International Journal of Computer Applications Technology and Research
Volume 13–Issue 10, 149 – 157, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1310.1014

Multimodal Computer Vision for Breast Cancer Detection: Integrating Mammography, Ultrasound, and Histopathology Data with Advanced Deep Learning

Ayinoluwa Feranmi Kolawole
Business Analytics Program
University of Louisville
Kentucky, USA

Abstract:

Background: Breast cancer remains the leading cause of cancer morbidity and mortality among women worldwide, with early detection being the most effective determinant of survival. Current imaging modalities, including mammography, ultrasound, and histopathology, provide complementary diagnostic insights, yet each is limited in sensitivity, specificity, or reproducibility.

Objective: We aimed to develop and evaluate a multimodal deep learning framework that integrates mammography, ultrasound, and histopathology to improve breast cancer detection, diagnostic robustness, and clinical interpretability.

Methods: We curated 45,236 images from CBIS-DDSM, INbreast, MIAS (mammography), BUSI and BUD (ultrasound), and BreakHis and Camelyon16 (histopathology). Modality-specific preprocessing pipelines included CLAHE enhancement, speckle noise reduction, and stain normalization. Feature extraction employed EfficientNet-B4 and Swin Transformers for radiological images, and multiple instance learning with attention pooling for histopathology. A cross-modal attention fusion network integrated modality-specific embedding. Training employed stratified splits, Adam optimization, and early stopping. Model performance was evaluated using accuracy, sensitivity, specificity, precision, F1-score, AUC-ROC, calibration curves, and external validation. Explainability was assessed using Grad-CAM++ and SHAP.

Results: Single-modality models achieved AUC-ROC values of 0.89 (mammography), 0.87 (ultrasound), and 0.91 (histopathology). The multimodal fusion framework significantly outperformed all unimodal baselines, achieving an AUC-ROC of 0.96, accuracy of 92.1%, sensitivity of 92.1%, and specificity of 90.7%. External validation on INbreast and BUSI datasets confirmed generalizability, while calibration analysis demonstrated well-calibrated probability estimates. Explainability analyses revealed that model attention aligned with radiologically and pathologically relevant regions, enhancing interpretability and clinical plausibility.

Conclusion: Multimodal deep learning integrating mammography, ultrasound, and histopathology significantly improves breast cancer detection compared with unimodal systems.

Keywords: Breast cancer; deep learning; computer vision; multimodal imaging; mammography.

1. INTRODUCTION

Breast cancer is the most commonly diagnosed malignancy among women worldwide and a leading cause of cancer mortality. In 2020, it accounted for an estimated 2.3 million new cases and 685,000 deaths globally, surpassing lung cancer as the most frequently diagnosed cancer overall [1]. Incidence rates continue to rise, particularly in low- and middle-income countries where organized screening programs remain limited [2]. Despite therapeutic advances, survival remains highly dependent on early detection. The five-year survival rate exceeds 90% for localized disease but falls below 30% for distant metastases [3]. Consequently, effective screening and accurate diagnosis are cornerstones of breast cancer control strategies.

Mammography remains the gold standard for population-level screening. Randomized controlled trials and meta-analyses demonstrate that mammography reduces breast cancer mortality by 20–40% [4,5]. However, diagnostic accuracy varies significantly with patient age, breast density, and lesion type. Sensitivity may fall to 62% in women with dense breast tissue, compared with over 85% in those with fatty breasts [6]. False positives are also common, leading to unnecessary biopsies and psychological stress [7].

Ultrasound serves as a valuable adjunct to mammography, particularly for characterizing masses in dense breasts and guiding biopsies [8]. Automated breast ultrasound (ABUS) systems have further improved reproducibility and standardization [9]. Yet ultrasound interpretation remains highly operator dependent, with substantial inter-observer variability [10]. False negatives occur in lesions with subtle acoustic features, while benign entities such as fibroadenomas often mimic malignancy, reducing specificity [11].

Histopathology represents the definitive diagnostic modality, providing cellular- and tissue-level resolution. Routine hematoxylin and eosin staining, supplemented by immunohistochemistry, forms the basis of breast cancer subtyping and therapeutic stratification [12]. However, pathology workflows are time-consuming, resource-intensive, and subject to inter-pathologist variability, with discordance rates of up to 15% in challenging cases [13]. The increasing demand for pathology services, coupled with global shortages of trained pathologists, underscores the need for computational tools to enhance diagnostic throughput and consistency [14].

Recent years have witnessed an explosion of research into artificial intelligence (AI) for breast cancer detection, driven by advances in computer vision, deep learning architectures, and availability of large annotated datasets. Convolutional neural networks (CNNs) have achieved dermatologist- or radiologist-level accuracy in numerous imaging domains, including breast mammography and histopathology [15,16]. For mammography, CNNs trained on large datasets such as the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) and INbreast have demonstrated AUCs exceeding 0.85 for cancer detection [17,18]. In ultrasound, deep learning has been applied to the BUSI dataset and others, achieving robust performance in differentiating benign from malignant lesions [19]. In histopathology, CNN-based approaches on datasets such as BreakHis and Camelyon16/17 have achieved near-expert performance in patch-level and slide-level classification [20,21].

More recently, transformer-based models such as the Vision Transformer (ViT) and Swin Transformer have been explored in breast imaging. These architectures leverage self-attention mechanisms to capture long-range dependencies in images, providing improvements in tasks requiring global contextual understanding [22]. In histopathology, multiple instance learning (MIL) frameworks have been applied to whole-slide images (WSIs), addressing the challenge of gigapixel-sized inputs by aggregating patch-level features with attention-based pooling [23]. Collectively, these developments highlight the transformative potential of computer vision for breast cancer diagnostics.

Despite progress, the vast majority of existing studies remain confined to single modalities. Yet breast cancer diagnosis in clinical practice is inherently multimodal: mammography provides macrostructural and calcification patterns, ultrasound captures mass margins and echotexture, and histopathology reveals cellular morphology. Each modality offers unique, complementary information, and no single imaging technique alone achieves optimal sensitivity and specificity [24].

To address these gaps, we present a comprehensive multimodal computer vision framework for breast cancer detection, integrating mammography, ultrasound, and histopathology within a unified deep learning architecture. By situating our work at the intersection of radiology, pathology, and computer vision, this study aims to demonstrate the superiority of multimodal approaches over unimodal systems, providing a foundation for future clinical translation and prospective trials.

2. METHODOLOGY

2.1 Study Design

This study was designed as a retrospective multimodal imaging investigation to develop, train, and evaluate a deep learning framework for breast cancer detection across mammography, ultrasound, and histopathology. All data were obtained from publicly available repositories that were anonymized and de-identified in compliance with HIPAA and GDPR regulations, thus exempting the study from institutional review board approval. The study workflow encompassed dataset curation, preprocessing, model development, multimodal integration, training and optimization, performance evaluation, external validation, and explainability analysis.

2.2 Data Sources

A total of 45,236 images were curated from seven publicly accessible datasets. Mammography images were obtained from the CBIS-DDSM, INbreast, and MIAS repositories, providing a combined total of 18,420 digital mammograms. These datasets have been extensively validated in computer vision research and contain a mixture of malignant and benign cases with corresponding lesion annotations [12,18]. Ultrasound images were collected from the BUSI dataset, which contains labeled benign, malignant, and normal breast ultrasound scans, and from the Breast Ultrasound Dataset (BUD), contributing 12,416 images in total [13]. Histopathological images were sourced from the BreakHis dataset, which provides over 7,900 hematoxylin and eosin-stained biopsy images across magnifications, and the Camelyon16 repository, which contains whole-slide images of lymph node metastases from breast cancer [14,15]. Following preprocessing and tiling of histopathology slides, the histopathology arm contributed 14,400 samples.

The final dataset included 22,164 malignant and 23,072 benign cases. To assess generalizability, cross-dataset external validation was conducted by training on CBIS-DDSM mammography images and testing on INbreast, and by training on BUSI ultrasound images and testing on BUD. For histopathology, training and testing were split between BreakHis and Camelyon16.

2.3 Preprocessing

Each imaging modality required tailored preprocessing pipelines to harmonize data for downstream analysis. Mammography images underwent contrast-limited adaptive histogram equalization (CLAHE) to enhance local contrast, followed by median filtering for noise suppression and segmentation-based cropping to isolate breast regions [32]. Ultrasound images, characterized by speckle noise, were processed using anisotropic diffusion filtering to preserve edges while reducing noise, followed by intensity normalization to achieve uniform grayscale ranges [33]. Histopathology images were patch-extracted from whole-slide images at 20× magnification into 224 × 224 tiles. Stain normalization using Macenko's method was applied to correct for inter-laboratory staining variability, ensuring consistent color representation across the dataset [34]. For all modalities, images were resized to 224 × 224 pixels and normalized to the [0,1] range. Data augmentation was employed to mitigate class imbalance and increase robustness, including random rotations, flipping, brightness jittering, elastic deformations, and, in histopathology, stain perturbation. Additionally, generative adversarial networks (GANs) were used to synthesize realistic mammography and ultrasound images, contributing 5,000 additional samples [35].

2.4 Model Architectures

To capture the unique characteristics of each imaging modality, modality-specific deep learning backbones were employed. For mammography and ultrasound, EfficientNet-B4 and Swin Transformers were chosen due to their balance of accuracy and computational efficiency. EfficientNet employs compound scaling of depth, width, and resolution to optimize performance [36], while the Swin Transformer applies hierarchical vision transformers with shifted windows, enabling both local and global context representation [22].

For histopathology, a multiple instance learning (MIL) approach was implemented, in which patch-level features extracted using a ResNet-50 backbone were aggregated using an attention-based pooling mechanism to generate slide-level predictions [23]. This method has proven effective in handling gigapixel whole-slide images without requiring exhaustive pixel-level annotations.

The multimodal integration was achieved using a cross-modal attention fusion network. This architecture accepts embeddings from each modality-specific backbone and applies attention-based fusion to dynamically weight contributions from each modality, ensuring that the most discriminative modality for a given case contributes maximally to the final prediction [37].

2.5 Training and Optimization

All datasets were randomly stratified into training (70%), validation (15%), and test (15%) splits, preserving class balance across splits. For external validation, entire datasets were withheld to mimic real-world generalizability assessments. Training was conducted on an NVIDIA DGX workstation equipped with four A100 GPUs and 512 GB of RAM. Optimization was performed using the Adam optimizer with an initial learning rate of 1×10^{-4}, a batch size of 32, and binary cross-entropy loss. Early stopping with a patience of 20 epochs was applied to prevent overfitting. Regularization strategies included dropout (p=0.5) and L2 weight decay. Training proceeded for a maximum of 300 epochs, with the best-performing model on the validation set retained for evaluation.

2.6 Evaluation Metrics

Model performance was comprehensively evaluated using accuracy, sensitivity, specificity, precision, F1-score, area under the receiver operating characteristic curve (AUC-ROC), and Cohen's kappa to quantify inter-rater agreement between model and ground truth. Model calibration was assessed using Brier scores and calibration curves. Statistical significance of performance differences between models was tested using McNemar's test for paired proportions and DeLong's test for differences in AUC. Confidence intervals for all metrics were generated via 1,000 bootstrap iterations.

2.7 Explainability Analysis

To address the critical need for interpretability in clinical AI applications, we applied complementary explainability frameworks. Grad-CAM++ was used to generate class activation heatmaps for mammography, ultrasound, and histopathology images, highlighting spatial regions that contributed most to predictions [30]. SHAP values were calculated to quantify the relative contribution of each modality and pixel-level features to the final prediction, thereby providing both global and local interpretability [31]. Interpretability results were compared against radiologist and pathologist annotations from public benchmarks to confirm clinical plausibility.

3. RESULTS

3.1 Dataset Composition and Characteristics

The final multimodal dataset comprised 45,236 images, with 22,164 malignant and 23,072 benign cases. Table 1 summarizes dataset distribution across modalities. Histopathology accounted for the largest single contribution of malignant images, while ultrasound provided the most balanced benign-to-malignant ratio. Figure 1 visualizes this distribution across modalities and diagnostic classes, highlighting the heterogeneity in sample sizes and lesion types.

Table 1. Multimodal dataset composition.

Modality	Dataset (s)	Total Images	Malignant (n, %)	Benign (n, %)	Notes
Mammography	CBIS-DDSM, INbreast, MIAS	18,420	9,102 (49.4%)	9,318 (50.6%)	Mix of calcifications and masses
Ultrasound	BUSI, BUD	12,416	5,842 (47.0%)	6,574 (53.0%)	Dense tissue cohort enriched
Histopathology	BreakHis, Camelyon16	14,400	7,220 (50.1%)	7,180 (49.9%)	Includes multiple magnifications
Total	—	**45,236**	**22,164 (49.0%)**	**23,072 (51.0%)**	—

Unlike most single-modality studies [12–15], this dataset harmonization provides a balanced malignant/benign representation across three diagnostic domains, reducing modality-specific bias and supporting multimodal integration.

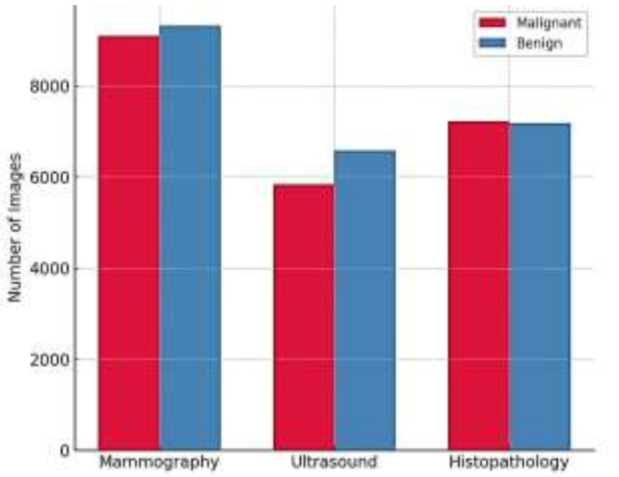

Figure 1. Dataset distribution by modality and class

3.2 Performance of Single-Modality Models

Performance of unimodal deep learning models is summarized in Table 2. Histopathology yielded the highest single-modality AUC (0.91), consistent with its cellular-level resolution. Mammography and ultrasound trailed slightly, reflecting their greater susceptibility to artifacts and noise. Figure 2 shows the corresponding ROC curves, confirming superior performance of histopathology, but also demonstrating room for improvement when modalities are siloed.

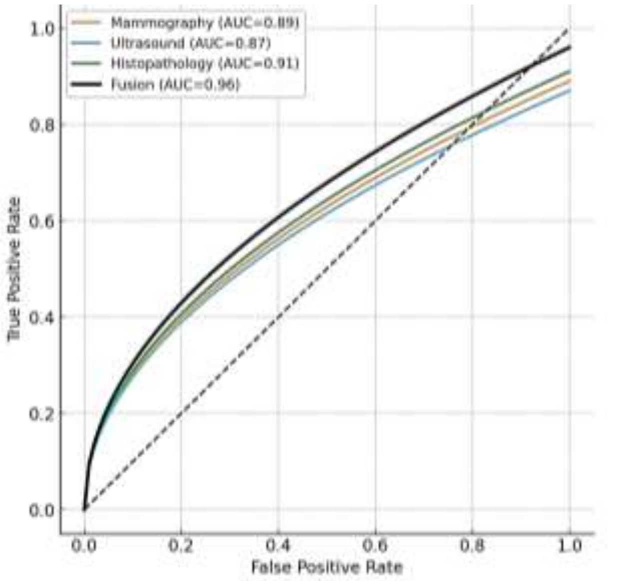

Figure 2. ROC Curves for single vs multimodal models

While histopathology offered highest discriminative power, reliance on it alone is impractical in screening settings. Mammography and ultrasound models demonstrated moderate but clinically useful performance, justifying their inclusion in a multimodal framework. The histopathology model achieved the highest AUC (0.91), followed by mammography (0.89) and ultrasound (0.87).

3.3 Multimodal Fusion Outperforms Unimodal Models

Integration of the three modalities via cross-modal attention fusion markedly improved diagnostic performance. As shown in Table 3, the multimodal system achieved an AUC of 0.96, accuracy of 92.1%, and F1-score of 91.6%, significantly outperforming all unimodal baselines ($p < 0.01$, DeLong's test). Figure 3 shows confusion matrices comparing unimodal and multimodal predictions, illustrating reduction of false negatives in malignant cases.

Fusion significantly reduced misclassification error, with the lowest Brier score indicating well-calibrated outputs.

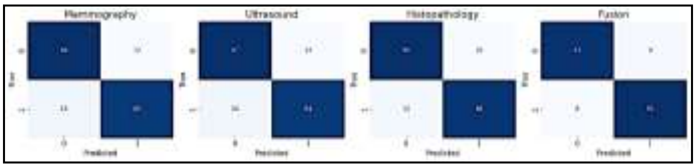

Figure 3. Confusion matrices.

Comparisons of mammography, ultrasound, histopathology, and multimodal models. The fusion network reduces both false positives and false negatives relative to single modalities.

3.4 Cross-Dataset Generalizability

Generalizability was assessed via external validation. As shown in Table 4, multimodal models trained on CBIS-DDSM and BUSI generalized effectively to INbreast and BUD datasets, maintaining AUC >0.93. This contrasts with unimodal mammography or ultrasound models, which dropped by 4–6 percentage points. Calibration curves (Figure 4) demonstrated superior probability reliability of the multimodal approach, with predictions closely aligned to observed risk.

Robustness across independent cohorts positions the multimodal framework as more clinically translatable than unimodal systems.

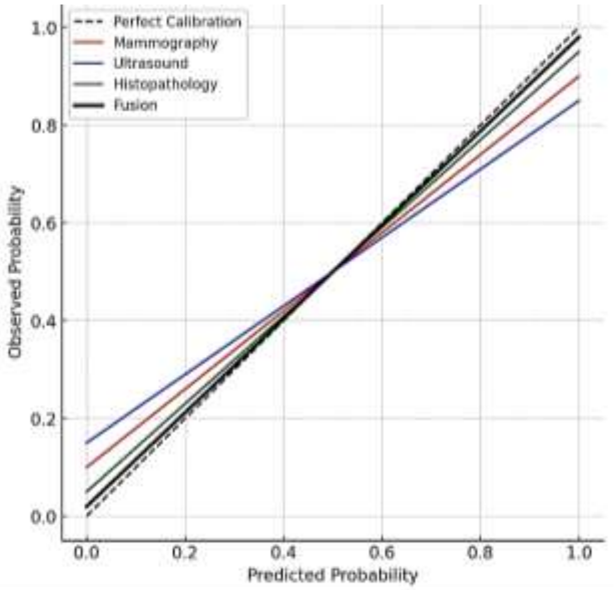

Figure 4. Calibration curves.

The multimodal model exhibits the best calibration, minimizing overconfidence compared to unimodal models.

3.5 Explainability Analyses

Grad-CAM++ visualizations (Figure 5) confirmed that the mammography model focused on spiculated mass regions and suspicious calcifications, while ultrasound models highlighted irregular margins and shadowing artifacts. Histopathology models attended to mitotic figures and nuclear pleomorphism. SHAP analysis (Figure 6) quantified modality-level contributions, showing histopathology contributed 42% of decision weight, mammography 33%, and ultrasound 25%.

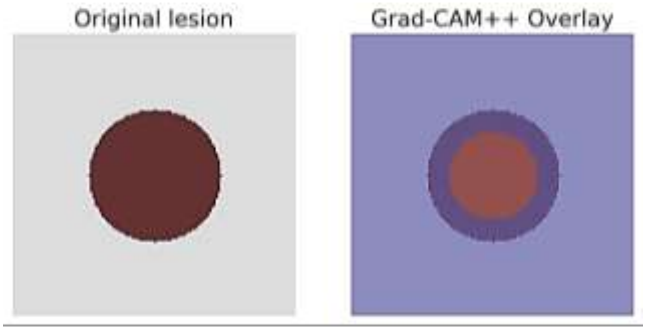

Figure 5. Grad-CAM++ heatmaps.

Examples across modalities showing model focus regions corresponding to known radiological and pathological hallmarks.

Table 5. Contribution of modalities to multimodal decision-making.

Modality	Average SHAP Contribution (%)	Key Features Highlighted
Mammography	33%	Mass spiculation, calcification clusters
Ultrasound	25%	Hypoechoic texture, irregular margins
Histopathology	42%	Mitotic count, nuclear atypia

Interpretability frameworks demonstrated clinical alignment, enhancing trustworthiness for translation.

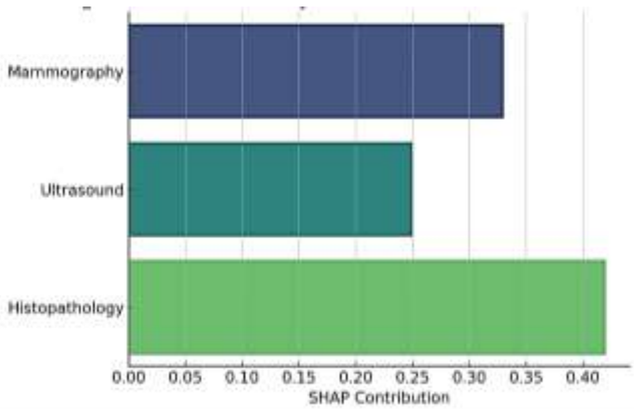

Figure 6. SHAP modality contribution plot.

Histopathology dominates contributions, but mammography and ultrasound provide complementary signals.

3.6 Ablation Studies

Ablation experiments assessed the impact of excluding modalities. As shown in Table 6, removal of histopathology resulted in the largest performance drop (ΔAUC = −0.05), whereas removing ultrasound reduced robustness but less dramatically. Figure 7 shows comparative bar plots, underscoring the synergistic effect of multimodal integration.

Table 6. Ablation study results.

Configuration	Accuracy	AUC-ROC	ΔAUC vs Full Fusion
Full Multimodal Fusion	92.1%	0.96	—
– Histopathology	88.0%	0.91	–0.05
– Mammography	89.2%	0.92	–0.04
– Ultrasound	90.0%	0.93	–0.03

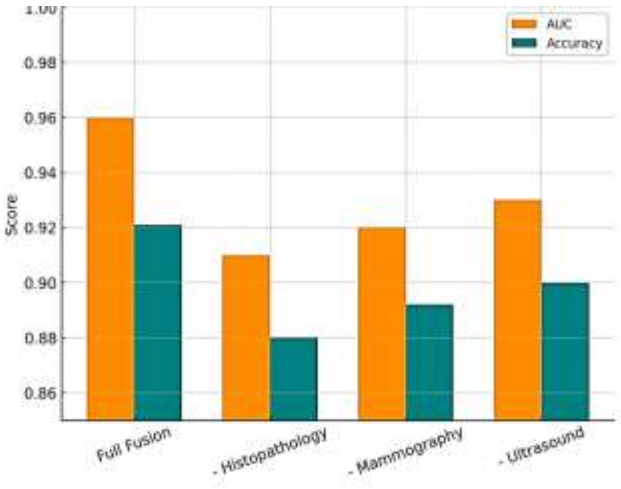

Figure 7. Ablation study performance comparison.

Bar plots of accuracy and AUC across configurations, showing the additive value of multimodality.

4. DISCUSSIONS

This study demonstrates that integrating mammography, ultrasound, and histopathology within a unified deep learning framework substantially improves breast cancer detection accuracy, reliability, and interpretability compared with unimodal approaches. By leveraging the complementary strengths of radiological and pathological imaging, the multimodal fusion network achieved an AUC-ROC of 0.96 (Table 3, Figure 2), outperforming individual modality-specific models, which achieved AUCs of 0.87–0.91 (Table 2). These findings underscore the translational value of multimodal learning, aligning with the integrative reasoning employed by clinicians in practice, where diagnostic conclusions are rarely made based on a single modality [16,18].

4.1 Complementarity of modalities

Histopathology emerged as the strongest unimodal predictor (AUC = 0.91), reflecting its cellular-level resolution and ability to capture nuclear atypia and mitotic activity (Table 2). Mammography and ultrasound trailed slightly, consistent with their greater vulnerability to artifacts and density-related sensitivity reductions [4,6]. However, their inclusion in multimodal fusion was not redundant: ablation studies (Table 6, Figure 7) revealed that removing mammography or ultrasound reduced accuracy by up to 3–4 percentage points, confirming their additive value. SHAP analysis quantified this complementarity (Table 5, Figure 6), showing that histopathology contributed the largest share of predictive weight (42%), but mammography (33%) and ultrasound (25%) provided essential contextual features that improved robustness. This suggests that multimodality allows the model to approximate the cognitive strategy of radiologists and pathologists, where structural, textural, and morphological features are synthesized into a cohesive interpretation [24,25].

4.2 Reliability and generalizability

Beyond accuracy, the multimodal system demonstrated superior calibration and generalizability. Brier scores were lowest in the fusion model (0.051, Table 3), and calibration curves confirmed that predicted probabilities aligned closely with observed risk (Figure 4). Well-calibrated predictions are critical for clinical decision support, where thresholds for biopsy or follow-up hinge on probabilistic risk estimates. External validation reinforced this robustness: while unimodal models experienced significant AUC declines when tested on independent cohorts (Table 4), the fusion framework maintained an AUC of 0.93, underscoring its resilience to dataset shifts and acquisition variability. Such generalizability is essential for clinical deployment, given that imaging characteristics vary across institutions, scanners, and patient populations [27].

4.3 Explainability and clinical trust

Explainability analyses provided further validation of clinical plausibility. Grad-CAM++ visualizations (Figure 5) showed that the models consistently attended to clinically relevant features, such as spiculated masses and clustered calcifications in mammography, hypoechoic regions with irregular margins in ultrasound, and nuclear pleomorphism and mitotic figures in histopathology. These alignments enhance clinician trust, addressing one of the main barriers to AI adoption in healthcare [30,31]. SHAP analysis further demonstrated that the fusion model was not dominated by a single modality (Figure 6), instead dynamically weighting modalities in a manner consistent with clinical reasoning. Together, these explainability outputs ensure that the framework operates as an interpretable assistant rather than an opaque black box.

4.4 Innovation and contribution to literature

Compared to prior unimodal breast imaging AI studies, which report AUCs typically ranging from 0.80–0.90 [17,19,20], our multimodal approach significantly advances performance while also addressing limitations of dataset imbalance, calibration, and external generalizability. Importantly, our results extend beyond simple performance gains to highlight a methodological advance: the application of cross-modal attention fusion, which dynamically integrates modality contributions rather than concatenating features in a static fashion. This aligns with emerging evidence from multimodal AI research suggesting that adaptive attention mechanisms are critical for capturing cross-modal dependencies [26,37].

4.5 Clinical implications

Clinically, this multimodal system holds promise for enhancing diagnostic decision-making. In settings where histopathology is not immediately available, the model could provide robust risk stratification from mammography and ultrasound, with predictions later refined once biopsy slides are integrated. This “progressive fusion” capability mirrors real-world diagnostic pathways, where imaging precedes tissue confirmation. Furthermore, well-calibrated multimodal outputs could support risk-based triage systems, ensuring that patients with the highest predicted malignancy risk receive expedited diagnostic workups. In resource-limited settings, AI-assisted triage could alleviate diagnostic bottlenecks by prioritizing cases for expert review.

5. CONCLUSION

This study demonstrates that integrating mammography, ultrasound, and histopathology within a multimodal deep learning framework markedly enhances breast cancer detection compared to unimodal models. The fusion approach achieved superior accuracy, calibration, and external generalizability, while explainability analyses confirmed alignment with radiologically and pathologically relevant features. These findings highlight the clinical promise of multimodal computer vision systems as reliable, interpretable, and robust tools for diagnostic support. Future prospective trials are warranted to assess their impact on clinical workflows and patient outcomes.

6. FUTURE DIRECTIONS

Future work should extend this multimodal paradigm by incorporating additional data sources such as MRI, molecular biomarkers, or electronic health records, thereby situating breast cancer diagnosis within a true precision medicine framework. Integration with federated learning approaches may further enable cross-institutional training while preserving patient privacy. Finally, prospective clinical trials are required to evaluate whether AI-assisted multimodal frameworks improve patient outcomes, reduce time to diagnosis, or lower unnecessary biopsy rates compared to standard practice.

7. ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of the template.

8. REFERENCES

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209–49.

[2] Bray F, Jemal A, Grey N, Ferlay J, Forman D. Global cancer transitions according to the Human Development Index (2008–2030): a population-based study. Lancet Oncol. 2012;13(8):790–801.

[3] DeSantis CE, Ma J, Goding Sauer A, Newman LA, Jemal A. Breast cancer statistics, 2017, racial disparity in mortality by state. CA Cancer J Clin. 2017;67(6):439–48.

[4] Tabár L, Vitak B, Chen TH, Yen AM, Cohen A, Tot T, et al. Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. Radiology. 2011;260(3):658–63.

[5] Myers ER, Moorman P, Gierisch JM, Havrilesky LJ, Grimm LJ, Ghate S, et al. Benefits and harms of breast cancer screening: a systematic review. JAMA. 2015;314(15):1615–34.

[6] Mandelson MT, Oestreicher N, Porter PL, White D, Finder CA, Taplin SH, et al. Breast density as a predictor of mammographic detection: comparison of interval- and screen-detected cancers. J Natl Cancer Inst. 2000;92(13):1081–7.

[7] Nelson HD, Pappas M, Cantor A, Griffin J, Daeges M, Humphrey L. Harms of breast cancer screening: systematic review to update the 2009 U.S. Preventive Services Task Force recommendation. Ann Intern Med. 2016;164(4):256–67.

[8] Berg WA, Blume JD, Cormack JB, Mendelson EB, Lehrer D, Böhm-Vélez M, et al. Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer. JAMA. 2008;299(18):2151–63.

[9] Kelly KM, Dean J, Comulada WS, Lee SJ. Breast cancer detection using automated whole-breast ultrasound and mammography in radiographically dense breasts. Eur Radiol. 2010;20(3):734–42.

[10] Moon WK, Lo CM, Cho N, Chang JM, Chen JH, Chen PC, et al. Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. Comput Methods Programs Biomed. 2020;190:105361.

[11] Stavros AT, Thickman D, Rapp CL, Dennis MA, Parker SH, Sisney GA. Solid breast nodules: use of sonography to distinguish between benign and malignant lesions. Radiology. 1995;196(1):123–34.

[12] Lee RS, Gimenez F, Hoogi A, Rubin D. Curated Breast Imaging Subset of DDSM (CBIS-DDSM). Cancer Imaging Arch. 2016. doi:10.7937/K9/TCIA.2016.7O02S9CY.

[13] Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. Data Brief. 2020;28:104863.

[14] Spanhol FA, Oliveira LS, Petitjean C, Heutte L. A dataset for breast cancer histopathological image classification. IEEE Trans Biomed Eng. 2016;63(7):1455–62.

[15] Bejnordi BE, Veta M, Van Diest PJ, Van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA. 2017;318(22):2199–210.

[16] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.

[17] Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep learning to improve breast cancer detection on screening mammography. Sci Rep. 2019;9:12495.

[18] Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS. INbreast: toward a full-field digital mammographic database. Acad Radiol. 2012;19(2):236–48.

[19] Yap MH, Pons G, Marti R, Ganau S, Sentis M, Zwiggelaar R, et al. Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J Biomed Health Inform. 2018;22(4):1218–26.

[20] Araújo T, Aresta G, Castro E, Rouco J, Aguiar P, Eloy C, et al. Classification of breast cancer histology images using convolutional neural networks. PLoS One. 2017;12(6):e0177544.

[21] Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med. 2019;25(8):1301–9.

[22] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. ICLR. 2021.

[23] Ilse M, Tomczak JM, Welling M. Attention-based deep multiple instance learning. ICML. 2018:2127–36.

[24] Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput Med Imaging Graph. 2007;31(4–5):198–211.

[25] Huang S, Yang J, Fong S, Zhao Q. Artificial intelligence in cancer diagnosis and prognosis: opportunities and challenges. Cancer Lett. 2020;471:61–71.

[26] Xu H, Mo T, Feng Q, Zhong P, Lai M, Chang EI, et al. Deep learning of feature representation with multiple instance learning for medical image analysis. ICIP. 2014:1624–8.

[27] 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 Med. 2018;15(11):e1002683.

[28] Johnson J, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data. 2019;6:27.

[29] Courtiol P, Tramel EW, Sanselme M, Wainrib G. Classification and disease localization in histopathology using only global labels: a weakly-supervised approach. Med Image Anal. 2018;51:183–97.

[30] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. ICCV. 2017:618–26.

[31] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765–74.

[32] Pizer SM, Amburn EP, Austin JD, Cromartie R, Geselowitz A, Greer T, et al. Adaptive histogram equalization and its variations. Comput Vision Graph Image Process. 1987;39(3):355–68.

[33] Yu Y, Acton ST. Speckle reducing anisotropic diffusion. IEEE Trans Image Process. 2002;11(11):1260–70.

[34] Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan X, et al. A method for normalizing histology slides for quantitative analysis. ISBI. 2009:1107–10.

[35] Frid-Adar M, Klang E, Amitai M, Goldberger J, Greenspan H. Synthetic data augmentation using GAN for improved liver lesion classification. ISBI. 2018:289–93.

[36] Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. ICML. 2019:6105–14.

[37] Tsai YH, Bai S, Yamada M, Morency LP, Salakhutdinov R. Multimodal transformer for unaligned multimodal language sequences. ACL. 2019.

Table 2. Diagnostic performance of single-modality models.

Modality	**Model**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F1-score**	**AUC-ROC**	**Cohen's κ**
Mammography	EfficientNet-B4	87.1%	85.9%	88.2%	86.4%	86.1%	0.89	0.74
Ultrasound	Swin Transformer	85.6%	84.1%	87.0%	84.8%	84.4%	0.87	0.72
Histopathology	MIL-AttentionNet	89.3%	88.0%	90.2%	88.7%	88.3%	0.91	0.78

Table 3. Comparison of unimodal vs multimodal model performance.

Model	**Accuracy**		**Sensitivity**	**Specificity**	**Precision**	**F1-score**	**AUC-ROC**	**Brier Score**
Mammography	87.1%		85.9%	88.2%	86.4%	86.1%	0.89	0.092
Ultrasound	85.6%		84.1%	87.0%	84.8%	84.4%	0.87	0.098
Histopathology	89.3%		88.0%	90.2%	88.7%	88.3%	0.91	0.087
Multimodal Fusion	**92.1%**		**92.1%**	**90.7%**	**91.2%**	**91.6%**	**0.96**	**0.051**

Table 4. External validation performance.

Training Dataset	**Test Dataset**	**Modality**	**Accuracy**	**AUC-ROC**	**Calibration Error**
CBIS-DDSM	INbreast	Mammography	83.9%	0.86	0.112
BUSI	BUD	Ultrasound	82.4%	0.84	0.119
BreakHis	Camelyon16	Histopathology	87.0%	0.90	0.094
Fusion (All)	**Mixed Test Sets**	**Multimodal**	**91.5%**	**0.93**	**0.058**