supervised Deep Learning for Breast Ultrasound Computer-Aided Diagnosis*. in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2019. IEEE.