

(RESEARCH ARTICLE)



EXTERNAL TESTING OF A FROZEN PUBLIC VISION TRANSFORMER FOR BREAST ULTRASOUND CLASSIFICATION ON BUS-UCLM: MALIGNANT SENSITIVITY, CALIBRATION, AND CONFIDENCE-BASED ABSTENTION

Abdullahi Oluwatobi Badamosi *, Ayodeji Emmanuel Popoola, Adeola Olusola Ekisola and Dabira Ghebemi Adefadipe

Department of Biomedical Engineering, University of Ilorin, Ilorin, Nigeria.

* Corresponding Author

ORCID Details

Abdullahi Oluwatobi Badamosi: <https://orcid.org/0009-0003-3212-2723>

Global Journal of Engineering and Technology Advances, 2026, 28(03), 288–297

Publication history: Received on 11 August 2026; revised on 24 September 2026; accepted on 26 September 2026

Article DOI: <https://doi.org/10.30574/gjeta.2026.28.3.0265>

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

Publicly released breast ultrasound artificial intelligence models may not retain their development performance when transferred unchanged to independently acquired data. This study externally tested the frozen three-class (benign, malignant, normal) Vision Transformer Parveshiiii/breast-cancer-detector, reported as trained on BUSI-derived data, on BUS-UCLM without retraining, fine-tuning, recalibration, or threshold optimization. The study followed a prespecified protocol frozen before full external inference. The analysis included 683 ultrasound images from 38 patients, including 90 malignant images; malignant sensitivity under the model's frozen decision rule was the primary outcome. Sensitivity was 43.3% (95% CI 24.2–61.4%) and specificity was 92.2% (87.9–95.8%), while malignant-score ROC-AUC was 0.785 (0.706–0.865), indicating moderate discrimination despite poor sensitivity at the frozen operating rule. Fifty-one malignant images were missed across 14 patients; 48/51 (94.1%) were classified as benign, and 27/51 (52.9%) had maximum-softmax confidence $\geq 80\%$. Overall three-class accuracy was 45.8%, whereas mean maximum-softmax confidence was 79.6%; Expected Calibration Error was 0.339 and multiclass Brier score was 0.808, consistent with marked overconfidence. At a 0.90 confidence-acceptance threshold, only 232/683 images (34.0%) were retained, accepted accuracy was 61.2%, and 16 malignant false negatives remained accepted. The model therefore showed limited external performance on BUS-UCLM, with low malignant sensitivity, miscalibration, and persistent high-confidence malignant misses. These findings highlight the value of reporting operating-point performance, calibration, patient-clustered confidence intervals, and selective-classification trade-offs alongside discrimination in external medical-imaging AI evaluation.

Keywords: Breast Ultrasound; Artificial Intelligence; External Validation; Vision Transformer; Calibration; Selective Classification

1 INTRODUCTION

Breast ultrasound is widely used to evaluate breast abnormalities, but interpretation can vary with operator technique, acquisition settings, lesion appearance, equipment, and patient population. These sources of variability have made breast ultrasound an active area for computer-aided diagnosis and deep-learning research, with numerous studies reporting strong lesion-classification performance [1,2]. Public datasets such as BUSI have accelerated model development by providing labeled benign, malignant, and normal ultrasound images [3].

Strong development-set performance does not establish that a model will behave similarly on independently acquired data. External testing is therefore important for assessing generalizability. A systematic review of radiologic deep-learning algorithms found that external evaluation remained relatively uncommon and that performance often decreased outside the development setting [4]. In breast ultrasound specifically, dataset shift and patient leakage can materially affect apparent performance [5]. Differences in scanner hardware, image processing, acquisition protocol, case mix, and institutional workflow may all change the input distribution without changing the nominal diagnostic task.

Reliability also cannot be summarized by discrimination alone. A classifier may preserve useful ranking information while producing poor hard-class decisions at its inherited operating rule. Neural networks may additionally assign high softmax scores to incorrect predictions, and uncertainty quality can deteriorate under distribution shift [6,7]. Selective classification attempts to address this problem by allowing a model to abstain on lower-confidence cases, but any apparent gain in accepted-case performance must be interpreted together with the proportion of cases deferred [8].

BUS-UCLM is an independently acquired public breast-ultrasound dataset from Ciudad Real General University Hospital / University of Castilla-La Mancha, Spain, containing 683 images from 38 patients and preserving patient identifiers for patient-aware analysis [9]. The present study independently tested a publicly released pretrained Vision Transformer on BUS-UCLM without retraining, fine-tuning, recalibration, threshold optimization, or other adaptation to the external data. The primary endpoint was malignant sensitivity under the frozen three-class decision rule. Secondary analyses examined three-class and binary performance, patient-clustered uncertainty, calibration, confidence behavior, selective classification, risk-coverage behavior, and clinically important failure patterns. Reporting decisions were informed by the CLAIM 2024 recommendations for medical-imaging AI studies [10]. The central question was not whether a new model could be trained to perform well on BUS-UCLM, but whether an already released public classifier would preserve useful and reliable behavior when transferred unchanged to an independent cohort.

2 MATERIALS AND METHODS

2.1 Study design and pretrained model

This was a prespecified independent external-testing study. Research Protocol v1.0 was frozen before complete inference on BUS-UCLM and defined the primary outcome, eligibility rules, analysis sequence, clustered bootstrap procedure, uncertainty measures, abstention thresholds, calibration analysis, and failure-analysis priorities. The study was a research benchmark and did not evaluate the model as a standalone clinical diagnostic device.

The evaluated model was Parveshiiii/breast-cancer-detector, obtained from Hugging Face and pinned to revision 7c4c1ac11f5f80d382cc641ec6f2d3989e7fa73c [11]. It is a ViTForImageClassification model fine-tuned from google/vit-base-patch16-224-in21k, with 224 x 224-pixel inputs, 16 x 16 patches, 12 transformer encoder layers, 12 attention heads, and three output classes: benign, malignant, and normal. Model weights remained frozen throughout the study. No BUS-UCLM image, label, patient identifier, or derived result was used to update parameters, select alternative weights, or recalibrate the classifier. The model documentation links its training data to the public BUSI/Al-Dhabyani dataset family [3].

2.2 External dataset, eligibility, and reference standard

BUS-UCLM contains 683 ultrasound images from 38 patients: 419 normal, 174 benign, and 90 malignant images. Images were acquired using a Siemens ACUSON S2000 system during 2022-2023; malignant lesions were reported by the dataset authors as biopsy-confirmed [9]. The image was the model prediction unit, while patient identifiers were retained for clustered uncertainty estimation because multiple images could originate from the same patient.

Before inference, the dataset was audited for file count, patient count, class distribution, dimensions, readability, label availability, exact duplicate hashes, masks, and available metadata flags. The primary analysis included every official source image that decoded successfully and mapped to a prespecified class. Outcome-based exclusions were prohibited, and segmentation masks were not used as classifier inputs. Official BUS-UCLM labels were used as the reference standard. For binary analysis, malignant was positive and benign plus normal were combined as non-malignant.

A prespecified clean-input sensitivity analysis excluded images flagged in BUS-UCLM metadata as containing visual marks/annotations, Doppler information, or combined/multi-panel content. The full eligible dataset remained the primary analysis population. Key model and cohort characteristics are summarized in Table 1.

Table 1: Model and external test cohort characteristics

Characteristic	Value
Evaluated model	Parveshiiii/breast-cancer-detector
Pinned revision	7c4c1ac11f5f80d382cc641ec6f2d3989e7fa73c
Base architecture	Vision Transformer (ViT-B/16)
Input / patch size	224 x 224 pixels / 16 x 16 pixels
Output classes	Benign, malignant, normal
External dataset	BUS-UCLM
External cohort	683 images from 38 patients
Class distribution	419 normal, 174 benign, 90 malignant
Ultrasound system	Siemens ACUSON S2000
Model adaptation on BUS-UCLM	None
Primary outcome	Malignant sensitivity
Uncertainty estimation	5,000 patient-cluster bootstrap replicates, seed 42

2.3 Preprocessing and frozen inference

All images were converted to RGB and processed with the ViTImageProcessor distributed with the pinned model revision. The processor resized images to 224 x 224 pixels using bilinear interpolation, rescaled pixel values by 1/255, and normalized channels using mean (0.5, 0.5, 0.5) and standard deviation (0.5, 0.5, 0.5). The same preprocessing was applied to every image.

Inference was implemented in PyTorch with Hugging Face Transformers. The classifier was placed in evaluation mode, gradients were disabled, logits were transformed with softmax, and the class with the highest score was recorded as the prediction. The frozen prediction table stored image and patient identifiers, reference and predicted classes, all class scores and logits, maximum-softmax confidence, top-two probability margin, model identifier and revision, execution device, and prespecified metadata flags. Integrity checks confirmed row count, unique images, complete outputs, valid class-score sums, and model/revision consistency before the prediction file was frozen for analysis.

2.4 Outcomes and statistical analysis

The primary outcome was image-level malignant sensitivity on the full external cohort, with a patient-clustered 95% confidence interval. Three-class outcomes included accuracy, balanced accuracy, macro precision, macro recall, macro F1 score, per-class metrics, and a confusion matrix. For malignant versus non-malignant analysis, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1 score, balanced accuracy, ROC-AUC, and PR-AUC were calculated. ROC-AUC and PR-AUC used the malignant softmax score as the continuous decision variable.

Confidence intervals were estimated using 5,000 nonparametric bootstrap replicates sampled at the patient-cluster level with random seed 42. This preserved within-patient image grouping and avoided treating all 683 images as statistically independent. Percentile 95% confidence intervals were reported.

Model confidence was defined as the maximum softmax score; prediction margin was the difference between the two highest softmax scores. Calibration was assessed using 10 equal-width confidence bins, Expected Calibration Error (ECE), and multiclass Brier score. Confidence was treated as a model score rather than a clinical probability.

Selective classification was evaluated at prespecified maximum-softmax acceptance thresholds from 0.50 to 0.95 in 0.05 increments. For each threshold, overall coverage, malignant coverage, accepted-case accuracy, accepted malignant sensitivity and specificity, accepted false negatives and false positives, and abstained malignant cases were recorded. Risk-coverage behavior was also evaluated continuously, with risk defined as one minus accepted-case accuracy. The empirical area under the risk-coverage curve (AURC) was descriptive only because no normalized comparator was

prespecified. Malignant false negatives and false-positive malignant predictions were audited for error direction, confidence level, patient concentration, and available metadata characteristics.

2.5 Reproducibility

The analysis was implemented in Python using PyTorch, Transformers, scikit-learn, pandas, NumPy, SciPy, Matplotlib, and Pillow. Model revision, preprocessing, random seed, frozen prediction outputs, analysis scripts, and generated tables and figures were recorded in the project repository. No post hoc retraining or recalibration was performed after observing external-test labels.

3 RESULTS

3.1 External cohort and classification performance

All 683 BUS-UCLM images from 38 patients were eligible for the primary analysis. The reference distribution was 419 normal, 174 benign, and 90 malignant images. The prespecified clean-input subset contained 521 images. Three-class accuracy was 0.4583 (95% CI 0.3756-0.5347), balanced accuracy was 0.5338 (95% CI 0.4564-0.6073), and macro F1 was 0.4544 (95% CI 0.3608-0.5346). The confusion matrix is shown in Figure 1.

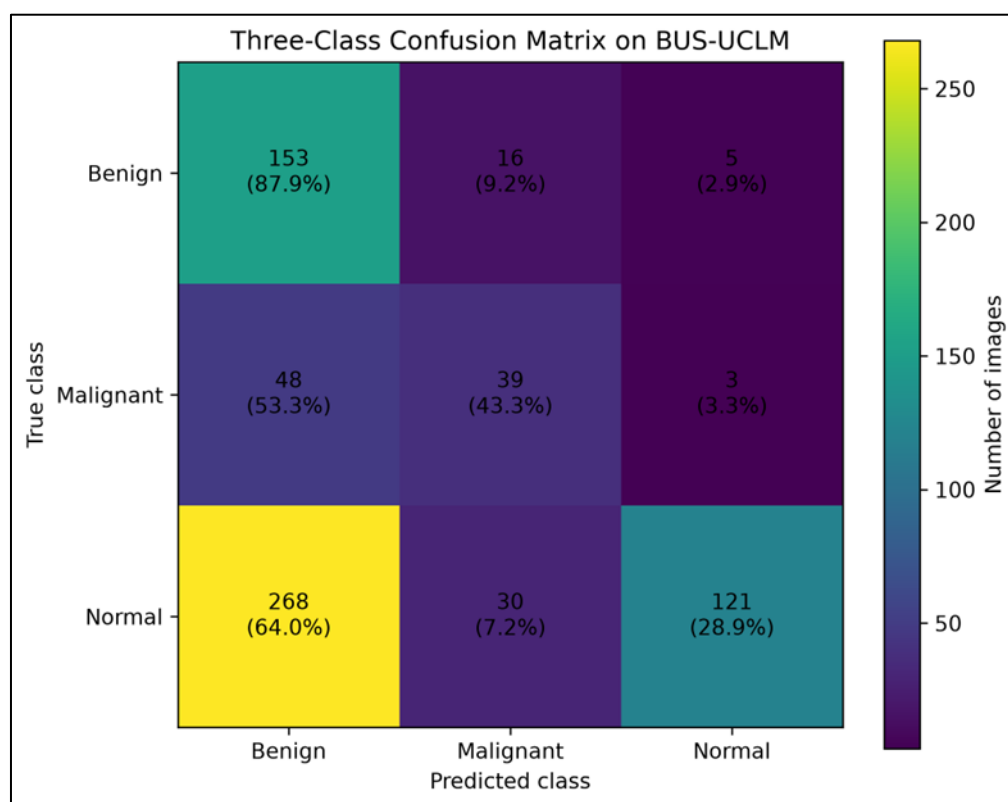


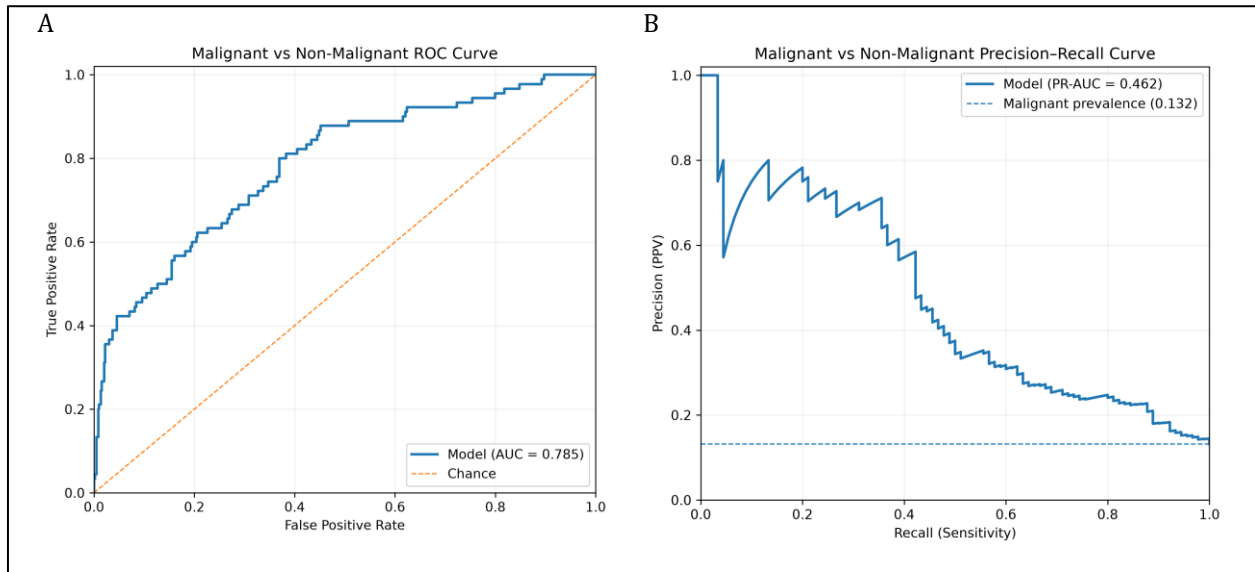
Figure 1: Three-class confusion matrix for external BUS-UCLM evaluation.

Among benign images, 153/174 were classified as benign, 16 as malignant, and 5 as normal. Among malignant images, 39/90 were classified as malignant, 48 as benign, and 3 as normal. Among normal images, 121/419 were classified as normal, 268 as benign, and 30 as malignant. Thus, malignant-to-benign classification was the dominant malignant error direction.

For malignant versus non-malignant classification, the frozen hard-class decision produced 39 true positives, 547 true negatives, 46 false positives, and 51 false negatives. Malignant sensitivity was 0.4333 (95% CI 0.2417-0.6136), specificity 0.9224 (95% CI 0.8788-0.9584), PPV 0.4588, NPV 0.9147, F1 0.4457, and balanced accuracy 0.6779. Despite the low hard-decision sensitivity, the continuous malignant score retained ROC-AUC 0.7855 (95% CI 0.7060-0.8645) and PR-AUC 0.4619 (95% CI 0.2500-0.6487). Performance estimates are summarized in Table 2, and the ROC and precision-recall curves are shown in Figure 2.

Table 2: External performance on BUS-UCLM with patient-clustered 95% confidence intervals

Outcome	Estimate	95% CI
Three-class accuracy	0.4583	0.3756-0.5347
Three-class balanced accuracy	0.5338	0.4564-0.6073
Macro F1	0.4544	0.3608-0.5346
Malignant sensitivity	0.4333	0.2417-0.6136
Malignant specificity	0.9224	0.8788-0.9584
Positive predictive value	0.4588	0.2222-0.6863
Negative predictive value	0.9147	0.8574-0.9589
Binary F1 score	0.4457	0.2446-0.6114
Binary balanced accuracy	0.6779	0.5776-0.7740
ROC-AUC	0.7855	0.7060-0.8645
PR-AUC	0.4619	0.2500-0.6487

**Figure 2: Malignant-versus-non-malignant discrimination on BUS-UCLM. (A) Receiver operating characteristic curve; ROC-AUC 0.7855 (95% CI 0.7060-0.8645). (B) Precision-recall curve; PR-AUC 0.4619 (95% CI 0.2500-0.6487). Both curves use the malignant softmax score as the continuous decision variable.**

In the clean-input subset ($n = 521$), three-class accuracy was 0.3973 and malignant sensitivity was 0.3889. Binary specificity was 0.9258, ROC-AUC 0.8018, and PR-AUC 0.2828. Because this subset contained only 36 malignant images and had a different case composition, these results were treated as a sensitivity analysis rather than evidence that visually cleaner images were intrinsically more difficult.

3.2 Confidence, calibration, and selective classification

Mean maximum-softmax confidence across all images was 0.7961 while overall accuracy was 0.4583. Correct predictions had higher mean confidence than incorrect predictions (0.8250 vs 0.7717). Malignant true positives were also more confident than malignant false negatives on average (0.8981 vs 0.8038), but clinically important errors were not confined to low-confidence cases.

Calibration was poor with respect to top-class correctness. ECE was 0.3386 and the multiclass Brier score was 0.8080. In the 0.80-0.90 confidence bin, mean confidence was 0.8531 while empirical accuracy was 0.3791; in the 0.90-1.00 bin, mean confidence was 0.9479 while accuracy was 0.6121. The reliability diagram is shown in Figure 4.

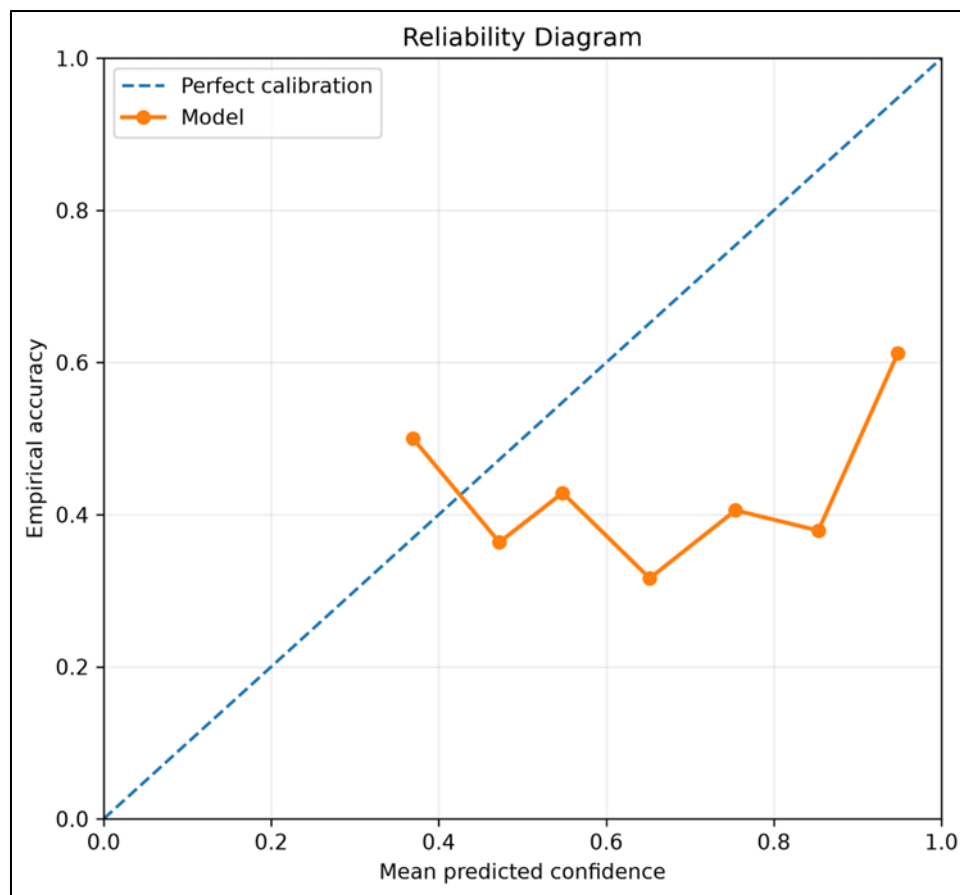


Figure 3: Reliability of maximum-softmax confidence during external testing.

Increasing the acceptance threshold reduced accepted-case risk but sharply reduced coverage. Representative operating points are shown in Table 3; the complete prespecified sweep is provided in Supplementary Table S1. At a threshold of 0.80, overall coverage was 0.5637 and 27 malignant false negatives remained accepted. At 0.90, coverage fell to 0.3397, accepted-case malignant sensitivity was 0.6279, and 16 malignant false negatives remained accepted. At 0.95, only 17.4% of all images and 18.9% of malignant images were accepted; five malignant false negatives remained among the accepted cases. Accepted-case sensitivity therefore increased partly because most malignant images were deferred, not because the full-cohort malignant detection rate approached the accepted-case estimate.

Table 3: Representative confidence-based abstention operating points

Threshold	Overall coverage	Malignant coverage	Malignant abstained, n	Accepted accuracy	Accepted malignant sensitivity*	Accepted malignant FN
0.50	0.9649	1.0000	0	0.4613	0.4333	51
0.70	0.7189	0.8556	13	0.4949	0.4935	39
0.80	0.5637	0.6667	30	0.5195	0.5500	27
0.90	0.3397	0.4778	47	0.6121	0.6279	16
0.95	0.1742	0.1889	73	0.8403	0.7059	5

Accepted-case metrics were calculated only among cases meeting the confidence threshold. The full 0.50-0.95 threshold sweep is reported in Supplementary Table S1.

The risk-coverage curve showed the same trade-off continuously (Figure 3). At approximately 25% coverage, accepted accuracy was 0.7135 (risk 0.2865); at 50% coverage, accuracy was 0.5380; at 75% coverage, accuracy was 0.4883; and at full coverage, accuracy was 0.4583. Descriptive AURC was 0.3831.

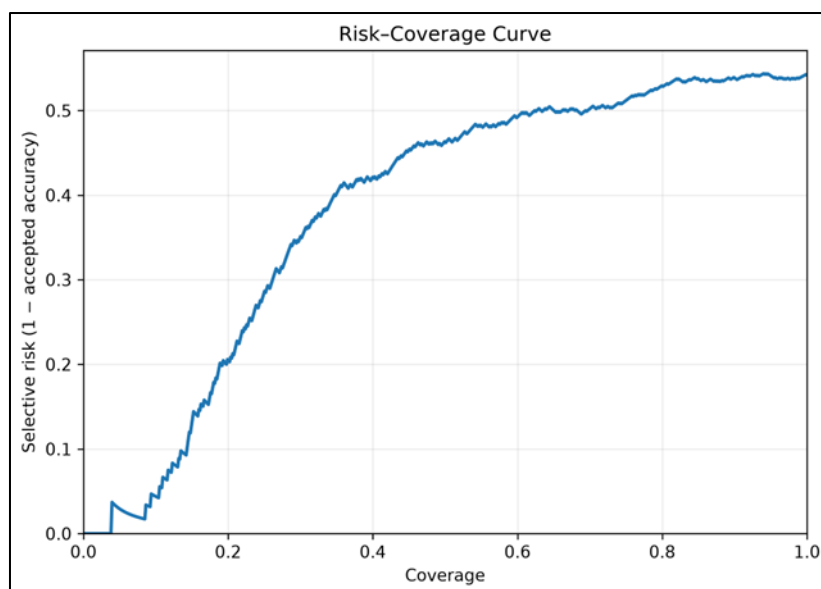


Figure 4: Risk-coverage relationship under confidence-based selective classification.

3.3 Failure patterns

Of 51 malignant false negatives, 48 (94.1%) were predicted benign and three (5.9%) were predicted normal. Twenty-seven false negatives (52.9%) had confidence at least 0.80, 16 (31.4%) at least 0.90, and five (9.8%) at least 0.95. False negatives occurred across 14 patients, with the five highest-count patients contributing 29/51 (56.9%). Of 46 false-positive malignant predictions, 30 arose from normal images and 16 from benign images; 21 had confidence at least 0.80. Detailed failure counts are provided in Supplementary Table S2.

4 DISCUSSION

Independent testing of the frozen public Vision Transformer on BUS-UCLM revealed a substantial gap between ranking performance and the inherited hard-class decision rule. Malignant sensitivity was 0.4333, meaning that 51 of 90 malignant images were not classified as malignant, while the malignant score retained ROC-AUC 0.7855. These metrics answer different questions: ROC-AUC reflects ranking across thresholds, whereas sensitivity reflects behavior at the model's fixed categorical decision rule. Similar divergence between discrimination and operating-point performance has been described under acquisition shift in medical imaging [12].

The most important error pattern was malignant-to-benign confusion. Forty-eight of 51 malignant false negatives were assigned to the benign class. The present study cannot establish why this occurred. Benign and malignant breast lesions can share overlapping sonographic appearances, including malignant tumors that mimic fibroadenomas [13]. Distributional differences between the model-development data and BUS-UCLM may also have altered the usefulness of learned features. Cross-institutional imaging studies have shown that models can exploit site- or acquisition-associated signals and lose reliability when those relationships change [14,15]. The BUSI lineage associated with the released model is benign-majority [3], but the exact composition of the model's effective training subset is insufficiently documented to attribute the observed errors to class imbalance. These possibilities should therefore be treated as hypotheses rather than demonstrated mechanisms.

The findings also show why confidence should not be interpreted as a safety guarantee. More than half of malignant false negatives had maximum-softmax confidence at least 0.80, almost one third remained above 0.90, and some exceeded 0.95. Mean confidence substantially exceeded observed top-class accuracy, with ECE 0.3386. This is consistent with evidence that neural-network softmax scores may be poorly calibrated and that uncertainty quality can deteriorate under dataset shift [6,7]. Failure-detection work in medical imaging similarly emphasizes that softmax values are not inherently calibrated confidence measures [16].

Confidence still contained useful relative information. Correct predictions were more confident than incorrect predictions on average, and selective classification identified lower-risk subsets. However, the benefit came at a substantial coverage cost. At a 0.90 threshold, only 34.0% of images were accepted and 47 of 90 malignant images were abstained; 16 malignant false negatives still remained accepted. At 0.95, accepted-case accuracy rose to 0.8403, but only 17.4% of all images and 17 of 90 malignant images were accepted. Thus, the increase in accepted-case malignant sensitivity should not be interpreted as equivalent to detecting approximately 70% of all malignant cases. Prior work

has also shown that the value of uncertainty-based referral depends strongly on calibration [17]. The present study did not recalibrate the model because doing so would have changed the prespecified frozen-model external-testing question.

The clean-input analysis did not restore malignant sensitivity, but this should not be interpreted causally. The subset had a different case mix and only 36 malignant images. It shows only that the main failure pattern was not confined to images carrying the prespecified visible metadata flags. Likewise, the concentration of errors within some patients is descriptive; the relatively small number of independent patients limits precision and motivated patient-clustered confidence intervals.

From a clinical-engineering perspective, these results do not support use of the evaluated frozen model as a standalone decision tool on data represented by this cohort. Public availability, a familiar architecture, or strong source-domain performance are not substitutes for local external testing. Larger multicenter breast-ultrasound studies have reported stronger performance under different development and validation designs [18,19], and clinical breast assessment ordinarily integrates more than a single isolated ultrasound image. Multimodal work has shown that clinical and demographic information can improve classification relative to ultrasound information alone [20]. The present study therefore evaluates one image-level public classifier, not a complete clinical diagnostic pathway.

The study has several limitations. First, BUS-UCLM contains 683 images but only 38 patients. The clustered bootstrap accounts for within-patient dependence, but the small number of independent patients produces wide uncertainty, particularly for malignant sensitivity. Second, only one external cohort and one scanner platform were tested, so the findings demonstrate a transfer limitation on BUS-UCLM rather than universal behavior across hospitals, devices, or populations. Third, the prediction unit was the image; patient-level diagnostic performance was not evaluated. Fourth, no retraining, threshold optimization, domain adaptation, or recalibration was performed, by design. Fifth, maximum-softmax confidence and top-two margin are simple uncertainty proxies, and no independently validated out-of-distribution detector was tested. Finally, lesion masks were not used as inputs and no attribution analysis was performed, so the study identifies error patterns but cannot determine their causal image features or representations.

Future work should evaluate the model and comparable systems across larger multi-institution and multivendor cohorts with patient-level separation. Recalibration or threshold-selection procedures should be developed on data separate from the final external test set. Stronger uncertainty and out-of-distribution methods should be externally validated rather than assumed to provide safeguards. Controlled comparisons across architectures, patch sizes, and image resolutions could test whether model design contributes to transfer behavior. These priorities align with lifecycle-oriented guidance for trustworthy healthcare AI, which emphasizes evaluation across development, validation, deployment, and monitoring stages [21].

5 CONCLUSION

Independent external testing of a frozen public breast-ultrasound Vision Transformer on BUS-UCLM identified low malignant sensitivity, predominant malignant-to-benign errors, poor calibration of maximum-softmax confidence, and clinically important high-confidence misses. The model retained useful ranking information, but confidence-based abstention reduced accepted-case risk only at substantial coverage cost and did not remove all high-confidence malignant false negatives. These findings support patient-aware external evaluation that examines calibration, selective-prediction trade-offs, and failure patterns in addition to discrimination metrics. The results should be interpreted as a reproducible external test of one public model on one independent cohort, not as clinical validation or a universal assessment of breast-ultrasound AI.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Acknowledgments

The authors would like to thank the Department of Biomedical Engineering, Faculty of Engineering and Technology, University of Ilorin, and the lecturers of the Department for their guidance and support.

Disclosure of conflict of interest

The authors declare that they have no competing interests.

Statement of ethical approval

This study involved secondary analysis of the publicly available BUS-UCLM dataset and did not involve new participant recruitment, image acquisition, or direct interaction with participants. The original BUS-UCLM study was reviewed and

approved by the Ethics Committee of Ciudad Real General University Hospital as part of project PID2021-127567NB-I00.

Statement of informed consent

Informed consent was obtained from all participants for the collection and sharing of their data.

Data availability

The BUS-UCLM dataset used for external testing is publicly available through Mendeley Data (Version 3; DOI: 10.17632/7fvvj4j7p7.3) under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence. No new participant data were collected for this study.

Code availability

The analysis code, reproducibility materials, evaluation scripts, and project documentation associated with this study are available at <https://github.com/Abdooorl/breast-ultrasound-ai-validation>.

Author contributions

Abdullahi Oluwatobi Badamosi: Conceptualization, Methodology, Investigation, Data curation, Formal analysis, and research execution. Ayodeji Emmanuel Popoola: Supervision, Methodology, Validation, Investigation, and Data curation. Adeola Olusola Ekiola: Supervision, Validation, Methodology, and Writing - original draft preparation. Dabira Gbebemi Adefadipe: Formal analysis, Visualization, and Writing - review and editing. All authors reviewed and approved the final version of the manuscript.

Disclosure of artificial intelligence assistance

Generative artificial intelligence tools, including OpenAI ChatGPT, were used to assist with manuscript organization, language editing, and programming and debugging support during preparation of this study. Artificial intelligence was not used to fabricate or alter research data or to determine the scientific conclusions. All AI-assisted code, analyses, references, statements, and manuscript content were reviewed and verified by the authors, who take full responsibility for the accuracy and integrity of the submitted work.

REFERENCES

- [1] Han S, Kang HK, Jeong JY, Park MH, Kim W, Bang WC, et al. A deep learning framework for supporting the classification of breast lesions in ultrasound images. *Physics in Medicine and Biology*. 2017;62(19):7714–28.
- [2] Ma T, Wang Z, Dong J, Cheng Y, Zhao H, Cui X. Ultrasound-based artificial intelligence for breast lesion classification. *Frontiers in Oncology*. 2026;16:1759194.
- [3] Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. *Data in Brief*. 2020;28:104863.
- [4] Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: A systematic review. *Radiology: Artificial Intelligence*. 2022;4(3):e210064.
- [5] Wang L. Breast ultrasound AI under dataset shift: A patient-leakage-aware benchmark. *Diagnostics*. 2026;16(10):1537.
- [6] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: *Proceedings of the 34th international conference on machine learning*. PMLR; 2017. p. 1321–30. (Proceedings of machine learning research; vol. 70).
- [7] Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In: *Advances in neural information processing systems*. 2019.
- [8] Geifman Y, El-Yaniv R. Selective classification for deep neural networks. In: *Advances in neural information processing systems*. 2017.
- [9] Vallez N, Bueno G, Deniz O, Rienda MA, Pastor C. BUS-UCLM: Breast ultrasound lesion segmentation dataset. *Scientific Data*. 2025;12(1):242.
- [10] Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al. Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. *Radiology: Artificial Intelligence*. 2024;6(4):e240300.

- [11] Parveshiiii. Breast cancer detector: A three-class breast ultrasound classifier [Internet]. Hugging Face model repository; 2026. Available from: <https://huggingface.co/Parveshiiii/breast-cancer-detector>
- [12] Roschewitz M, Khara G, Yearsley J, Sharma N, James JJ, Ambrózay É, et al. Automatic correction of performance drift under acquisition shift in medical image classification. *Nature Communications*. 2023;14:6608.
- [13] Swoboda M, Deeg J, Egle D, Ladenhauf V, Galijasevic M, Plöbst C, et al. Identification of differentiating sonographic features between fibroadenomas and malignant tumors of the breast mimicking fibroadenomas: 10-year experience in 421 histologically verified cases. *Ultraschall in der Medizin*. 2025;46(3):291–8.
- [14] Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*. 2018;15(11):e1002683.
- [15] Ong Ly C, Unnikrishnan B, Tadic T, Patel T, Duhamel J, Kandel S, et al. Shortcut learning in medical AI hinders generalization: Method for estimating AI model generalization without external data. *npj Digital Medicine*. 2024;7:124.
- [16] Ahmed S, Dera D, Hassan SU, Bouaynaya N, Rasool G. Failure detection in deep neural networks for medical imaging. *Frontiers in Medical Technology*. 2022;4:919046.
- [17] Zhang R, Gatsonis C, Steingrimsson JA. Role of calibration in uncertainty-based referral for deep learning. *Statistical Methods in Medical Research*. 2023;32(5):927–43.
- [18] Gu Y, Xu W, Lin B, An X, Tian J, Ran H, et al. Deep learning based on ultrasound images assists breast lesion diagnosis in china: A multicenter diagnostic study. *Insights into Imaging*. 2022;13(1):124.
- [19] Xiang H, Wang X, Xu M, Zhang Y, Zeng S, Li C, et al. Deep learning-assisted diagnosis of breast lesions on US images: A multivendor, multicenter study. *Radiology: Artificial Intelligence*. 2023;5(5):e220185.
- [20] Pfob A, Sidey-Gibbons C, Barr RG, Duda V, Alwafai Z, Balleyguier C, et al. The importance of multi-modal imaging and clinical information for humans and AI-based algorithms to classify breast masses (INSPIRED 003): An international, multicenter analysis. *European Radiology*. 2022;32(6):4101–15.
- [21] Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*. 2025;388:e081554.