Prediction of mammographic breast density based on clinical breast ultrasound images using deep learning: a retrospective analysis

Arianna Bunnell,[a,b] Dustin Valdez,[b] Thomas K. Wolfgruber,[b] Brandon Quon,[b] Kailee Hung,[a] Brenda Y. Hernandez,[b] Todd B. Seto,[c] Jeffrey Killeen,[d] Marshall Miyoshi,[e] Peter Sadowski,[a] and John A. Shepherd[b,*]

[a]Department of Information and Computer Sciences, University of Hawai'i at Mānoa, POST Bldg, 1860 East-West Rd, Honolulu, HI, 96822, USA
[b]University of Hawai'i Cancer Center, 701 Ilalo St, Honolulu, HI, 96813, USA
[c]The Queen's Medical Center, 1301 Punchbowl Street, Honolulu, HI, 96813, USA
[d]Hawai'i Pacific Health, 55 Merchant St., Honolulu, HI, 96813, USA
[e]Hawai'i Diagnostic Radiology Services (St. Francis), 2230 Liliha Street, Suite 106, Honolulu, HI, 96817, USA

The Lancet Regional Health - Americas 2025;46: 101096

Published Online xxx
https://doi.org/10.1016/j.lana.2025.101096

Summary

Background Breast density, as derived from mammographic images and defined by the Breast Imaging Reporting & Data System (BI-RADS), is one of the strongest risk factors for breast cancer. Breast ultrasound is an alternative breast cancer screening modality, particularly useful in low-resource, rural contexts. To date, breast ultrasound has not been used to inform risk models that need breast density. The purpose of this study is to explore the use of artificial intelligence (AI) to predict BI-RADS breast density category from clinical breast ultrasound imaging.

Methods We compared deep learning methods for predicting breast density directly from breast ultrasound imaging, as well as machine learning models from breast ultrasound image gray-level histograms alone. The use of AI-derived breast ultrasound breast density as a breast cancer risk factor was compared to clinical BI-RADS breast density. Retrospective (2009–2022) breast ultrasound data were split by individual into 70/20/10% groups for training, validation, and held-out testing for reporting results.

Findings 405,120 clinical breast ultrasound images from 14,066 women (mean age 53 years, range 18–99 years) with clinical breast ultrasound exams were retrospectively selected for inclusion from three institutions: 10,393 training (302,574 images), 2593 validation (69,842), and 1074 testing (28,616). The AI model achieves AUROC 0.854 in breast density classification and statistically significantly outperforms all image statistic-based methods. In an existing clinical 5-year breast cancer risk model, breast ultrasound AI and clinical breast density predict 5-year breast cancer risk with 0.606 and 0.599 AUROC (DeLong's test p-value: 0.67), respectively.

Interpretation BI-RADS breast density can be estimated from breast ultrasound imaging with high accuracy. The AI model provided superior estimates to other machine learning approaches. Furthermore, we demonstrate that age-adjusted, AI-derived breast ultrasound breast density provides similar predictive power to mammographic breast density in our population. Estimated breast density from ultrasound may be useful in performing breast cancer risk assessment in areas where mammography may not be available.

Funding National Cancer Institute.

Keywords: Breast cancer; Ultrasound; Breast cancer screening; Breast density; BI-RADS; Breast cancer risk; Artificial intelligence; Low- and middle-income countries

Abbreviations: AUROC, area under the receiver operating characteristic curve; AI, artificial intelligence; BI-RADS, breast imaging-reporting and data system; ABUS, automated breast ultrasound; HIPIMR, Hawai'i and Pacific Islands Mammography Registry

*Corresponding author.
E-mail address: jshepherd@cc.hawaii.edu (J.A. Shepherd).

Research in context

Evidence before this study

Age-adjusted mammographic breast density is one of the strongest risk factors for breast cancer and several artificial intelligence (AI) approaches exist for automatic breast density estimation from digital mammography imaging. We reviewed the available evidence on breast density estimation from breast imaging by performing a literature search in PubMed and Google Scholar (2010–2024) using the following terms: "breast" AND "density" AND ("ai" or "artificial intelligence" OR "convolutional" OR "deep learning" OR "machine learning" OR "network"). While numerous existing studies explore the use of deep learning for estimation of breast density from other breast imaging modalities, we did not find any studies using deep learning for estimation of breast density from clinical breast ultrasound imaging. When using breast ultrasound for primary screening, such as in low-resource medical contexts, breast density is not available clinically, or using deep learning approaches. Lack of mammographic breast density limits the breast cancer risk evaluation which can be done. Prior work has shown that breast density can be estimated from modalities other than mammography, with varying levels of success. In ultrasound, speed-of-sound using specialized hardware or methods using automated/3D ultrasound (ABUS) have been the most successful. These methods have limited applicability in low-resource contexts (for ABUS) or require specialized hardware (speed of sound).

Added value of this study

Our study proposes an AI method for estimation of BI-RADS mammographic breast density category from handheld breast ultrasound imaging and demonstrates the efficacy of breast density estimation from ultrasound in one existing 5-year breast cancer risk model. This method could be used to provide more complete breast cancer risk evaluation in rural and low-resource areas where mammography is not available, or in any context where women are primarily being screened with breast ultrasound.

Implications of all the available evidence

Breast ultrasound imaging contains information about mammographic breast density. Our study shows that BI-RADS mammographic breast density can be accurately estimated from clinical breast ultrasound data using artificial intelligence. Estimated breast density may be useful in performing breast cancer risk assessment in rural and low-resource areas, where mammography may not be available.

Introduction

Other than age and sex, mammographic breast density is one of the strongest risk factors for breast cancer. Breast density has been extensively studied over the past 30 years and is included in many breast cancer risk models.[1–3] A recent meta-analysis showed that having extremely dense breasts increased lifetime breast cancer risk by 2.11 times over women with scattered dense breast tissue (Breast Imaging Reporting & Data System; BI-RADS density B), even when adjusted for body mass index (BMI) and age.[4] High breast density lowers the sensitivity of mammography leading to a federal mandate to report breast density to all women undergoing screening mammography.[5] Thus, many women have direct knowledge of their breast density and can apply it to readily available risk models. Women at high risk of developing breast cancer have multiple options for lowering their risk including lifestyle choices,[6] and chemoprevention.[7–9] Women at high risk may also elect to participate in supplemental screening with magnetic resonance imaging or contrast-enhanced mammography in high-resource settings.[10,11] However, there are many parts of the world that are either too resource-limited or too remote to have access to screening mammography programs. Without mammography, clinical breast density has not been available for comprehensive breast cancer risk assessment. Estimation of breast density has been explored through other modalities, including ultrasound tomosynthesis,[12–14] microwave breast imaging,[15] dual X-ray absorptiometry,[16,17] optical spectroscopy,[17] and bioimpedance.[18,19] However, unlike clinical breast ultrasound, these methods either do not provide cancer detection capabilities in addition to density assessment, are too expensive for practical use in resource-limited areas, or are stationary, making them unsuitable for use in remote locations.

Breast ultrasound is a sensitive method for detecting breast cancer and is broadly used around the world as primary screening in lower resource settings,[20–22] secondary screening for women with dense breast tissue,[23–25] and follow-up imaging.[26,27] However, mammography is preferred, when available, due to its lower false-positive rate,[28] early detection capabilities, and imaging consistency. Ideally, breast density would be available from accessible breast imaging technologies, like breast ultrasound, with a calibration equivalent to mammographic density.

Image contrast in B-mode breast ultrasound is based on the boundaries where density changes. A change in brightness on the beamformed image corresponds to a change in acoustic impedance in the examined tissue, which implies a change in density. See Fig. 1 for examples of matched breast ultrasound and mammography images at different mammographic densities. These boundary features have been investigated in the past to derive density-like measures. Jud et al.[29] used linear regression from the gray-level values of breast ultrasound imaging normalized into 16 bins to predict

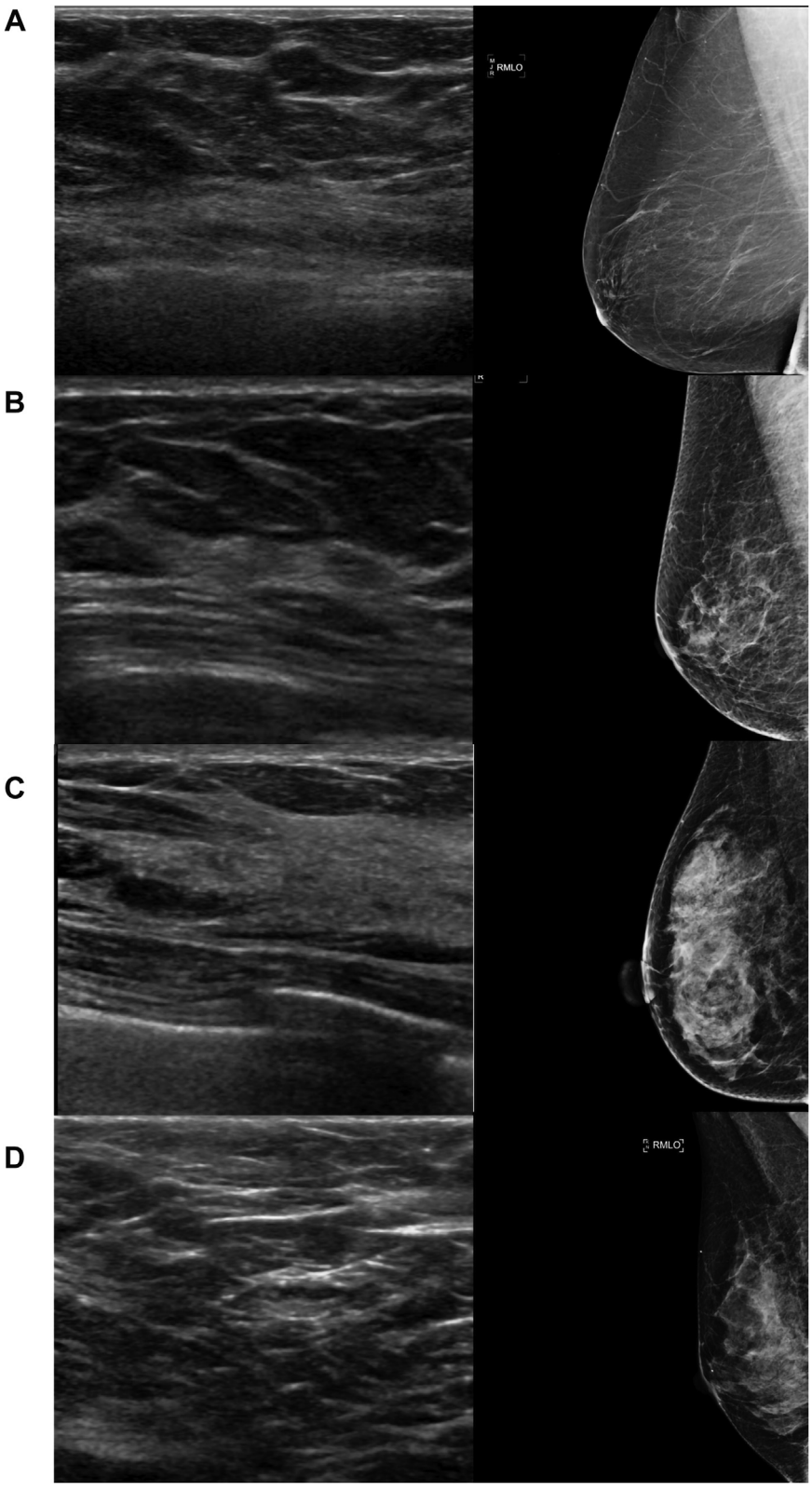

Fig. 1: Example clinical breast ultrasound (left) and right mediolateral oblique digital 2D mammography (right) images, selected from the dataset included in this work. Image sets are labeled with their BI-RADS breast density category. Images have been cropped for ease of display.

percent mammographic density. Their model was able to estimate percent mammographic density with an accuracy of $R^2 = 0.67$ but suffered from calibration issues in external datasets.[30]

In the past decade, there has been a revolution in machine learning and convolutional neural deep learning networks with substantial literature on predicting mammographic breast density from mammograms using artificial intelligence (AI)[31–33] and deriving breast cancer risk from mammograms directly.[34–37] The purpose of this study is to explore the use of AI to estimate breast density from breast ultrasound images. We further explore the use of our AI-derived breast ultrasound breast density in place of mammographic density in an existing breast cancer risk model.[1] Our hypothesis is that AI-derived breast ultrasound breast density is highly associated to mammographic density and is similar to mammographic density in its value in breast cancer risk assessment.

Methods

Study sample

This study follows a cohort design. All women included in this study participated in either screening or diagnostic breast ultrasound imaging at one of the three clinical sites in the Hawai'i and Pacific Islands Mammography Registry (HIPIMR; see Supplementary Methods) from 2009 to 2022. Women were identified by the patient's legal sex in their clinical site's electronic medical record. The HIPIMR prospectively collects breast imaging, demographics, and clinical risk factors. Race/ethnicity data were not retrieved due to data incompleteness for some clinical sites. Women were retrospectively selected for inclusion if they met all the following criteria: (a) had at least one negative screening 2-dimensional mammography visit (not tomosynthesis); (b) had a negative, benign, or probably benign breast ultrasound visit within one year of their mammograms; (c) had a clinical BI-RADS breast density from their examining radiologist; (d) had the standard four views captured at their screening mammogram; and (e) had no history of breast cancer prior to screening mammogram (See Supplementary Methods for complete inclusion/exclusion criteria). From these, cases were defined as women diagnosed with invasive breast cancer at least 6 months and no more than 5 years from their extracted breast ultrasound exam. Non-cases were selected from women who did not develop cancer following their screening examination, matched to cases at a 5:1 non-case:case ratio on year of screening mammography. These cases and non-cases comprise the testing set. All other women meeting the inclusion criteria comprise the development dataset. See Fig. 2 for the complete data flow diagram.

Imaging data preparation

Clinical breast ultrasound data are highly noisy data, with many artifacts such as Color Doppler blood flow highlighting, sonographer text annotations, and lesion caliper markers which can interfere with AI model learning. We implement a cleaning pipeline developed in-house called BUSClean[38] to remove these artifacts and standardize breast ultrasound images converted from DICOM to PNG.

Data splitting

The development dataset was randomly split by woman, stratified by clinical BI-RADS breast density category, with 80% assigned to the training data set and 20% assigned to the validation split for hyperparameter tuning and early stopping determination. The matched testing set was reserved for performance evaluation. Each woman and her images are only present in a single data split. We duplicated the validation dataset to create *curated* and *uncurated* validation sets. In brief, the *curated* validation set has extensive preprocessing applied to remove clinical artifacts, while the *uncurated* validation set had only invalid scans removed and scans with multiple views split. See Supplementary Methods for details. χ^2 tests ($\alpha = 0.05$) are used to test the similarity of distributions of patient-level categorial variables (clinical BI-RADS density, menopausal status) across data splits. Kruskal-Wallis[39] tests ($\alpha = 0.05$) are used to test the similarity of distributions of continuous patient-level variables (age at breast ultrasound, age at mammogram, median gap between mammogram and breast ultrasound) across data splits.

Deep learning model

Three deep learning models were trained in PyTorch (1.12.1)[40] and Lightning-AI (1.9.4)[41] using three distinct ImageNet-pretrained architectures: DenseNet121,[42] ViT-B/32,[43] and ResNet50,[44] selected to represent a diverse range of deep learning architecture designs. Hyperparameters for each were optimized over 25 trials using the TPESampler in Optuna (3.1.0).[45] Supplementary Table S1 displays the search space. The best-performing model architecture and hyperparameters on the curated validation set were selected and retrained for the final model (see Supplementary Methods for training configuration). Models were trained to predict clinical BI-RADS breast density.

Image histogram models

Prior work in estimating percentage mammographic breast density from B-mode breast ultrasound imaging from Jud et al. makes use of gray-level image histograms with equally-sized intervals in a linear regression model.[29] We implement a version of their method adapted for the categorical BI-RADS breast density measure (see Supplementary Methods). Logistic regression, random forest, and multi-layer perceptron (MLP) models were constructed using the gray-level image histogram intervals as features to predict clinical BI-RADS breast density.

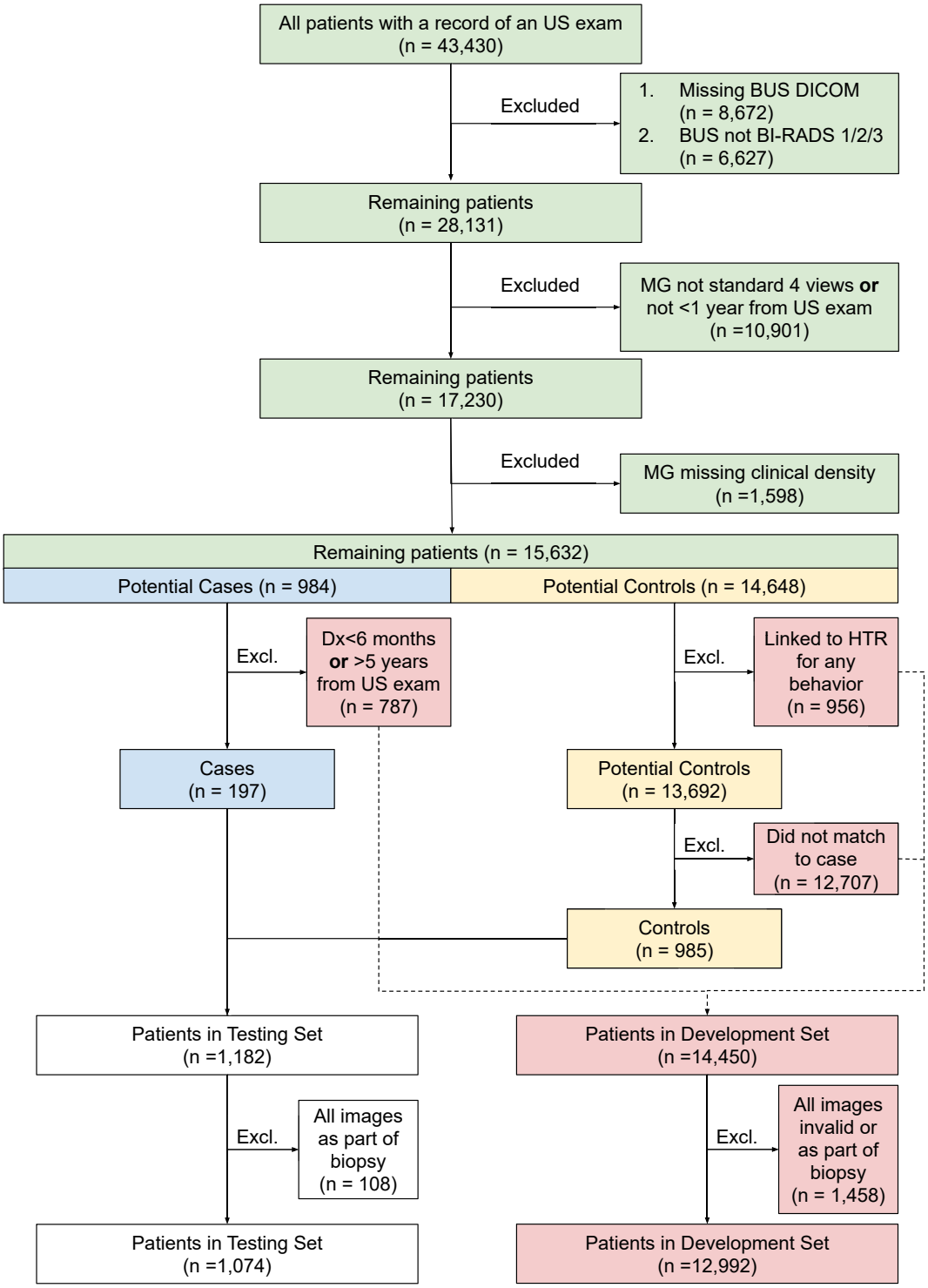

Fig. 2: Flow diagram showing data selection process from the HIPIMR. Cases and controls were selected from the same pool of 15,632 patients with matched mammograms with density records and BUS exams. 197 cases and 965 controls, matched on mammogram year, were selected for the testing set. After exclusion of invalid images, 1,074 women remained. The remaining 12,992 eligible women with valid images were allocated to the development set. Controls were excluded for linkage to the HTR for any behavior if there was record of ductal carcinoma in situ (even if they did not match for invasive cancer) in the HTR. Dx = diagnosis; MG = mammogram; BUS = breast ultrasound; Excl. = excluded; HTR = Hawai'i Tumor Registry.

Statistical analysis

Statistical analysis was undertaken in two steps. The first step demonstrates AI model accuracy in predicting clinical breast density category from breast ultrasound. The second step demonstrates utility of predicted breast ultrasound BI-RADS breast density category in breast cancer risk modeling. See the Supplementary Methods for implementation details. To assess AI model performance in predicting breast density from breast ultrasound, all models are evaluated by their micro-averaged AUROC over all density categories with 95% confidence intervals estimated using DeLong's method[46,47] on the testing set. We use DeLong's test[48] for comparing AUROC values pairwise between the image histogram and deep learning models at $\alpha = 0.004$ to account for multiple testing (12 total tests). We use micro-averaged AUROC for increased stability under class imbalance. We present AUROC on the overall testing set as well by breast ultrasound machine manufacturer, binned patient age at breast ultrasound exam, cancer status, and breast ultrasound exam BI-RADS category. We additionally compute Kendall's τ-b,[49] a non-parametric statistic ranging from –1 to 1 used to measure ordinal agreement, between clinical density and predicted density from breast ultrasound. As a comparator, we compute mammographic density using the deep learning method presented in Wu et al.[31] on all included women from the four standard mammography views and compute Kendall's τ-b between these predictions and clinical density using pre-existing interpretation guidelines where [–1, 0] represents a lack of agreement; (0, 0.20] represents poor agreement; (0.20, 0.40] represents fair agreement; (0.40, 0.60] represents moderate agreement; (0.60, 0.80] represents good agreement; and (0.80, 1] represents strong agreement.[50] Statistical analysis was done in Python using the confidence interval,[46] scikit-learn (1.2.1),[51] and scipy (1.10.0)[52] packages.

To demonstrate utility in cancer risk prediction, we use patient age at breast ultrasound exam and predicted breast density in the Tyrer-Cuzick breast cancer risk model.[1,53] We use DeLong's test[48] for comparing AUROC values pairwise between clinical density, breast ultrasound AI density, and Wu et al. AI density from mammograms.[31] For comparison of performance in cancer risk prediction, we exclude all non-cases with less than 5 years of follow-up and all women >80 years old or <40 years old at breast ultrasound imaging from the testing set.

The funders of the study had no role in the study design, data collection, data analysis, data interpretation or writing of the report.

Results

Study sample characteristics

The characteristics of all 14,066 women (15,632 identified for inclusion from the HIPIMR) included in the study (mean age at breast ultrasound exam ± standard deviation, 53 ± 12 years) are described in Table 1. The median time from first visit in the HIPIMR until diagnosis was 5.2 years (interquartile range: 4.3 years). Controls had a median observation from first visit in the HIPIMR of 9.5 years (interquartile range: 4.3 years). See the Supplementary Methods for complete inclusion/exclusion criteria-specific counts. 405,120 total breast ultrasound images were included in this study. No variables were found to be significantly differently distributed between the data splits on the patient level.

Deep learning model

Based on superior performance in the curated validation set, DenseNet121 was chosen as the final deep learning model architecture (see Supplementary Table S1). Performance with confidence interval estimates on the complete testing set are reported in Table 2. DenseNet121 identified patients with clinical BI-RADS breast density category A with AUROC 0.84, B with AUROC 0.79, C with AUROC 0.72, and D with AUROC 0.88 on the unseen test set. Additionally, the DenseNet121 identified dense (BI-RADS C & D) from non-dense (BI-RADS A & B) women with AUROC 0.823 (0.797, 0.848). Overall micro-averaged AUROC was 0.854 (0.842, 0.866). Fig. 3 displays receiver operating characteristic curves for the DenseNet121 on the unseen test set, separated by clinical BI-RADS mammographic density category. Model performance stratified by patient age, breast ultrasound machine manufacturer, cancer status, and breast ultrasound exam diagnostic BI-RADS category can be found in Supplementary Table S3. Supplementary Fig. S1 shows example saliency maps highlighting the areas of the breast ultrasound scans which may be of particular importance for the deep learning model's predictions on the unseen test set, computed using integrated gradients.[54,55]

Image histogram models

Performance results for the baseline random forest and logistic regression models (see Supplementary Table S3 for model selection and tuning details) are presented in Table 2. Logistic regression from the gray-level image histograms identified patients in the held-out testing set with AUROCs ranging from 0.621 (BI-RADS C) to 0.668 (BI-RADS D). The image histogram random forest model identified patients in the held-out testing set with AUROCs ranging from 0.575 (BI-RADS C) to 0.727 (BI-RADS D). See Table 2.

Two-sided DeLong's tests[48] were performed ($\alpha = 0.004$ to account for multiple testing) to compare AUROC values for each clinical BI-RADS breast density category (A, B, C, D) between all three models (logistic regression, random forest, and DenseNet121). Results are reported in Table 3. DenseNet121 was found to perform significantly better in breast density classification (measured by AUROC) compared to both image histogram models, across all density categories.

	Overall	Training set	*Curated* validation set[a]	Testing set	p-value
Women, N	14,066	10,393	2593	1074	N/A
Women with benign findings, N (%)	–	–	–	896 (83.4)	
Women with malignant findings, N (%)	–	–	–	178 (16.6)	
Mean age at US, years (SD)	53.4 (11.9)	53.4 (11.9)	53.3 (12.0)	53.8 (12.1)	0.42
Mean age at MG, years (SD)	53.4 (11.9)	53.4 (11.9)	53.2 (12.0)	53.8 (12.1)	0.42
Mean age at diagnosis, years (SD)	–	–	–	60.4 (11.4)	N/A
Women with fatty/A breasts, N (%)	478 (3.4)	344 (3.3)	99 (3.8)	35 (3.3)	0.70
Women with scattered/B breasts, N (%)	5486 (39.0)	4088 (39.3)	987 (38.1)	409 (38.1)	
Women with heterogen./C breasts, N (%)	6291 (44.7)	4615 (44.4)	1178 (45.4)	494 (45.9)	
Women with dense/D breasts, N (%)	1811 (12.9)	1346 (13.0)	329 (12.7)	136 (12.7)	
Menopausal women, N (%)	4365 (31.0)	3210 (30.9)	806 (31.1)	345 (32.1)	0.19
Premenopausal women, N (%)	480 (3.4)	347 (3.3)	83 (3.2)	49 (4.6)	
Unknown menopausal status, N (%)	9221 (65.6)	6836 (65.8)	1704 (65.7)	680 (63.3)	
Median gap btw. MAM & US, days (IQR)	5.0 (20.0)	5.0 (20.0)	4.0 (19.0)	6.0 (19.8)	0.37
Median gap btw. MAM & Dx, days (IQR)	–	–	–	822.0 (816.8)	N/A
Median gap btw. US & Dx, days (IQR)	–	–	–	808.5 (808.5)	
Images, N	405,120	302,574	69,842	28,616	
Mean no. of images per woman, N (SD)	28.8 (22.9)	29.1 (23.1)	26.9 (21.8)	26.6 (21.9)	
Images with benign findings, N (%)	–	–	–	23,901 (83.5)	
Images with malignant findings, N (%)	–	–	–	4715 (16.5)	
US BI-RADS 1 images, N (%)	58,428 (14.4)	43,163 (14.3)	10,856 (15.5)	4042 (14.1)	
US BI-RADS 2 images, N (%)	238,268 (58.8)	177,280 (58.6)	41,313 (59.2)	17,419 (60.9)	
US BI-RADS 3 images, N (%)	108,424 (26.8)	82,131 (27.1)	17,673 (25.3)	7155 (25.0)	
Images of fatty/A breasts, N (%)	11,525 (2.9)	8416 (2.9)	2223 (3.2)	886 (3.1)	
Images of scattered/B breasts, N (%)	151,015 (37.6)	115,394 (38.1)	25,794 (36.9)	9827 (34.3)	
Images of heterogeneous/C breasts, N (%)	180,927 (45.1)	135,622 (44.8)	31,941 (45.7)	13,364 (46.7)	
Images of dense/D breasts, N (%)	57,565 (14.4)	43,142 (14.2)	9884 (14.2)	4539 (15.9)	
Images on PHILIPS System, N (%)	363,308 (89.7)	271,905 (89.9)	64,204 (91.9)	24,687 (86.3)	
Images on ATL System, N (%)	25,297 (6.2)	17,877 (5.9)	3723 (5.3)	3208 (11.2)	
Images on SIEMENS System, N (%)	12,617 (3.1)	9795 (3.2)	1495 (2.1)	443 (1.5)	
Images on TOSHIBA System, N (%)	3763 (0.9)	2862 (0.9)	420 (0.6)	278 (1.0)	
Images on GE System, N (%)	126 (0.0)	126 (0.0)	0 (0.0)	0 (0.0)	
Images on ALOKA System, N (%)	9 (0.0)	9 (0.0)	0 (0.0)	0 (0.0)	

See the supplement for counts of the uncurated validation set. BUS = breast ultrasound; MG = mammogram; Dx = diagnosis; IQR = inter-quartile range; SD = standard deviation. p-values from tests of homogeneity are shown in the rightmost column. N/A represents a row where a test was not performed. [a]The curated validation and testing sets had the complete preprocessing pipeline applied, while the training and uncurated validation sets had only invalid scans removed and dual-view scans split. The curated validation set is a proper subset of the uncurated validation set. The uncurated validation set includes 73,930 images from 2599 women. Complete category-specific counts can be found in the Supplement.

Table 1: **Study sample characteristics.**

	AUROC[a] (95% C.I.)	by clinical BI-RADS breast density category			
	Overall	Fatty/A	Scattered/B	Heterogen./C	Dense/D
DenseNet121					
Per-image	0.840 (0.838, 0.842)	0.857 (0.847, 0.867)	0.761 (0.755, 0.766)	0.673 (0.667, 0.679)	0.839 (0.833, 0.846)
Per-patient	0.854 (0.842, 0.866)	0.841 (0.790, 0.891)	0.787 (0.759, 0.815)	0.719 (0.689, 0.749)	0.879 (0.848, 0.910)
Logistic Reg.					
Per-image	0.771 (0.768, 0.774)	0.594 (0.574, 0.613)	0.607 (0.601, 0.614)	0.584 (0.577, 0.590)	0.583 (0.574, 0.592)
Per-patient	0.795 (0.781, 0.809)	0.642 (0.552, 0.732)	0.654 (0.620, 0.687)	0.621 (0.588, 0.654)	0.668 (0.619, 0.716)
Random forest					
Per-image	0.759 (0.756, 0.762)	0.604 (0.586, 0.623)	0.597 (0.590, 0.604)	0.540 (0.533, 0.546)	0.625 (0.617, 0.634)
Per-patient	0.794 (0.780, 0.808)	0.698 (0.620, 0.776)	0.639 (0.605, 0.673)	0.575 (0.541, 0.609)	0.727 (0.680, 0.774)

See Supplementary Table S3 for subgroup performance. [a]AUROC is a micro-averaged one vs. rest AUROC over all density categories, except for density subgroups. For density subgroups, one vs. rest AUROC for each category is shown. 95% confidence intervals are calculated using DeLong's method.[1,2]

Table 2: **Overall performance statistics for the DenseNet121, logistic regression, and random forest models.**

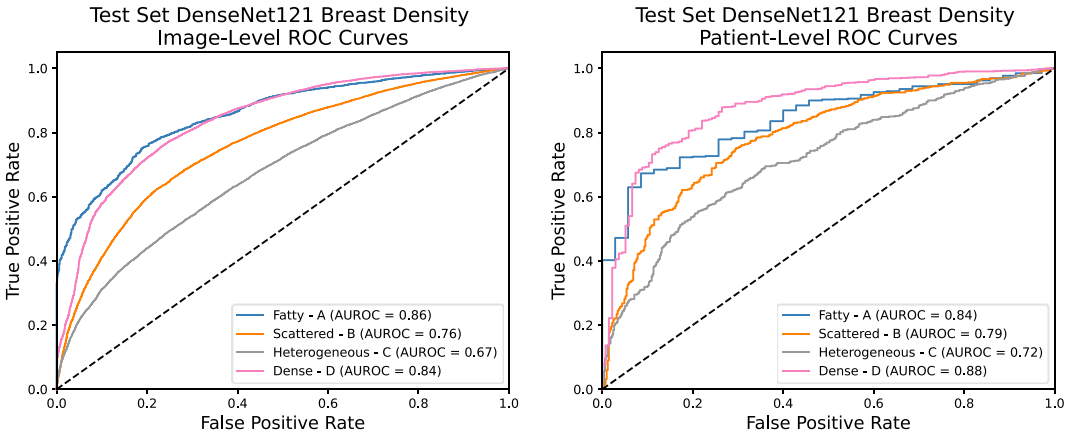

Fig. 3: BUS AI model receiver operating characteristic (ROC) curves for the image-level predictions (left) and when aggregated through arithmetic averaging to patient-level (right).

Agreement to clinical density

Kendall's τ-b between predicted breast ultrasound AI density and clinical density was found to be 0.47 (moderate agreement). Predicted breast ultrasound AI density was calculated per-patient by taking an arithmetic mean over the predicted class with the largest probability from each image and rounding to the nearest integer. Kendall's τ-b between Wu et al.[31] mammographic AI density and clinical density was found to be 0.60 (moderate agreement). Predicted mammographic AI density from Wu et al.[31] was defined as the predicted class with the largest probability.

Cancer risk modeling

The performance benchmark for AI-derived breast ultrasound breast density was the performance of clinical breast density in 5-year breast cancer risk prediction using the Tyrer-Cuzick risk calculator.[1,53] For comparison of performance in cancer risk prediction, we exclude all non-cases with less than 5 years of follow-up (n = 103). We also exclude all women >80 years of age at breast ultrasound examination (n = 27) from the testing set, as the Tyrer-Cuzick model is not validated for women >85 years old. We also exclude women <40 years old (n = 75). 869 women (164 cases) were used in risk prediction. Predefined "missing" placeholder values were used for all other variables except patient age and breast density. In the age alone model, risk was predicted using the predefined "missing" value for breast density. Predicted breast ultrasound AI density was calculated per-patient by taking an arithmetic mean over the predicted class with the largest probability from each image and rounding to the nearest integer. Predicted Wu et al.[31] mammographic AI density was defined as the predicted class with the largest probability.

In the Tyrer-Cuzick risk model, clinical breast density predicted 5-year breast cancer risk with 0.599 AUROC (95% confidence interval (0.552, 0.647)). In the Tyrer-Cuzick risk model, breast ultrasound AI density predicted 5-year breast cancer risk with AUROC 0.606 (0.559, 0.653). In the Tyrer-Cuzick risk model, Wu et al.[31] mammographic AI density predicted 5-year breast cancer risk with AUROC 0.614 (0.568, 0.661). In the Tyrer-Cuzick risk model, age alone predicted 5-year breast cancer risk with AUROC 0.612 (0.564,

	DeLong's test results by clinical BI-RADS breast density category			
	Fatty/A	Scattered/B	Heterogeneous/C	Dense/D
DenseNet121 vs. Logistic Reg.	5.71 (p < 0.0001)**	7.77 (p < 0.0001)**	4.77 (p < 0.0001)**	9.72 (p < 0.0001)**
Logistic Reg. vs. RF	−1.47 (0.14)	0.87 (0.39)	2.25 (0.024)	−3.74 (0.0002)*
DenseNet121 vs. RF	4.21 (p < 0.0001)**	9.54 (p < 0.0001)**	7.57 (p < 0.0001)**	7.29 (p < 0.0001)**

Two-sided DeLong's test statistics are presented followed by their corresponding p-value in parentheses in each cell. p-values less than 0.0001 are represented as "p < 0.0001." RF = Random Forest. For density subgroups, one vs. rest AUROC for each category was calculated. Significant results at the $\alpha = 4 \times 10^{-3}$ level are indicated with a*. p-values less than 4×10^{-5} are indicated with a**. Test statistics can be interpreted as the number of standard deviations (standard normal distribution) by which the observed difference in model AUROCs deviates from zero (null hypothesis).

Table 3: **Results of two-sided DeLong's tests on patient-level, clinical BI-RADS density category-specific AUROC values for each of the image histogram (RF and logistic regression) and deep learning (DenseNet121) models.**

0.659). Supplementary Table S4 presents AUROC values. Two-sided DeLong's tests[48] were performed ($\alpha = 0.05$) to compare AUROC values for predicting 5-year breast cancer risk using age along with one of clinical, breast ultrasound AI, or Wu et al.[31] mammographic AI breast density as well as age alone in the Tyrer-Cuzick risk model.[1,53] DeLong's tests revealed no significant differences in model performance. Supplementary Table S4 presents test statistics and p-values.

Discussion

We found that BI-RADS breast density can be estimated through machine learning methods from breast ultrasound images and that our deep learning model outperformed other approaches in breast density estimation. Our deep learning method was developed on a training dataset of 10,393 women (302,574 images) with hyperparameters chosen on a validation set of 2593 women (69,842 images). On an unseen test dataset of 1074 women, the breast ultrasound AI density model had similar agreement (moderate agreement; τ-b = 0.47) with clinical density as Wu et al.[31] mammographic AI density, an open-source method for predicting density from mammogram directly, had with clinical density (moderate agreement, τ-b = 0.60) On an unseen test dataset of 869 women (164 cases), we found that the top performing breast ultrasound AI breast density estimate had comparable performance to clinical BI-RADS breast density for predicting 5-year breast cancer risk when used in the Tyrer-Cuzick breast cancer risk model.[1,31,53]

Limited prior work has investigated the prediction of mammographic breast density from ultrasound. Tissue speed-of-sound using a dedicated non-diagnostic probe has been proposed for measuring mammographic breast density.[56] It was found that non-dense (BI-RADS classes A & B) could be predicted from dense (BI-RADS C & D) breasts with AUROC = 0.887. However, the application of a dedicated breast ultrasound device for breast density has limited clinical applications or benefit over being able to get breast ultrasound from a clinical hand-held system. Our method can identify patients with extremely dense breasts (class D) with similar performance of AUROC 0.879. When dichotomized, our method identified dense (BI-RADS C & D) from non-dense (BI-RADS A & B) women with AUROC 0.823 (0.797, 0.848). Mammographic percent density has also been measured using tissue speed of sound on a SoftVue (Delphinus Medical Technologies; MI, USA) ultrasound tomosynthesis system with an accuracy of $R^2 = 0.96$.[12–14] Ultrasound tomosynthesis systems are highly specialized equipment with high installation and maintenance costs which may not be appropriate for low-resource and rural areas. Handheld breast ultrasound is most directly applicable to rural and resource-limited scenarios due to portability and low cost. Additional non-mammographic methods of measuring breast density have also been explored, including microwave breast imaging,[15] dual X-ray absorptiometry,[16,17] optical spectroscopy,[17] and bioimpedance.[18,19] These methods, no matter their performance, are less desirable than integrating breast density into clinical breast ultrasound since they do not provide cancer detection in addition to density assessment the way clinical breast ultrasound does. If used in breast cancer screening, these alternative methods would need to be coupled with breast ultrasound or mammography to provide both detection and risk assessment.

Alternative breast ultrasound imaging-based biomarkers of breast cancer risk have been explored other than breast density. Breast parenchymal pattern from ultrasound measures the distribution of fat and ductal tissue in the breast. It can be classified by a breast radiologist into four categories: ductal, heterogeneous, mixed, and fibrous and has been found to be associated with breast cancer risk in the Chinese population.[57,58] Another biomarker of breast cancer risk identified from breast ultrasound imaging is glandular tissue component. Glandular tissue component is also a four-category classification which measures the degree of lobular involution observed in *screening* breast ultrasound.[59] High glandular tissue component is associated with increased risk of breast cancer in women with extremely dense (BI-RADS D) breasts.[59,60] It may be that these measures have complementary risk information to breast density but this has yet to be explored. We pursued breast density since it may be integrated into existing breast cancer risk models such as the Breast Cancer Surveillance Consortium, Tyrer-Cuzick and others. Background echotexture, as defined by the American College of Radiology, is a three-category classification which has been found to be associated with both mammographic density and parity, known risk factors for breast cancer.[61] AI combined with automated breast ultrasound (ABUS) performed with AUROC 0.979 for identifying background echotexture categories,[62] however ABUS may be prohibitively costly for remote and low-resource rural areas.

We tested the association of breast ultrasound AI density to clinical mammographic density by using both measures in an existing clinical breast cancer risk model. No statistically significant pairwise differences were found between performance in predicting 5-year breast cancer risk using clinical density with age, Wu et al. mammographic AI density with age,[31] age alone, or breast ultrasound AI breast density with age. The reported performance for 5-year breast cancer risk prediction on an external cohort is in line with expectations from the literature.[63]

Although our findings are promising, there are limitations to consider. First, there were relatively few

women present in the lowest density category A (4%). This imbalance may be due to the ethnic breakdown of the population of the HIPIMR, with an estimated 60% of women in Hawai'i identifying as Asian (alone or in combination with another race) in 2023.[64] Asian women have been found to have denser breasts than women of other races.[65,66] Future work will explore sample enrichment with category A women as well as inclusion of a geographically external testing set. Another limitation of this work is that evaluation using a clinical risk model only used age and BI-RADS breast density. Inclusion of other breast cancer risk factors may provide a more comprehensive assessment and validation of our method.

The lack of observed effect of BI-RADS breast density in our study in the Tyrer-Cuzick risk model when combined with age, as compared to age alone, may be attributed to the imaging being from diagnostic visits. breast ultrasound is not a primary screening modality in Hawai'i, and all exams included in this study are diagnostic exams requested after an initial mammogram. The characteristics of the diagnostic breast ultrasound population are likely different from those of a screening mammography population, on which most breast cancer risk evaluation is based. Future, external validation is needed in a population screened with breast ultrasound to verify our findings and confirm clinical utility of breast ultrasound -derived breast density in traditional breast cancer risk models.

In conclusion, we found that BI-RADS breast density can be estimated from breast ultrasound BUS images using deep learning. Furthermore, on an unseen test dataset of 869 women (164 cases), we found that the top performing breast ultrasound AI breast density estimate had comparable performance to clinical BI-RADS breast density for predicting 5-year breast cancer risk when used in an existing 5-year breast cancer risk model. Estimated breast density from breast ultrasound may be useful for breast cancer risk assessment in areas where mammography is not available.

Contributors

JAS and AB conceptualized the study. MM, TBS, JK, BQ, KH, and TKW contributed to data curation. BQ, TKW, and AB contributed to formal analysis. JAS and BYH contributed to funding acquisition. AB and DV contributed to methodology. JAS and PS contributed to project administration. MM, TBS, JK, BQ, JAS, BYH, and TKW provided resources. KH, TKW, and AB contributed to software. JAS and PS supervised the study. AB drafted the manuscript. BV, TKW, BYH, JAS, and PS contributed to review and editing. AB, TKW, and BQ have accessed and verified the underlying data reported in the manuscript. JAS, PS, and AB were responsible for the decision to submit the manuscript.

Data sharing statement

The data that support the findings of this study are available from the Hawai'i and Pacific Islands Mammography Registry (HIPIMR), but restrictions apply to the availability of these data. Deidentified data can be requested for research use at any time through the HIPIMR's website at https://hipimr.shepherdresearchlab.org/. All code for running the trained models, creating the gray-level comparator method, and evaluation is available at https://github.com/hawaii-ai/bus-density.

Ethics approval and consent to participate

This study was approved by the WCG IRB (Study Number 1264170). The requirement for informed consent was waived. A protocol was not registered for this study. There was no patient or public involvement in the design of this study.

Consent for publication

Not applicable.

Declaration of interests

JAS has received grants from the National Cancer Institute, National Institutes of Health, and Johns Hopkins University. JAS has received consulting fees from the University of Washington and Gilead Sciences, Inc. JAS has received support for attending meetings/travel from the International Energy Agency and the International Society of Clinical Densitometry. JAS participated on the Scientific Advisory Board of the International Breast Density and Breast Cancer Risk Assessment Workshop and of Gilead Sciences, Inc. JAS has received materials for research from Hologic, Inc. and Advanced Micro Devices, Inc. All other authors declare no conflicts of interest or financial interest.

Acknowledgements

This study was funded with NIH grants U54CA143728 and R01CA263491.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.lana.2025.101096.

References

1 Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. *Stat Med*. 2004;23(7):1111–1130.
2 Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst*. 1989;81(24):1879–1886.
3 Tice JA, Cummings SR, Smith-Bindman R, Ichikawa L, Barlow WE, Kerlikowske K. Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. *Ann Intern Med*. 2008;148(5):337–347.
4 Bodewes F, Van Asselt A, Dorrius M, Greuter M, De Bock G. Mammographic breast density and the risk of breast cancer: a systematic review and meta-analysis. *Breast*. 2022;66:62–68.
5 Berg WA, Seitzman RL, Pushkin J. Implementing the national dense breast reporting standard, expanding supplemental screening using current guidelines, and the proposed Find it Early Act. *J Breast Imaging*. 2023;5(6):712–723.
6 Maramara T, Matos MC, Ardila S, Phantana-angkool A, Henry D. Minimizing breast cancer risk with diet and exercise. *Curr Breast Cancer Rep*. 2024;16(1):45–52.
7 Zaluzec EK, Sempere LF. Systemic and local strategies for primary prevention of breast cancer. *Cancers*. 2024;16(2):248.
8 Cuzick J, Chu K, Keevil B, et al. Effect of baseline oestradiol serum concentration on the efficacy of anastrozole for preventing breast cancer in postmenopausal women at high risk: a case-control study of the IBIS-II prevention trial. *Lancet Oncol*. 2024;25(1):108–116.
9 Yogendran L, Meis L, Burnside E, Schrager S. Management of women at high risk for breast cancer. *J Am Board Fam Med*. 2024;36(6):1029–1032.
10 Monticciolo DL, Newell MS, Moy L, Lee CS, Destounis SV. Breast cancer screening for women at higher-than-average risk: updated recommendations from the ACR. *J Am Coll Radiol*. 2023;20(9):902–914.

11 Onishi N, Kataoka M. Breast cancer screening for women at high risk: review of current guidelines from leading specialty societies. *Breast Cancer*. 2021;28(6):1195–1211.
12 Duric N, Sak M, Fan S, et al. Using whole breast ultrasound tomography to improve breast cancer risk assessment: a novel risk factor based on the quantitative tissue property of sound speed. *J Clin Med*. 2020;9(2):367.
13 Glide C, Duric N, Littrup P. Novel approach to evaluating breast density utilizing ultrasound tomography. *Med Phys*. 2007;34(2):744–753.
14 O'Flynn EAM, Fromageau J, Ledger AE, et al. Ultrasound tomography evaluation of breast density. *Investig Radiol*. 2017;52(6):343–348.
15 Fear E, Bourqui J, Mojabi P, et al. Abstract PO3-07-08: breast density estimation with a microwave-frequency imaging system. *Cancer Res*. 2024;84(9_Supplement):PO3-07-8–PO3–8.
16 Shepherd JA, Malkov S, Fan B, Laidevant A, Novotny R, Maskarinec G. Breast density assessment in adolescent girls using dual-energy X-ray absorptiometry: a feasibility study. *Cancer Epidemiol Biomark Prev*. 2008;17(7):1709–1713.
17 Lloyd R, Pirikahu S, Walter J, et al. Alternative methods to measure breast density in younger women. *Br J Cancer*. 2023;128(9):1701–1709.
18 Maskarinec G, Morimoto Y, Laguaña MB, Novotny R, Guerrero RTL. Bioimpedence to assess breast density as a risk factor for breast cancer in adult women and adolescent girls. *Asian Pac J Cancer Prev*. 2016;17(1):65.
19 Gutiérrez-Lopez M, Prado-Olivarez J, Matheus-Troconis C, et al. A case study in breast density evaluation using bioimpedance measurements. *Sensors*. 2022;22(7):2747.
20 Yang L, Wang S, Zhang L, et al. Performance of ultrasonography screening for breast cancer: a systematic review and meta-analysis. *BMC Cancer*. 2020;20(1).
21 Sood R, Rositch AF, Shakoor D, et al. Ultrasound for breast cancer detection globally: a systematic review and meta-analysis. *J Glob Oncol*. 2019;(5):1–17.
22 He J, Chen W, Li N, et al. China guideline for the screening and early detection of female breast cancer (2021, Beijing). *Zhonghua Zhong Liu Za Zhi*. 2021;43(4):357–382.
23 Mainiero MB, Lourenco A, Mahoney MC, et al. ACR appropriateness criteria breast cancer screening. *J Am Coll Radiol*. 2016;13(11):R45–R49.
24 Urban LABD, Chala LF, Bauab SdP, et al. Breast cancer screening: updated recommendations of the Brazilian College of Radiology and diagnostic imaging, Brazilian breast disease society, and Brazilian federation of gynecological and obstetrical Associations. *Radiol Bras*. 2017;50:244–249.
25 Ren W, Chen M, Qiao Y, Zhao F. Global guidelines for breast cancer screening: a systematic review. *Breast*. 2022;64:85–99.
26 Evans A, Trimboli RM, Athanasiou A, et al. Breast ultrasound: recommendations for information to women and referring physicians by the European Society of Breast Imaging. *Insights Imaging*. 2018;9:449–461.
27 Radiology ACo, *ACR practice parameter for the performance of a diagnostic breast ultrasound examination*. American College of Radiology Website; 2021.
28 Berg WA, Bandos AI, Mendelson EB, Lehrer D, Jong RA, Pisano ED. Ultrasound as the primary screening test for breast cancer: analysis from ACRIN 6666. *J Natl Cancer Inst*. 2016;108(4):djv367.
29 Jud SM, Häberle L, Fasching PA, et al. Correlates of mammographic density in B-mode ultrasound and real time elastography. *Eur J Cancer Prev*. 2012;21(4):343–349.
30 Behrens A, Fasching PA, Schwenke E, et al. Predicting mammographic density with linear ultrasound transducers. *Eur J Med Res*. 2023;28(1):384.
31 Wu N, Geras KJ, Shen Y, et al. Breast density classification with deep convolutional neural networks. In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP); 2018 2018-04-01*. IEEE; 2018.
32 Ciatto S, Bernardi D, Calabrese M, et al. A first evaluation of breast radiological density assessment by QUANTRA software as compared to visual classification. *Breast*. 2012;21(4):503–506.
33 Lehman CD, Yala A, Schuster T, et al. Mammographic breast density assessment using deep learning: clinical implementation. *Radiology*. 2019;290(1):52–58.
34 Yala A, Mikhael PG, Strand F, et al. Toward robust mammography-based models for breast cancer risk. *Sci Transl Med*. 2021;13(578):eaba4373.
35 Yeoh HH, Liew A, Phan R, et al. RADIFUSION: a multi-radiomics deep learning based breast cancer risk prediction model using sequential mammographic images with image attention and bilateral asymmetry refinement. *arXiv*. 2023. preprint arXiv:230400257.
36 Pertuz S, Torres GF, Tamimi R, Kamarainen J. Open framework for mammography-based breast cancer risk assessment. In: *2019 IEEE EMBS international conference on biomedical & Health informatics (BHI); 2019*. IEEE; 2019:1–4.
37 Leong L, Wolfgruber T, Quon B, et al. Image-based models for predicting advanced breast cancer risk. In: *10th international breast density & cancer risk assessment Workshop; 2023 06/07/2023*. Kailua-Kona, HI; 2023.
38 Bunnell A, Hung K, Shepherd JA, Sadowski P. BUSClean: open-source software for breast ultrasound image pre-processing and knowledge extraction for medical AI. *PLoS One*. 2024;19(12):e0315434.
39 Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *J Am Stat Assoc*. 1952;47(260):583–621.
40 Paszke A, Gross S, Massa F, et al. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, eds. *PyTorch: an imperative style, high-performance deep learning library*. 2019.
41 Falcon W. *PyTorch lightning. 1.4 ed*; 2019. https://github.com/PyTorchLightning/pytorch-lightning. The lightweight PyTorch wrapper for high-performance AI research. Scale your models, not the boilerplate.
42 Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition; 2017*. 2017:4700–4708.
43 Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. *arXiv*. 2020. preprint arXiv:201011929.
44 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition; 2016*. 2016:770–778.
45 Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. In: *International conference on knowledge discovery and data mining*. ACM; 2019:2623–2631.
46 Jacob G. *A python library for confidence intervals*. 2023.
47 Sun X, Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process Lett*. 2014;21(11):1389–1393.
48 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44:837–845.
49 Kendall MG. A new measure of rank correlation. *Biometrika*. 1938;30(1–2):81–93.
50 Abadi M, Moore DR. Selection of circular proposals in building projects: an MCDM model for lifecycle circularity assessments using AHP. *Buildings*. 2022;12(8):1110.
51 Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.
52 Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17(3):261–272.
53 *Tyrer-cuzick risk assessment calculator*; 2021. https://ibis-risk-calculator.magview.com/. Accessed February 21, 2023.
54 Kokhlikyan N, Miglani V, Martin M, et al. Captum: a unified and generic model interpretability library for pytorch. *arXiv*. 2020. preprint arXiv:200907896.
55 Qi Z, Khorram S, Li F. Visualizing deep networks by optimizing with integrated gradients. In: *CVPR workshops; 2019*. 2019:1–4.
56 Sanabria SJ, Goksel O, Martini K, et al. Breast-density assessment with hand-held ultrasound: a novel biomarker to assess breast cancer risk and to tailor screening? *Eur Radiol*. 2018;28:3165–3175.
57 Kaizer L, Fishell EK, Hunt JW, Foster FS, Boyd NF. Ultrasonographically defined parenchymal patterns of the breast: relationship to mammographic patterns and other risk factors for breast cancer. *Br J Radiol*. 1988;61(722):118–124.
58 Hou X-Y, Niu H-Y, Huang X-L, Gao Y. Correlation of breast ultrasound classifications with breast cancer in Chinese women. *Ultrasound Med Biol*. 2016;42(11):2616–2621.

59 Lee SH, Moon WK. Glandular tissue component on breast ultrasound in dense breasts: a new imaging biomarker for breast cancer risk. *Korean J Radiol*. 2022;23(6):574.
60 Lee SH, Ryu H-S, Jang M-J, et al. Glandular tissue component and breast cancer risk in mammographically dense breasts at screening breast US. *Radiology*. 2021;301(1):57–65.
61 Kim WH, Lee SH, Chang JM, Cho N, Moon WK. Background echotexture classification in breast ultrasound: inter-observer agreement study. *Acta Radiol*. 2017;58(12):1427–1433.
62 Chang RF, Hou YL, Lo CM, et al. Quantitative analysis of breast echotexture patterns in automated breast ultrasound images. *Med Phys*. 2015;42(8):4566–4578.
63 Brentnall AR, Harkness EF, Astley SM, et al. Mammographic density adds accuracy to both the Tyrer-Cuzick and Gail breast cancer risk models in a prospective UK screening cohort. *Breast Cancer Res*. 2015;17:1–10.
64 State of Hawaii Department of Business Economic Development and Tourism, *Annual estimates of the resident population by sex, race, and hispanic origin for Hawaii: April 1, 2020 to July 1, 2023 (SC-EST2023-SR11H-15)*. 2024.
65 Maskarinec G, Meng L, Ursin G. Ethnic differences in mammographic densities. *Int J Epidemiol*. 2001;30(5):959–965.
66 Habel LA, Capra AM, Oestreicher N, et al. Mammographic density in a multiethnic cohort. *Menopause*. 2007;14(5):891–899.