This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Domain-Specific Augmentation and Uncertainty Quantification for Robust Breast Ultrasound Lesion Segmentation and Classification

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Breast ultrasound holds great potential for cancer screening in resource-limited settings, yet lesion characterization remains challenging due to speckle noise, acoustic shadowing, and limited specialist access. We propose an uncertainty-aware dual-task framework for joint lesion segmentation and malignancy classification on 2,773 images from three public datasets (5-fold cross-validation), comparing nine augmentation combinations across ResNet34, ResNet50, and EfficientNet-B3. We also present DUSpeckNet, a lightweight dual-task network trained from scratch. With ultrasound-specific augmentation, ResNet50 achieved Dice= 0.88 ± 0.01 and EfficientNet-B3 achieved AUC= 0.94 ± 0.02, Accuracy= 0.89 ± 0.02. DUSpeckNet achieved Dice= 0.85 ± 0.01, AUC= 0.92 ± 0.01 at 51.8% fewer parameters with no pretraining. On the ABreast cohort (a small African handheld point-of-care dataset, n=20, all benign), transfer is assessed for segmentation only, as the benign-only composition prevents malignancy-detection evaluation. EfficientNet-B3 and DUSpeckNet achieved segmentation Dice= 0.85 ± 0.00 and 0.82 ± 0.02 respectively. Classification uncertainty tracked error for EfficientNet-B3 ($r = 0.34$, $p = 0.0005$) but not for DUSpeckNet, identifying uncertainty calibration under pretraining-free transfer as a key future direction.

Keywords: Breast ultrasound · Lesion segmentation · Malignancy classification · Uncertainty quantification · Domain shift

1 Introduction

Breast cancer is the most common cancer in women worldwide, accounting for an estimated 2.3 million new cases and 685 000 deaths annually [5,4]. Early detection through population-based screening is linked to enhanced survival and reduced mortality [17,16]. In high-income countries, mammography has achieved a 20% or greater reduction in breast cancer mortality relative to unscreened populations [5,19], yet sensitivity is considerably reduced in women with dense breast tissue due to overlapping radiodensity of tumours and fibroglandular tissue [5,9].

In low- and middle-income countries (LMICs), mammography infrastructure is lacking due to cost, equipment scarcity, and limited trained personnel, resulting in late detection and poorer survival [4,18,6]. Ultrasound (US) is a portable,

inexpensive, radiation-free alternative and is the standard of care in LMICs that enhances lesion detection in dense breasts and resource-limited settings [2,3], though its interpretation remains highly operator-specific with significant inter-observer variability [25]. Speckle noise, acoustic shadowing, and posterior enhancement further complicate visual assessment [2].

Deep learning (DL) has shown great promise for breast lesion detection, classification, and segmentation [5,10], yet clinical translation remains limited. Most methods address segmentation or classification separately, yielding suboptimal spatial decision-making [11], and rely on general augmentation strategies that fail to simulate ultrasound-specific artifacts. Calibrated uncertainty quantification is rarely incorporated, limiting safe clinical deployment [11]. Crucially, existing models are rarely tested on African populations or portable device data, leaving a critical gap in resource-limited settings.

To address these limitations, we present an uncertainty-aware deep learning framework for breast ultrasound analysis. Our contributions are: (1) a dual-task architecture for simultaneous segmentation and classification from a shared encoder; (2) ultrasound-specific augmentation strategies that produce consistent gains across all encoder architectures; (3) a Monte Carlo Dropout uncertainty framework for identifying ambiguous predictions; (4) cross-domain evaluation on the ABreast cohort, a novel African point-of-care breast ultrasound dataset, assessing segmentation transfer on a small, entirely benign sample (n=20) and examining Monte Carlo Dropout uncertainty as a domain-shift signal, which relates to error for the pretrained EfficientNet-B3 but not for the from-scratch DUSpeckNet; and (5) DUSpeckNet, a Dual-task Ultrasound Speckle-aware Network and lightweight instantiation of this framework (6,522,507 parameters) with a speckle-aware stem and temperature-scaled uncertainty calibration, trained entirely from scratch for resource-limited deployment.

2 Related Work

Deep learning has significantly advanced segmentation of anatomical structures and lesions in medical images. U-Net [24] introduced skip connections for preserving spatial resolution, with U-Net++ [27] improving feature propagation via nested pathways, and attention mechanisms further enhancing boundary delineation [22]. In breast ultrasound, early CNN-based methods improved lesion detection over hand-crafted features [26], while later work coupled encoder-decoder segmentation with separate classification pipelines [13]. Most studies remain confined to single datasets without cross-centre validation, and largely ignore ultrasound-specific artifacts such as speckle noise, acoustic shadowing, and gain variation [21]. Beyond architectural advances, uncertainty quantification and multi-task learning have emerged as key directions for robust clinical deployment. Monte Carlo Dropout [7] and Bayesian deep learning [15] enable epistemic uncertainty estimation to flag out-of-distribution cases, while multi-task learning exploits shared representations for joint segmentation and classification [11,12]. Musah et al. [20] demonstrated that MC Dropout uncertainty can signal

domain shift in breast ultrasound segmentation, yet cross-domain generalisation to diverse device types and patient demographics remains an open challenge that our work takes a preliminary step toward.

3 Method

3.1 Model Architecture

We present a dual-task encoder-decoder for simultaneous lesion segmentation and malignancy classification via a shared encoder with task-specific heads. Three ImageNet-pretrained encoders are considered: ResNet50 (∼25M), ResNet34 (∼21M), and EfficientNet-B3 (∼12M). The segmentation branch uses a U-Net-like decoder (transposed convolution, batch normalisation, ReLU, skip connections) with a final 1×1 convolution and sigmoid output. The classification branch applies global average pooling, fully connected layers (512, 256, 128, 2), dropout ($p = 0.3$), and softmax output. Epistemic uncertainty is estimated via Monte Carlo Dropout ($N = 20$). Per-image classification uncertainty is taken as the summed across-pass variance of the class probabilities, and per-image segmentation uncertainty as the mean over pixels of the across-pass variance of the predicted probability.

3.2 DUSpeckNet: Dual-task Ultrasound Speckle-aware Network

DUSpeckNet is a lightweight dual-task architecture for resource-constrained deployment, trained entirely from scratch, comprising: (1) a **Speckle-Aware Stem** (depthwise separable convolution with channel attention) to suppress speckle noise; (2) a **Mobile Inverted Bottleneck Encoder** (6,522,507 parameters) with five stages of SE-enhanced MBConv blocks; (3) a **Spatial Attention Decoder** directing segmentation towards lesion boundaries; and (4) a **Temperature-Scaled Classification Head** for calibrated uncertainty estimation without large-scale pretraining.

3.3 Training and Validation Datasets

We combined three publicly available breast ultrasound datasets: BUSI [1] (646 images, aged 25–75, Egypt; normal cases excluded, multiple annotations merged via pixel-wise logical union), BrEaST [23] (252 images, aged 18–87, Poland; normal cases excluded), and BUS-BRA [8] (1,875 images, 1,064 female patients, aged 16–89, Brazil). The combined dataset comprises 2,773 images verified as benign (1,859) or malignant (914), evaluated using stratified five-fold cross-validation at the image level. Among the three sources, only BUS-BRA contains multiple images per patient (1,875 images from 1,064 patients); a subset of these patients may therefore be represented in both training and validation folds, whereas BrEaST and BUSI are unaffected.

3.4 Test (ABreast) Dataset

The ABreast dataset (African Breast Imaging Dataset for Equitable Cancer Care) is a prospective handheld ultrasound dataset acquired in community settings in Nigeria and Uganda [14,6]. Of 132 enrolled participants, the 20 with radiologist-delineated lesion annotations were included as held-out test data; inclusion was determined by annotation availability rather than image-quality selection. Images were acquired using a Philips Lumify L12-4 device (linear array, 4–12 MHz) [14]. All 20 cases are benign (BI-RADS 1–3), which permits evaluation of segmentation transfer but not of the malignancy classification branch in this cohort.

3.5 Data Preprocessing and Augmentation

Images were resized to 256×256 pixels via bilinear interpolation and normalised using ImageNet channel statistics; grayscale images were replicated to three channels. Masks were binarised at threshold 127; 17 BUSI images with multiple annotations were merged via pixel-wise logical union. Standard augmentations included horizontal flip ($p = 0.5$), rotation ($\pm 15°$), scaling (0.9–1.1), and brightness/contrast adjustment. Three ultrasound-specific augmentations ($p = 0.5$) were applied: (1) speckle noise ($\alpha \in [0.05, 0.2]$), (2) acoustic shadowing ($\beta \in [0.3, 0.6]$), and (3) gain variation as vertical intensity gradients. Test-Time Augmentation (TTA) averaged predictions over 7 fixed transforms per image: horizontal flip; rotations of 90°, 180°, and 270°; scaling of -10% and $+10\%$; and the untransformed original.

4 Experimental Setup

4.1 Experimental Design

All models were trained with a weighted multi-task loss combining segmentation (Dice and binary cross-entropy equally) and classification (categorical cross-entropy):

$$\mathcal{L} = 0.6 \cdot \mathcal{L}_{\text{seg}} + 0.4 \cdot \mathcal{L}_{\text{cls}}. \tag{1}$$

All experiments were conducted on a Kaggle-hosted NVIDIA Tesla P100 GPU (16 GB VRAM, CUDA 12.8, PyTorch 2.11), using AdamW (learning rate 10^{-4}, weight decay 10^{-5}), cosine annealing (minimum 10^{-6}), batch size 16, gradient clipping 1.0, and early stopping (patience 15) over 100 epochs. ResNet and EfficientNet encoders were ImageNet-pretrained; DUSpeckNet was trained from scratch. Segmentation was assessed via Dice and IoU; classification via accuracy, precision, recall, F1, and AUC-ROC. Classification uncertainty (summed across-pass variance of the class probabilities) was correlated with classification error via Spearman rank correlation. TTA used 7 variants per image with averaged predictions. Per-epoch training time ranged from 34 s (ResNet34) to 91 s (EfficientNet-B3 with TTA); DUSpeckNet ran at 51–52 s per epoch (6.5 M parameters, no pretraining).

4.2 ABreast Cohort Testing Protocol

Three experiments evaluated cross-domain segmentation transfer and uncertainty behaviour under domain shift on the ABreast cohort. **Experiment A: Zero-Shot Transfer.** Each trained model was applied directly to all 20 ABreast cases without retraining. **Experiment B: Uncertainty under Domain Shift.** Uncertainty distributions from the CV folds and the ABreast cohort were compared using Mann-Whitney U tests to assess whether uncertainty increases out of distribution; separately, Spearman rank correlation quantified the relationship between classification uncertainty and classification error within the ABreast cohort. **Experiment C: Decoder Fine-Tuning.** Leave-one-out cross-validation was applied across the 20 ABreast cases, fine-tuning the decoder and classification head for 30 epochs (learning rate 10^{-5}) with encoder weights frozen.

5 Results

5.1 Overall Performance

Nine configurations were evaluated under 5-fold cross-validation (Table 1). ResNet50 + US-Aug achieved the highest Dice (0.8814 ± 0.0058) and IoU (0.7879 ± 0.0093); EffNet-B3 + US-Aug achieved the highest AUC (0.9432 ± 0.0151), Accuracy (0.8936 ± 0.0178), and F1 (0.8365 ± 0.0272). Domain-specific augmentation consistently outperformed baseline and TTA configurations across both encoders. DUSpeckNet (6,522,507 parameters, no pretraining) achieved Dice$= 0.8539 \pm 0.0073$, AUC$= 0.9155 \pm 0.0135$, remaining competitive with all ResNet baselines at 51.8% fewer parameters.

Table 1. Overall performance (mean ± std, 5-fold CV). US-Aug: ultrasound-specific augmentation; TTA: test-time augmentation; DUSpeckNet trained from scratch. Bold: highest per column.

Configuration	Dice	IoU	Acc.	Recall	F1	AUC
ResNet50	0.8675 ± 0.0044	0.7660 ± 0.0069	0.8608 ± 0.0120	0.7112 ± 0.0804	0.7692 ± 0.0307	0.9178 ± 0.0108
ResNet50 + TTA	0.7660 ± 0.0219	0.6213 ± 0.0282	0.8392 ± 0.0496	0.6402 ± 0.2205	0.6972 ± 0.1667	0.8893 ± 0.0286
ResNet50 + US-Aug	$\mathbf{0.8814 \pm 0.0058}$	$\mathbf{0.7879 \pm 0.0093}$	0.8860 ± 0.0121	0.7911 ± 0.0236	0.8207 ± 0.0183	0.9396 ± 0.0127
ResNet34	0.8682 ± 0.0050	0.7660 ± 0.0100	0.8745 ± 0.0183	0.7899 ± 0.0253	0.8042 ± 0.0274	0.9267 ± 0.0188
EffNet-B3	0.8656 ± 0.0025	0.7630 ± 0.0039	0.8767 ± 0.0237	0.7845 ± 0.0335	0.8080 ± 0.0320	0.9310 ± 0.0192
EffNet-B3 + TTA	0.7777 ± 0.0133	0.6365 ± 0.0177	0.8525 ± 0.0257	0.7954 ± 0.0570	0.7810 ± 0.0290	0.8977 ± 0.0177
EffNet-B3 + US-Aug	0.8774 ± 0.0045	0.7815 ± 0.0071	$\mathbf{0.8936 \pm 0.0178}$	0.8260 ± 0.0359	$\mathbf{0.8365 \pm 0.0272}$	$\mathbf{0.9432 \pm 0.0151}$
EffNet-B3 + US-Aug + TTA	0.8031 ± 0.0057	0.6710 ± 0.0079	0.8702 ± 0.0316	$\mathbf{0.8577 \pm 0.0228}$	0.8145 ± 0.0379	0.9288 ± 0.0151
DUSpeckNet (Ours)	0.8539 ± 0.0073	0.7451 ± 0.0111	0.8615 ± 0.0094	0.7538 ± 0.0117	0.7822 ± 0.0124	0.9155 ± 0.0135

5.2 Segmentation and Classification Performance

ResNet50 + US-Aug achieved the highest Dice (0.8814) while EfficientNet-B3 + US-Aug showed the lowest variance (±0.0045), indicating more stable generalisation (Table 2). TTA consistently degraded performance across both encoders. This degradation is consistent with a mismatch between the test-time transforms and ultrasound acquisition geometry: four of the seven TTA

variants apply 90°, 180°, or 270° rotations, which reorient the image away from the fixed transducer-to-depth axis of B-mode ultrasound, where echo intensity, acoustic shadowing, and posterior enhancement all vary systematically with depth. Averaging predictions over these anatomically implausible orientations lowers segmentation quality, whereas the ultrasound-specific training augmentations, which preserve this geometry, improve it. EfficientNet-B3 + US-Aug achieved the highest AUC (0.9432), marginally above ResNet50 + US-Aug (0.9396); domain-specific augmentation outperformed architectural depth across all encoders. High recall variance for ResNet50 + TTA (±0.2205) reflects TTA pipeline instability.

Table 2. Per-fold Dice across configurations (5-fold CV). Bold: highest per fold. Per-fold AUC and recall follow the same trends; see Table 1 for their means.

Configuration	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**
ResNet50 (Baseline)	0.8714	0.8610	0.8695	0.8721	0.8635
ResNet50 + TTA	0.7879	0.7762	0.7241	0.7690	0.7728
ResNet50 + US-Aug	**0.8887**	**0.8709**	**0.8808**	**0.8836**	**0.8827**
ResNet34 (Baseline)	0.8735	0.8621	0.8647	0.8746	0.8662
EfficientNet-B3 (Baseline)	0.8693	0.8637	0.8652	0.8675	0.8622
EfficientNet-B3 + TTA	0.7911	0.7550	0.7792	0.7905	0.7729
EfficientNet-B3 + US-Aug	0.8832	0.8701	0.8752	0.8797	0.8786
EfficientNet-B3 + US-Aug + TTA	0.8031	0.8044	0.7940	0.8020	0.8118
DUSpeckNet (Ours)	0.8634	0.8468	0.8575	0.8578	0.8441

5.3 Testing on the ABreast Cohort

Zero-Shot Transfer. EfficientNet-B3 + US-Aug achieved Dice= 0.8485±0.0049, IoU= 0.7369 ± 0.0074; DUSpeckNet achieved Dice= 0.8190 ± 0.0242, IoU= 0.6942 ± 0.0344 on the ABreast cohort.

Uncertainty under Domain Shift. For EfficientNet-B3 + US-Aug, predictive uncertainty was higher on the ABreast cohort (0.0065 ± 0.0190) than on the CV folds (0.0042±0.0157), a shift confirmed by a Mann-Whitney U test and consistent with epistemic uncertainty rising out of distribution. Within the ABreast cohort, classification uncertainty correlated with classification error (Spearman $r = 0.341$, $p = 0.0005$); as all cases are benign, this error reflects false positives on benign tissue rather than missed malignancies. DUSpeckNet showed neither effect: its uncertainty did not rise on the cohort (0.0009 ± 0.0057 vs 0.0031 ± 0.0105) and its uncertainty-error correlation was not significant ($r = 0.148$, $p = 0.1418$), indicating that its from-scratch uncertainty estimates are not a reliable error signal under domain shift. Classification performance metrics are not reported, as the benign-only composition (BI-RADS 1–3) precludes evaluating malignancy detection.

Decoder Fine-Tuning. Fine-tuning achieved Dice= 0.8556 ± 0.1016, IoU= 0.7606 ± 0.1467 (EfficientNet-B3 + US-Aug) and Dice= 0.7698 ± 0.2645, IoU= 0.6805 ± 0.2597 (DUSpeckNet); Table 3 summarises results.

Table 3. Results on the ABreast cohort. Zero-Shot: direct inference. Fine-Tune: leave-one-out decoder adaptation with frozen encoder. r: Spearman correlation between classification uncertainty and classification error (false positives, as all cases are benign).

Model	Zero-Shot Dice	Zero-Shot IoU	Fine-Tune Dice	Fine-Tune IoU	UQ Spearman r
EfficientNet-B3 + US-Aug	0.8485 ± 0.0049	0.7369 ± 0.0074	0.8556 ± 0.1016	0.7606 ± 0.1467	$r = 0.3410$ ($p = 0.0005$)
DUSpeckNet (Ours)	0.8190 ± 0.0242	0.6942 ± 0.0344	0.7698 ± 0.2645	0.6805 ± 0.2597	$r = 0.1480$ ($p = 0.1418$)

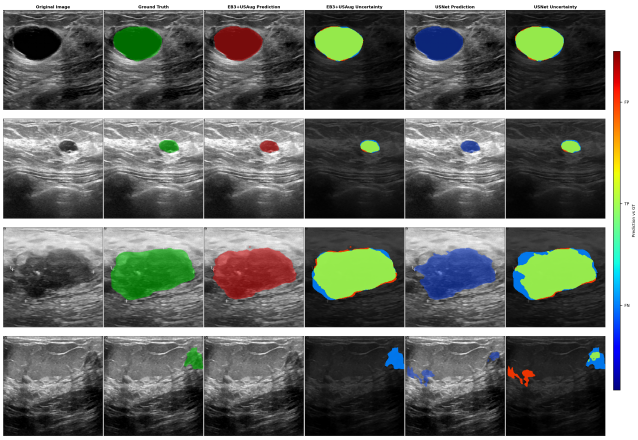


Fig. 1. Qualitative segmentation on the combined dataset. Rows: benign easy, benign hard, malignant easy, malignant hard (ranked by Dice). Columns: (1) image, (2) ground truth (green), (3) EffNet-B3 + US-Aug (red), (4) error map, (5) DUSpeckNet (blue), (6) error map. TP (green), FP (red), FN (cyan).

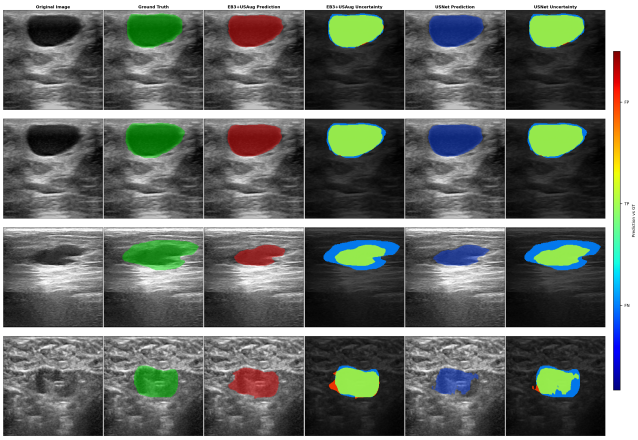


Fig. 2. Qualitative segmentation on the ABreast cohort (Philips Lumify L12-4, zero-shot). Rows: best, second best, most challenging, highest uncertainty. Same column layout as Fig. 1.

6 Impact in Resource-Constrained Settings

Breast cancer incidence in sub-Saharan Africa is projected to double by 2040, with Black African women facing the highest mortality globally, driven by late diagnosis and limited diagnostic access [6]. Handheld point-of-care ultrasound is a low-cost alternative to mammography but needs tools that generalise without site-specific retraining. As a preliminary step, we evaluate segmentation transfer on the ABreast cohort [14], a community-acquired breast imaging dataset from Nigeria and Uganda, on which both models transferred without retraining. For the pretrained EfficientNet-B3, uncertainty rose under domain shift and tracked error, suggesting a signal for prioritising radiologist review; this signal was unreliable for the from-scratch DUSpeckNet, indicating that lightweight, pretraining-free models need further calibration before their uncertainty can be trusted in the field.

7 Conclusion

We proposed an uncertainty-aware dual-task framework for joint breast ultrasound lesion segmentation and malignancy classification. Domain-specific augmentation consistently outperformed generic strategies across all encoders: ResNet50 + US-Aug achieved the highest Dice (0.8814) and IoU (0.7879), and EfficientNet-B3 + US-Aug the best AUC (0.9432), Accuracy (0.8936), and F1 (0.8365). On the ABreast cohort, both models transferred with competitive segmentation (EfficientNet-B3 Dice= 0.8485; DUSpeckNet Dice= 0.8190), and classification uncertainty tracked error for the pretrained model ($r = 0.341$, $p = 0.0005$) but not for DUSpeckNet ($r = 0.148$, $p = 0.1418$). DUSpeckNet (6.5M parameters, no pretraining) thus matches heavier baselines on segmentation while its uncertainty remains unreliable under domain shift, identifying calibration for pretraining-free transfer as a key future direction.

Limitations and Future Work

Cross-domain evaluation is limited to the ABreast cohort's 20 annotated benign cases, permitting assessment of segmentation transfer but not malignancy detection; larger prospective cohorts with malignant cases are needed. Cross-validation splits were image-level, so some BUS-BRA patients with multiple images may appear in both training and validation folds. We report DUSpeckNet as a single configuration and do not ablate its individual components, including the Speckle-Aware Stem; a component-level analysis is left to future work. DUSpeckNet's uncertainty calibration also degrades under domain shift and may benefit from deep ensembles or self-supervised pretraining. Future work includes multi-centre testing, conformal prediction, DICOM integration, and extension to community settings.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in Brief **28**, 104863 (2019). https://doi.org/10.1016/j.dib.2019.104863
2. Baad, M., Lu, Z.F., Reiser, I., Paushter, D.: Clinical significance of US artifacts. Radiographics **37**(5), 1408–1423 (2017). https://doi.org/10.1148/rg.2017160175
3. Bunnell, A., Valdez, D., Strand, F., Glaser, Y., Sadowski, P., Shepherd, J.A.: Artificial intelligence-informed handheld breast ultrasound for screening: A systematic review of diagnostic test accuracy. ArXiv (2024), https://arxiv.org/abs/2411.07322
4. Dan, Q., Xu, Z., Burrows, H., et al.: Diagnostic performance of deep learning in ultrasound diagnosis of breast cancer: a systematic review. npj Precision Oncology **8**, 21 (2024). https://doi.org/10.1038/s41698-024-00514-z
5. Dan, Q., Zheng, T., Liu, L., Sun, D., Chen, Y.: Ultrasound for breast cancer screening in resource-limited settings: Current practice and future directions. Cancers (Basel) **15**(7), 2112 (2023). https://doi.org/10.3390/cancers15072112
6. Emegoakor, A., Raymond, C., Gbadamosi, Y., Malumba, R., Umoren, C., Betancourt, D.S., Iorumbur, A.M., Kalaiwo, C., Muhammad, A.R., Musinguzi, D., Atukunda, P.: Bridging the gap: A community driven and AI-enabled approach to early breast cancer detection in black african women. In: Artificial Intelligence and Imaging for Diagnostic and Treatment Challenges in Breast Care. Deep-Breath 2025. Lecture Notes in Computer Science, vol. 16142, pp. 289–299. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-05559-0_29
7. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: Appendix. ArXiv (2015), https://arxiv.org/abs/1506.02157
8. Gómez-Flores, W., Gregorio-Calas, M.J., Coelho de Albuquerque Pereira, W.: BUS-BRA: A breast ultrasound dataset for assessing computer-aided diagnosis systems. Medical Physics **51**(4), 3110–3123 (2024). https://doi.org/10.1002/mp.16812
9. Gordon, P.B.: The impact of dense breasts on the stage of breast cancer at diagnosis: A review and options for supplemental screening. Current Oncology **29**(5), 3595 (2022). https://doi.org/10.3390/curroncol29050291
10. Hasan, A.H., Khalid, U.A.R., Abid, M.A.: Leveraging artificial intelligence in breast cancer screening and diagnosis. Cureus **17**(2), e79177 (2025). https://doi.org/10.7759/cureus.79177
11. He, Q., Yang, Q., Su, H., Wang, Y.: Multi-task learning for segmentation and classification of breast tumors from ultrasound images. Computers in Biology and Medicine **173**, 108319 (2024). https://doi.org/10.1016/j.compbiomed.2024.108319
12. Jiménez-Gaona, Y., Rodríguez-Álvarez, M.J., Lakshminarayanan, V.: Deep learning based computer-aided systems for breast cancer imaging: A critical review. ArXiv (2020), https://arxiv.org/abs/2010.00961
13. Kalafi, E.Y., Jodeiri, A., Setarehdan, S.K., Lin, N.W., Rahmat, K., Taib, N.A., Ganggayah, M.D., Dhillon, S.K.: Classification of breast cancer lesions in ultrasound images by using attention layer and loss ensemble in deep convolutional neural networks. Diagnostics (Basel) **11**(10), 1859 (2021). https://doi.org/10.3390/diagnostics11101859
14. Kawooya, M.G., Malumba, R., Jatho, A., Katumba, A., Nakatumba-Nabende, J., Musinguzi, D., Achuka, S.A., Adewole, M., Anazodo, U.C.: The african breast imaging dataset for equitable cancer care: Protocol for an open mammogram and

ultrasound breast cancer detection dataset. medRxiv (2025). https://doi.org/10.1101/2025.08.26.25334483
15. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? ArXiv (2017), https://arxiv.org/abs/1703.04977
16. Loud, J.T., Murphy, J.: Cancer screening and early detection in the 21st century. Seminars in Oncology Nursing **33**(2), 121–128 (2017). https://doi.org/10.1016/j.soncn.2017.02.002
17. Ma, Z., He, W., Zhang, Y., et al.: First mammography screening participation and breast cancer incidence and mortality in the subsequent 25 years: population based cohort study. BMJ **390**, e085029 (2025). https://doi.org/10.1136/bmj-2025-085029
18. Manson, E.N., Achel, D.G.: Fighting breast cancer in low-and-middle-income countries – what must we do to get every woman screened on regular basis? Scientific African (2023), https://api.semanticscholar.org/CorpusID:260693911
19. Marcon, M., Fuchsjäger, M.H., Clauser, P., Mann, R.M.: ESR essentials: screening for breast cancer – general recommendations by EUSOBI. European Radiology **34**(10), 6348–6357 (2024). https://doi.org/10.1007/s00330-024-10740-5
20. Musah, T., Kalaiwo, C., Akram, M., Abdulai, U.N., Adewole, M., Dako, F., Emegoakor, A.C., Anazodo, U.C., Adjei, P.E., Raymond, C.: Towards trustworthy breast tumor segmentation in ultrasound using monte carlo dropout and deep ensembles for epistemic uncertainty estimation. ArXiv (2025), https://arxiv.org/abs/2508.17768
21. Noble, J.A., Boukerroui, D.: Ultrasound image segmentation: a survey. IEEE Transactions on Medical Imaging **25**(8), 987–1010 (2006). https://doi.org/10.1109/tmi.2006.877092
22. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D.: Attention U-Net: Learning where to look for the pancreas. ArXiv (2018), https://arxiv.org/abs/1804.03999
23. Pawłowska, A., Ćwierz-Pieńkowska, A., Domalik, A., et al.: Curated benchmark dataset for ultrasound based breast lesion analysis. Scientific Data **11**(1), 148 (2024). https://doi.org/10.1038/s41597-024-02984-z
24. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. ArXiv (2015), https://arxiv.org/abs/1505.04597
25. Webb, J.M., Adusei, S.A., Wang, Y., Samreen, N., Adler, K., Meixner, D.D., Fazzio, R.T., Fatemi, M., Alizad, A.: Comparing deep learning-based automatic segmentation of breast masses to expert interobserver variability in ultrasound imaging. Computers in Biology and Medicine **139**, 104966 (2021). https://doi.org/10.1016/j.compbiomed.2021.104966
26. Yap, M.H., Pons, G., Marti, J., et al.: Automated breast ultrasound lesions detection using convolutional neural networks. IEEE Journal of Biomedical and Health Informatics **22**(4), 1218–1226 (2018). https://doi.org/10.1109/JBHI.2017.2731873
27. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: A nested U-Net architecture for medical image segmentation. Lecture Notes in Computer Science **11045**, 3–11 (2018). https://doi.org/10.1007/978-3-030-00889-5_1