Beyond Dice: Evaluating Reliability and Robustness in Breast Ultrasound Lesion Segmentation Under Progressive Speckle Corruption

Ummulkhair Osman-Watara[1][0009−0007−4721−9113], Robert Owusu Afiakwa[2,3][0009−0006−3132−8633], Frederick Wiafe Ababio[2], Nafisat Abdulsalam[4], Benjamin Appiah[2,5][0009−0002−4925−7547], John Takal[2], Felix Osei Bonsu[2], Toufiq Musah[2], Aondona Moses Iorumbur[6], Raymond Confidence[7], Udunna Anazodo[8], and Livingstone Eli Ayivor[2,3][0009−0008−3032−204X]

[1] Spectra Health Imaging and Interventional Radiology Center, Ghana
[2] Department of Computer Engineering, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
[3] Global Health and Infectious Disease Group, Kumasi Centre for Collaborative Research in Tropical Medicine, Kumasi, Ghana
[4] Department of Mechatronics Engineering, Federal University of Technology, Minna, Nigeria
[5] Biomedical Technologies Lab, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
[6] Department of Physics, Federal University of Technology, Minna, Nigeria
[7] Department of Biomedical Engineering, McGill University, Montreal, Canada
[8] Montreal Neurological Institute, McGill University, Montreal, Canada
leayivor@st.knust.edu.gh

Abstract. Breast ultrasound lesion segmentation is challenging due to speckle noise, acoustic artifacts, and high variability in lesion appearance. Although deep learning methods achieve strong performance on benchmark datasets, evaluation is often dominated by overlap-based metrics that may not fully reflect clinically relevant segmentation quality. In this work, we study the effect of training strategy on segmentation performance and reliability under progressive speckle corruption. We compare standard supervised training with fixed augmentation against a curriculum that gradually introduces spatially correlated multiplicative speckle noise during optimization. Experiments are conducted on the PRECISE/PACE breast ultrasound dataset using U-Net, Attention U-Net, TransUNet, and SegFormer-B0. Evaluation includes overlap-based metrics, boundary-sensitive measures, lesion-level analysis, uncertainty estimation with Monte Carlo dropout, and calibration assessment. Progressive speckle corruption improves Dice across architectures, with the best performance achieved by Attention U-Net (0.8099). However, it reduces IoU and Surface Dice and generally increases HD95. It also leads to higher failure rates and over-segmentation, particularly in malignant and normal cases. In contrast, standard training yields better boundary accuracy, improved calibration, and more reliable uncertainty estimates.

The findings highlight the need for multi-metric evaluation beyond overlap measures when developing segmentation models for artifact-prone ultrasound imaging.

Keywords: Breast Ultrasound · Lesion Segmentation; · Deep Learning · Curriculum Learning · Predictive Uncertainty.

1 Introduction

Breast ultrasonography is a widely used imaging modality for lesion detection and characterization due to its accessibility, portability, and lack of ionizing radiation [4,19,16,10]. In clinical workflows, it supports not only detection but also lesion characterization, treatment monitoring, and image-guided interventions

Despite its clinical utility, automated analysis of breast ultrasound remains challenging. Image formation is affected by speckle noise, acoustic shadowing, attenuation artifacts, and operator-dependent variability, often resulting in low contrast and poorly defined lesion boundaries [9,21,12,24]. These factors make accurate lesion delineation difficult, and can directly impact clinically relevant measurements derived from segmentation outputs [12,24,20].

Deep learning methods, including convolutional, attention, and transformer-based architectures, have significantly advanced medical image segmentation [18,15,3,22,2,8]. In breast ultrasound, recent approaches have improved lesion localization through enhanced feature modelling and hybrid representations [2,17]. However, evaluation remains dominated by overlap-based metrics such as Dice and IoU, which do not fully capture boundary accuracy, robustness under degradation, or predictive reliability [11,23]. Consequently, improvements in benchmark scores may not reflect clinically meaningful performance gains.

This limitation is particularly critical in ultrasound imaging, where lesion appearance is highly sensitive to speckle and other acquisition artifacts. While standard training relies on fixed or randomly sampled augmentations, it does not control exposure to progressively increasing degradation. Curriculum learning offers an alternative by gradually increasing task difficulty during training, improving robustness under distribution shift [1,13]. However, its application to progressive speckle corruption in breast ultrasound segmentation remains underexplored, particularly in relation to model reliability and uncertainty.

Reliable deployment of segmentation systems also requires uncertainty-aware predictions. Monte Carlo dropout enables approximate Bayesian inference by performing stochastic forward passes at test time [6,14], while calibration metrics assess the alignment between predicted confidence and actual performance [7]. These measures are essential in ultrasound imaging, where ambiguity arises from weak boundaries and low-contrast regions.

In this work, we evaluate how training strategy under progressive speckle corruption influences both segmentation performance and reliability in breast ultrasound lesion segmentation. We compare standard augmentation-based training with a curriculum that progressively introduces ultrasound-specific speckle degradation during optimization. Evaluation is designed to go beyond Dice by incorporating boundary-sensitive metrics, lesion-level performance analysis, robustness under artifact conditions, and reliability assessment using uncertainty estimation and calibration measures.

2 Methods

2.1 Dataset and Preprocessing

Experiments were conducted on the PRECISE/PACE Breast Ultrasound Challenge dataset, consisting of 3,116 images with corresponding expert-annotated lesion masks [5]. The dataset was split into 70% training, 15% validation, and 15% testing using a fixed partition to ensure fair comparison across methods. All images were resized to 256×256 pixels and normalized. Contrast enhancement was applied using Contrast Limited Adaptive Histogram Equalization (CLAHE). During training, data augmentation included random flips, rotations and intensity variations. Identical geometric transformations were applied to the corresponding segmentation masks to preserve spatial alignment.

2.2 Training Paradigms

Standard Training The first paradigm follows conventional supervised training, where images are sampled from a fixed augmentation distribution throughout optimization. This serves as the baseline training strategy against which progressive corruption is evaluated.

Progressive Speckle Corruption The second paradigm progressively introduces ultrasound-specific image degradation during training. Rather than exposing the network to the full range of corruption levels from the outset, training begins with uncorrupted images and gradually increases corruption intensity as optimization progresses.

Speckle corruption was modelled as a spatially correlated multiplicative process, $I_c = I(1 + \sigma S)$,, where I denotes the original image, I_c represents the corrupted image, S is a spatially correlated noise field generated by Gaussian smoothing of white noise, and σ controls the intensity of corruption. The intensity of corruption was progressively increased according to a predefined schedule, ranging from noise-free images ($\sigma = 0$) to heavily corrupted images ($\sigma = 0.12$). Lesion masks remained unchanged throughout training, ensuring that only image appearance was modified while anatomical structures were preserved.

2.3 Implementation Details

Segmentation Architectures To assess whether the effects of progressive corruption generalize across different segmentation paradigms, experiments were conducted using four representative architectures: U-Net [18], Attention U-Net [15], TransUNet [3], and SegFormer-B0 [22]. These architectures collectively represent convolutional, attention-enhanced, hybrid CNN-transformer, and transformer-based segmentation approaches. All models were configured for binary lesion segmentation and trained using identical data partitions and optimization settings.

Optimization Details All models were optimized using AdamW with a combined Dice and Binary Cross-Entropy objective. Learning rates were updated using cosine annealing to promote stable convergence. A background suppression strategy was incorporated to reduce the influence of spurious activations in non-lesion regions. Apart from the corruption schedule used in the progressive training paradigm, all optimization settings were kept identical across experiments to ensure that observed differences could be attributed solely to the training strategy.

2.4 Evaluation Framework

Segmentation performance was evaluated using both overlap-based and boundary-sensitive metrics. Region overlap was quantified using the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU). Boundary quality was assessed using the 95th percentile Hausdorff Distance (HD95), Boundary F1 Score, and Surface Dice. To evaluate robustness beyond conventional segmentation accuracy, additional analyses were performed across lesion types and imaging conditions. Performance was assessed separately for benign and malignant lesions, and artifact-specific evaluations were conducted under varying levels of speckle corruption and acoustic shadowing.

2.5 Uncertainty and Calibration Analysis

Predictive uncertainty was estimated using Monte Carlo Dropout. During inference, dropout layers were kept active and T stochastic forward passes were performed for each input image, yielding a set of probabilistic predictions $p_t{}_{t=1}^{T}$. The predictive mean was computed as:

$$\bar{p} = \frac{1}{T} \sum_{t=1}^{T} p_t,$$

where p_t denotes the predicted probability map from the t-th stochastic forward pass.

Epistemic uncertainty was quantified using the predictive variance:

$$U_{epi} = \frac{1}{T} \sum_{t=1}^{T} (p_t - \bar{p})^2,$$

which captures uncertainty arising from model parameter variability.

Table 1. Overall segmentation performance under standard and progressive training.

Model	Training	Dice ↑	IoU ↑	HD95 ↓	Surface Dice ↑
U-Net	Standard	0.7552	0.7322	26.75	0.6551
U-Net	Progressive	0.8064	0.6825	26.37	0.6210
Attention U-Net	Standard	0.7546	0.7340	**25.62**	0.6586
Attention U-Net	Progressive	**0.8099**	0.6883	26.47	0.6336
TransUNet	Standard	0.7630	**0.7401**	**18.92**	**0.6708**
TransUNet	Progressive	0.7939	0.6686	24.19	0.6039
SegFormer-B0	Standard	0.6705	0.5935	30.84	0.4995
SegFormer-B0	Progressive	0.7214	0.5756	32.07	0.4522

In addition, predictive entropy was computed to quantify overall uncertainty in segmentation outputs:

$$H(p) = -\sum_{c} p_c \log(p_c),$$

where p_c denotes the predicted probability for class c. Mutual information was used to isolate epistemic uncertainty by measuring the divergence between predictive entropy and expected entropy across stochastic forward passes. Calibration performance was evaluated using Expected Calibration Error (ECE), which quantifies the discrepancy between predicted confidence and empirical accuracy across confidence bins.

3 Results

3.1 Overall Segmentation Performance

Table 1 summarizes performance under standard and progressive training across all architectures. Progressive speckle corruption improved Dice scores for all models, with gains of 0.0512 for U-Net, 0.0553 for Attention U-Net, 0.0309 for TransUNet, and 0.0509 for SegFormer-B0. The highest Dice score was achieved by Attention U-Net under the progressive training.

However, improvements in Dice did not consistently translate to improvements in boundary-sensitive metrics. IoU and Surface Dice decreased across most architectures under progressive training. Similarly, HD95 exhibited mixed behaviour: while a slight improvement was observed for U-Net, increased boundary errors were recorded for Attention U-Net, TransUNet, and SegFormer-B0. **Fig. 1.** presents qualitative segmentation results for representative test cases under both training paradigms.

3.2 Lesion-Specific Performance

Across both training paradigms, performance was consistently higher on benign lesions compared to malignant lesions. Under standard training, Dice scores were

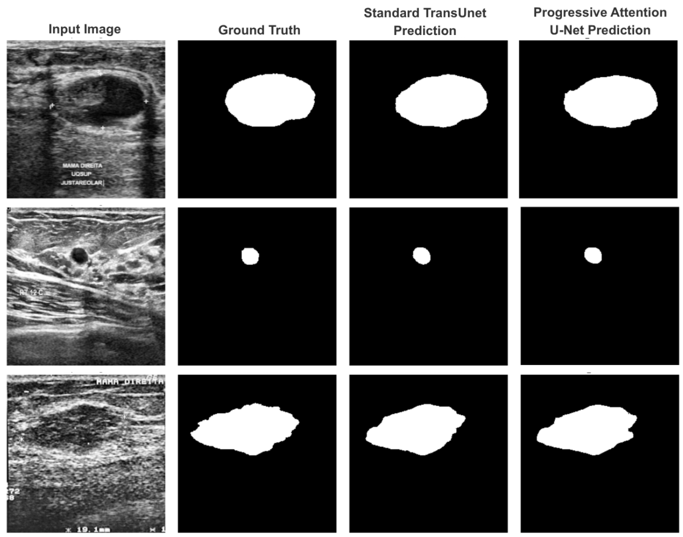

Fig. 1. Qualitative predictions for test cases using best model under each paradigm.

Table 2. Lesion-specific segmentation performance (core metrics).

Training	Lesion Type	Dice ↑	Failure Rate ↓	Surface Dice ↑
Standard	Benign	**0.8433**	**12.07%**	**0.7856**
Progressive	Benign	0.8047	16.90%	0.7351
Standard	Malignant	**0.7812**	**18.06%**	**0.5978**
Progressive	Malignant	0.7684	25.69%	0.5670

0.8433 for benign lesions and 0.7812 for malignant lesions. Progressive training preserved this trend but resulted in a moderate reduction in lesion-level accuracy.

Standard training achieved lower failure rates and improved boundary accuracy for both lesion categories. HD95 increased substantially under progressive training, rising from 14.48 to 28.98 for benign lesions and from 27.85 to 33.18 for malignant lesions.

Under-segmentation remained the dominant failure mode across all settings, especially for malignant lesions, which exhibited more irregular and heterogeneous boundaries compared to benign cases.

Normal cases were excluded from lesion-specific evaluations due to the absence of lesion masks. Instead, false-positive behaviour was analysed separately, where over-segmentation increased from 32.35% under standard training to 50.00% under progressive training, indicating a higher tendency toward spurious lesion predictions under strong corruption exposure.

Table 3. Lesion-specific error characteristics under standard and progressive training.

Training	Lesion Type	Under-seg. ↓	Over-seg. ↓	HD95 ↓
Standard	Benign	**8.62%**	**3.45%**	**14.48**
Progressive	Benign	8.62%	7.24%	28.98
Standard	Malignant	**13.19%**	**4.17%**	**27.85**
Progressive	Malignant	18.75%	6.25%	33.18

Table 4. Segmentation performance under different artifact conditions

Training	Artifact Category	Mean Dice ↑	Δ vs. No Artifact
Standard	No artifact	**0.8181**	—
Standard	Speckle	0.7621	**-0.0560**
Standard	Shadowing	0.6357	-0.1824
Progressive	No artifact	0.7938	—
Progressive	Speckle	0.7342	**-0.0596**
Progressive	Shadowing	0.5682	-0.2256

3.3 Robustness under Ultrasound-Specific Artifacts

Table 4 shows the effect of ultrasound artifacts on segmentation performance. Both training paradigms exhibited performance degradation under artifact-contaminated images. Under standard training, Dice score decreased from 0.8181 on artifact-free images to 0.7621 under speckle corruption and further to 0.6357 under acoustic shadowing. A similar trend was observed for progressive training, where Dice de-creased from 0.7938 (no artifact) to 0.7342 (speckle) and 0.5682 (shadowing).

When comparing relative performance drops, progressive training showed slightly better robustness to speckle corruption, with a smaller reduction in Dice compared to standard training. However, both paradigms exhibited substantial degradation under acoustic shadowing, with larger performance drops observed in the progressive setting.

3.4 Uncertainty and Reliability Analysis

To evaluate reliability beyond segmentation accuracy, uncertainty and calibration analyses were performed on the best model from each training paradigm.

Progressive training exhibited higher uncertainty compared to standard training, as reflected by increased entropy and reduced mutual information. This suggests greater variability in model predictions under stochastic inference.

Calibration analysis further revealed a degradation in confidence alignment under progressive training. The Expected Calibration Error (ECE) increased relative to standard training, indicating a larger mismatch between predicted confidence and observed segmentation accuracy.

Table 5. Uncertainty and reliability metrics for best-performing models.

Training	Best Model	Entropy ↓	MI ↓	ECE ↓	Pixel Acc. ↑
Standard	TransUNet	**0.0231**	0.0011	**0.0185**	**0.9741**
Progressive	Attention U-Net	0.0279	**0.0005**	0.0227	0.9690

Overall, the standard TransUNet demonstrated more stable and better-calibrated predictions, whereas the progressively trained model showed improved robustness in Dice score but reduced reliability in confidence estimation.

4 Discussion and Conclusion

The results demonstrate that training strategy significantly influences segmentation behaviour in breast ultrasound lesion segmentation. Across all evaluated architectures, progressive speckle corruption consistently improved Dice scores, indicating enhanced region-level overlap with ground truth annotations. However, these gains were not consistently reflected in boundary-sensitive or reliability-oriented evaluation metrics. In particular, progressive training led to reduced Surface Dice and increased HD95 in most architectures, suggesting reduced boundary precision despite improved volumetric agreement. Lesion-specific analysis further revealed higher failure rates and increased over-segmentation under progressive training compared with standard training. These findings indicate that improvements in Dice do not necessarily translate into improved geometric accuracy, particularly in regions where lesion boundaries are poorly defined.

Malignant lesions remained consistently more challenging to segment than benign lesions under both training paradigms, reflecting their greater heterogeneity and less distinct boundary characteristics. Although progressive training slightly reduced performance disparities in some metrics, it did not improve absolute segmentation performance for malignant lesions.

Beyond segmentation accuracy, artifact and reliability analyses highlight additional limitations of progressive speckle corruption. While this strategy improved robustness to speckle-like intensity variations, it did not consistently generalize to more severe structural artifacts such as acoustic shadowing. Moreover, progressively trained models exhibited higher false-positive rates in normal cases and less favourable calibration behaviour, indicating reduced reliability despite improved Dice performance. In contrast, standard training produced more stable boundary predictions and better uncertainty estimates, as reflected by lower ECE and more consistent predictive behaviour. This suggests that while progressive corruption improves robustness to stochastic texture variations, it does not consistently improve reliability under structurally complex imaging conditions.

Together, these results highlight the limitations of relying solely on overlap-based metrics to evaluate segmentation performance in ultrasound imaging. More broadly, they suggest that curriculum-based speckle augmentation alone is insufficient to capture the full spectrum of ultrasound degradation. Future work should explore multi-type artifact modelling and adaptive corruption strategies that better reflect the diversity of clinical ultrasound conditions.

Limitations Experiments were conducted on a single dataset, and used fixed progressive corruption schedule. Uncertainty estimation was also limited to Monte Carlo dropout. Future work should include multi-center validation, broader artifact modelling, and additional uncertainty quantification strategies.

Acknowledgments. This work was part of the Sprint AI Training for African Medical Imaging Knowledge Translation (SPARK) Academy 2026 summer school on deep learning in medical imaging. The authors would like to thank the instructors of the summer for providing insightful background knowledge on brain tumours that informed the research presented here, most notably: Maruf Adewole, Mohannad Barakat, Craig Jones, Noha Magdy, Amal Saleh, Aondona Iorumbur, Bernes Atabonfack, Confidence Raymond, Craig Jones, Ilerioluwakiiye Abolade, Jeremiah Fadugba, Juampablo Heras Rivera, Kenneth Aguh, Lukman E. Ismaila, Maruf Adewole, Mohtady Barakat, Nourou Dine Bankole, Saikat Roy, Toufiq Musah, Willem Boonzaier.. The authors acknowledge the funding support provided to SPARK through the Lacuna Fund for Health and Equity (PI: Udunna Anazodo), the RSNAR & E Foundation (PI: Farouk Dako), the University of Washington Population Health Initiative Tier2 Grant (PI: Mehmet Kurt), McGill University Healthy Brain and Healthy Lives (HBHL, Anazodo), and the National Science and Engineering Research Council of Canada (NSERC) Discovery Launch Supplement (PI:UdunnaAnazodo, DGECR-2022-00136).

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

References

1. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning. p. 41–48. ICML '09, Association for Computing Machinery, New York, NY, USA (2009). https://doi.org/10.1145/1553374.1553380, https://doi.org/10.1145/1553374.1553380
2. Chen, G., Dai, Y., Zhang, J., Yap, M.H.: Aau-net: An adaptive attention u-net for breast lesions segmentation in ultrasound images (04 2022). https://doi.org/10.48550/arXiv.2204.12077
3. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. CoRR **abs/2102.04306** (2021), https://arxiv.org/abs/2102.04306
4. 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 **15**(7) (2023). https://doi.org/10.3390/cancers15072112, https://www.mdpi.com/2072-6694/15/7/2112
5. 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., Makhoul, P., Asiimwe, R., Jatho, A., Nnamani, A., Eze, F., Toyobo, O., Fatade, A., Anazodo, U.C., Dako, F., Kawooya, M., Adewole, M.: Bridging the

gap: A community driven and ai-enabled approach to early breast cancer detection in black african women. In: Zhang, T., Saldanha, O.L., Han, L., Rasoolzadeh, N., Garrucho Moras, L., van Dijk, J., Tan, T., Kather, J.N., Mann, R. (eds.) Artificial Intelligence and Imaging for Diagnostic and Treatment Challenges in Breast Care. pp. 289–299. Springer Nature Switzerland, Cham (2026)
6. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48. p. 1050–1059. ICML'16, JMLR.org (2016)
7. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. CoRR **abs/1706.04599** (2017), http://arxiv.org/abs/1706.04599
8. He, Q., Yang, Q., Xie, M.: Hctnet: A hybrid cnn-transformer network for breast ultrasound image segmentation. Comput. Biol. Med. **155**(C) (Mar 2023). https://doi.org/10.1016/j.compbiomed.2023.106629, https://doi.org/10.1016/j.compbiomed.2023.106629
9. Huang, Q., Luo, Y., Zhang, Q.: Breast ultrasound image segmentation: a survey. International Journal of Computer Assisted Radiology and Surgery **12** (01 2017). https://doi.org/10.1007/s11548-016-1513-1
10. Iacob, R., Iacob, E.R., Stoicescu, E.R., Ghenciu, D.M., Cocolea, D.M., Constantinescu, A., Ghenciu, L.A., Manolescu, D.L.: Evaluating the role of breast ultrasound in early detection of breast cancer in low- and middle-income countries: A comprehensive narrative review. Bioengineering (Basel) **11**(3), 262 (Mar 2024). https://doi.org/https://doi.org/10.3390/bioengineering11030262
11. Jiang, J., Jiang, X., Xu, L., Zhang, Y., Zheng, Y., Kong, D.: Noise-robustness test for ultrasound breast nodule neural network models as medical devices. Frontiers in Oncology **Volume 13 - 2023** (2023). https://doi.org/10.3389/fonc.2023.1177225, https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2023.1177225
12. Lei, B., Huang, S., Li, R., Bian, C., Li, H., Chou, Y.H., Cheng, J.Z.: Segmentation of breast anatomy for automated whole breast ultrasound images with boundary regularized convolutional encoder–decoder network. Neurocomputing **321**, 178–186 (2018). https://doi.org/https://doi.org/10.1016/j.neucom.2018.09.043, https://www.sciencedirect.com/science/article/pii/S092523121831107X
13. Liu, Z., Manh, V., Yang, X., Huang, X., Lekadir, K., Campello, V., Ravikumar, N., Frangi, A., Ni, D.: Style curriculum learning for robust medical image segmentation. In: {de Bruijne}, M., {de Bruijne}, M., Cattin, P., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021 - 24th International Conference, Proceedings. pp. 451–460. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Nature, United States (Sep 2021). https://doi.org/10.1007/978-3-030-87193-2_43
14. 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 preprint arXiv:2508.17768 (2025)
15. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M.C.H., Heinrich, M.P., Misawa, K., Mori, K., McDonagh, S.G., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D.:

Attention u-net: Learning where to look for the pancreas. CoRR **abs/1804.03999** (2018), http://arxiv.org/abs/1804.03999
16. Pan, H.B.: The role of breast ultrasound in early cancer detection. Journal of Medical Ultrasound **24**(4), 138–141 (2016). https://doi.org/https://doi.org/10.1016/j.jmu.2016.10.001, https://www.sciencedirect.com/science/article/pii/S0929644116300686
17. Qu, X., Zhou, J., Jiang, J., Wang, W., Wang, H., Wang, S., Tang, W., Lin, X.: Eh-former: Regional easy-hard-aware transformer for breast lesion segmentation in ultrasound images. Inf. Fusion **109**(C) (Sep 2024). https://doi.org/10.1016/j.inffus.2024.102430, https://doi.org/10.1016/j.inffus.2024.102430
18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer International Publishing, Cham (2015)
19. Sood, R., Rositch, A.F., Shakoor, D., Ambinder, E., Pool, K.L., Pollack, E., Mollura, D.J., Mullen, L.A., Harvey, S.C.: Ultrasound for breast cancer detection globally: A systematic review and meta-analysis. Journal of Global Oncology (5), 1–17 (2019). https://doi.org/10.1200/JGO.19.00127, https://ascopubs.org/doi/abs/10.1200/JGO.19.00127, pMID: 31454282
20. Spak, D.A., Plaxco, J.S., Santiago, L., Dryden, M.J., Dogan, B.E.: BI-RADS® fifth edition: A summary of changes. Diagn. Interv. Imaging **98**(3), 179–190 (Mar 2017). https://doi.org/https://doi.org/10.1016/j.diii.2017.01.001
21. Xian, M., Zhang, Y., Cheng, H., Xu, F., Zhang, B., Ding, J.: Automatic breast ultrasound image segmentation: A survey. Pattern Recognition **79**, 340–355 (2018). https://doi.org/https://doi.org/10.1016/j.patcog.2018.02.012, https://www.sciencedirect.com/science/article/pii/S0031320318300645
22. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Álvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. CoRR **abs/2105.15203** (2021), https://arxiv.org/abs/2105.15203
23. Zeevi, T., Lieffrig, E.V., Staib, L.H., Onofrey, J.A.: Spatially-aware evaluation of segmentation uncertainty (2025), https://arxiv.org/abs/2506.16589
24. Zhou, Z., Wu, S., Chang, K.J., Chen, W.R., Chen, Y.S., Kuo, W.H., Lin, C.C., Tsui, P.H.: Classification of benign and malignant breast tumors in ultrasound images with posterior acoustic shadowing using half-contour features. J. Med. Biol. Eng. **35**(2), 178–187 (Apr 2015). https://doi.org/https://doi.org/10.1007/s40846-015-0031-x