Multimodal Deep Learning for Phyllodes Tumor Classification from Ultrasound and Clinical Data

Farhan Fuad Abir[1], Abigail Elliott Daly[3], Kyle Anderman[3], Tolga Ozmen[3], and Laura J. Brattain[2]

[1]Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL, USA
[2]Department of Medicine, University of Central Florida College of Medicine, Orlando, FL, USA
[3]Massachusetts General Hospital, Department of Surgery, Section of Breast Surgery, MA, USA

***Abstract*—Phyllodes tumors (PTs) are rare fibroepithelial breast lesions that can be malignant but are difficult to classify preoperatively due to their radiological similarity to benign fibroadenomas. This often leads to unnecessary surgical excisions. To address this, we propose a multimodal deep learning framework that integrates breast ultrasound (BUS) images with structured clinical data to improve diagnostic accuracy. We developed a dual-branch neural network that extracts and fuses features from ultrasound images and patient metadata from 81 subjects with confirmed PTs. Class-aware sampling and subject-stratified 5-fold cross-validation were applied to mitigate class imbalance and data leakage. The results show that our proposed multimodal method outperforms unimodal baselines in classifying benign versus borderline/malignant PTs. Among six deep learning-based image encoders, ConvNeXt and ResNet18 achieved the best performance in the multimodal setting, with AUC-ROC scores of 0.9427 and 0.9349, and F1-scores of 0.6720 and 0.7294, respectively. This study demonstrates the potential of multimodal AI to serve as a non-invasive diagnostic tool, reducing unnecessary surgical excisions and improving clinical decision-making in breast cancer care.**

***Index Terms*—phyllodes tumor, breast ultrasound, multimodal AI, feature fusion, class-aware sampling**

I. Introduction

Breast fibroepithelial lesions are a diverse group of biphasic tumors consisting of fibroadenomas (FAs) and phyllodes tumors (PTs). While FAs are common benign tumors with a low risk of recurrence or metastasis, PTs are rare lesions, accounting for less than 1% of breast tumors, that exhibit heterogeneous characteristics [1]. However, their clinical significance lies in the uncertain progression and the diagnostic challenges [2]. Moreover, the World Health Organization (WHO) classifies phyllodes tumors into benign, borderline, and malignant subtypes with reported local recurrence rates of 7.1%, 16.7%, and 25.1%, respectively [3]. Therefore, accurate preoperative grading of PTs into benign versus borderline/malignant is crucial for appropriate treatment planning.

The current standard diagnostic pathway often involves a core needle biopsy (CNB), followed by surgical excision if the pathology is ambiguous. However, only up to 57% of CNB-diagnosed fibroepithelial lesions are later upgraded to phyllodes tumors after excision, leading to unnecessary surgeries with potential complications, scarring, and increased healthcare costs. In the United States alone, excisional biopsies for suspected phyllodes tumors cost over $2 million annually [1], [4]. Artificial Intelligence (AI) could potentially mitigate this challenge by improving the differentiation of malignant PTs from the benign ones.

Early efforts focused on ultrasound (US) imaging alone. Yan et al. developed a deep learning (DL) model differentiating PTs from fibroadenomas (FAs) with 87.3% accuracy using multicenter ultrasound data [3]. Similarly, the Xception architecture achieved an area under the curve (AUC) of 0.87 for PT-FA discrimination [5]. However, these models lacked clinical context, limiting their diagnostic utility in ambiguous cases [6].

Multimodal AI approaches often integrate imaging with clinical data, and they show significant promise in broader breast cancer diagnostics. Ben et al. incorporated patient metadata with mammography, a tissue characterization technique, which improved molecular subtyping AUC by up to 27.6% over image-only models [7]. Similarly, fusing B-mode ultrasound with Nakagami parametric imaging increased AUC by 10.75% [8], and combining mammography and ultrasound boosted specificity to 96.41% [9]. Xu et al. integrated morphological and texture features that reduced false positives in BI-RADS 4 lesions by 15% [10]. Despite the progress, multimodal strategies have rarely been applied to phyllodes tumors, where models continue to rely on ultrasound or histopathological features alone. This is a critical gap, as clinical data, such as age, tumor size, and menopausal status, are central to the overall evaluation and could significantly enhance diagnostic performance in AI models.

In this work, we propose a multimodal deep learning framework integrating US images and structured clinical data for PT classification. Our dual-branch architecture fuses image and tabular features, while class-aware sampling and subject-stratified 5-fold cross-validation address class imbalance and data leakage. We also evaluated the method on multiple deep learning-based image encoders, all of which outperformed unimodal baselines, highlighting its potential as a non-invasive decision support tool in clinical workflows.

II. Methods

A. Dataset

The phyllodes tumor dataset was retrospectively collected and anonymized at the Massachusetts General Hospital (MGH) and was approved by the Institutional Review Board.

The deidentified dataset comprises clinical data from 106 subjects, breast ultrasound (BUS) images from 71, and BUS videos from 10 subjects in DICOM format. Not all subjects had complete clinical and imaging data. After data matching, a subset of 81 subjects (an average age of 38.01 $\pm$ 13.37 years) with both modalities was selected for this study. In terms of phyllodes tumor class, there were 65 benign, 10 borderline, and 6 malignant subjects. For this study, we combined the borderline and malignant cases to formulate a binary classification task.

B. Preprocessing

The BUS images and videos, originally in DICOM format, were converted to PNG format. Duplicate images were removed, and only frames containing tumors were retained. Since most BUS images included text along the image borders (e.g., probe position, frequency, and other metadata), we cropped the borders to remove these artifacts. Additionally, the extracted images were in RGB, and we converted them to grayscale to ensure a consistent color profile across the dataset. The resulting image dataset consisted of 1,213 benign and 425 malignant images, resulting in a class ratio of approximately 3:1.

The clinical dataset was then preprocessed based on feature availability, retaining three numeric (age at diagnosis, BMI, and tumor size) and three categorical features (race, menopausal status, and tumor echogenicity). Categorical variables were processed using one-hot encoding, yielding a total of 10 clinical features.

C. Class-aware sampling

In highly imbalanced datasets, conventional random sampling often results in mini-batches dominated by the majority class, which can lead to biased model updates and poor generalization to minority classes. Class-aware sampling addresses this issue by decoupling the sampling process into two stages. First, a class is selected uniformly at random, ensuring equal sampling probability for each class regardless of its prevalence in the dataset. Second, a sample is drawn uniformly from within the selected class. By repeating this process, the constructed mini-batches contain a more equitable distribution of classes over time [11]. This approach has effectively improved the training dynamics of deep learning models on long-tailed and imbalanced datasets [12].

D. Multimodal Deep Learning Framework

Our multimodal architecture consists of two parallel branches as illustrated in Figure 1.

The imaging branch is a convolutional neural network (CNN)-based image encoder to extract one-dimensional feature embeddings from grayscale ultrasound images. Transfer learning was used during training to finetune the ImageNet-pretrained image encoders. In parallel, the clinical data (10-dimensional input) was processed by a two-layer multilayer perceptron (MLP) with 10$\rightarrow$16$\rightarrow$8 units and ReLU activations to generate a compact embedding of the tabular features. The two embeddings were then concatenated at the feature level. Lastly, the fused vector was passed through a final linear classifier with a softmax activation function to predict the class label.

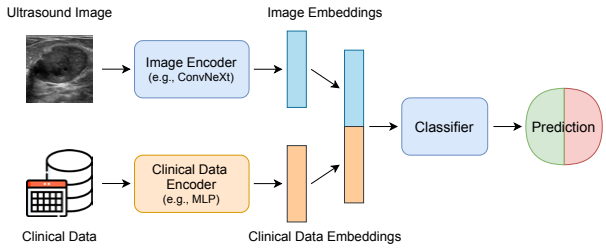

Fig. 1. Multimodal AI model architecture with feature-level fusion. The image branch consists of an image encoder (e.g., VGG16, ConvNeXt), while the clinical data branch uses an MLP. The resulting embeddings are concatenated and passed to a deep learning classifier (fully-connected layer) for the final prediction.

This design enables the model to jointly reason over visual and tabular inputs, enhancing robustness and interpretability.

III. Results

A. Experimental Setup

We performed subject-stratified 5-fold cross-validation to prevent data leakage across training and validation sets. Moreover, class-aware sampling was applied during training to address class imbalance. Ultrasound images were resized to 224$\times$224 pixels, and the intensity values were normalized to the [0, 1] range. During training, we applied random horizontal flipping and small rotation ($\pm 10°$). Models were trained for 30 epochs using stochastic gradient descent (SGD) with a learning rate of 0.001, momentum of 0.9, and a batch size of 4. Each model was evaluated using average accuracy, F1-score, and its 95% confidence interval (CI) across folds. All experiments were conducted using an NVIDIA RTX A4000 GPU (16GB VRAM) and a system with 32GB RAM.

B. Evaluation of multimodal image encoders

We evaluated six ImageNet-pretrained image encoders: VGG16 [13], ResNet18, ResNet50 [14], EfficientNet-B7 [15], CCT [16], and ConvNeXt [17] within the multimodal architecture. Table I summarizes the average accuracy, F1-score, and AUC-ROC over five subject-stratified cross-validation folds.

TABLE I
EVALUATION OF THE MULTIMODAL FRAMEWORK WITH DIFFERENT IMAGE ENCODERS (SORTED BY AUC-ROC)

Image Encoder	Mean Accuracy	Mean F1-Score	Mean AUC	95% CI (AUC)
ConvNeXt	0.9098	0.6720	**0.9427**	**[0.8842, 1.0000]**
CCT	0.9018	0.6683	0.9406	[0.8633, 1.0000]
ResNet18	**0.9179**	**0.7294**	0.9349	[0.8775, 0.9923]
ResNet50	0.9100	0.7065	0.9168	[0.8253, 1.0000]
EfficientNet-B7	0.8252	0.5473	0.8963	[0.8203, 0.9722]
VGG16	0.8885	0.6389	0.8886	[0.7376, 1.0000]

Among all models, **ConvNeXt** achieved the highest AUC-ROC of 0.9427, closely followed by CCT (0.9406) and ResNet18 (0.9349). In addition, ConvNeXt achieved a sensitivity of 0.71, a specificity of 0.95, a PPV of 0.87, and an NPV of 0.81. **ResNet18** obtained the best average accuracy of 0.9179 and an F1-score of 0.7294. In contrast, **EfficientNet-B7** and **VGG16** exhibited lower performance, particularly in F1-score (0.5473 and 0.6389, respectively), likely reflecting their sensitivity to class imbalance or limited generalization on small datasets.

C. Attribution analysis

To improve explainability, we used Score-CAM to visualize salient regions in ultrasound images, highlighting the model's decision process. Score-CAM generates class-specific attribution maps by weighting activation maps with their corresponding class scores [18]. Figure 2 shows examples of true positive, true negative, and misclassified cases using ConvNeXt as the image encoder.

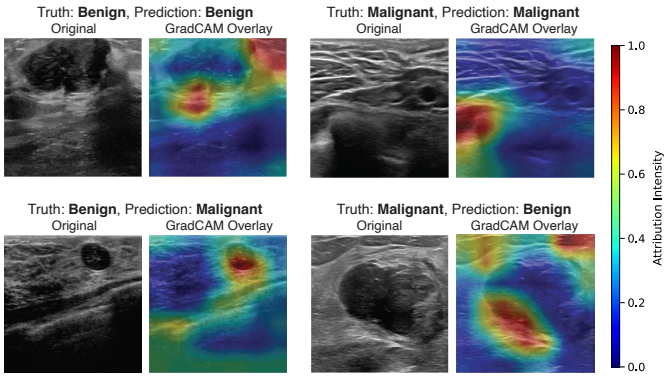

Fig. 2. Score-CAM attributes of the ConvNeXt image encoder benign and malignant cases. Warmer colors indicate regions contributing more strongly to the predicted class. Correctly predicted cases show focused attention on relevant tumor regions, whereas misclassified cases often exhibit diffuse or misplaced activations, suggesting visual ambiguity or model uncertainty.

In the correctly predicted malignant cases, the image encoder focused on tumor regions and the surroundings with heterogeneous echotexture, while the misclassified samples showed dispersed or ambiguous activation patterns. These insights suggest that visual cues alone are sometimes insufficient, which further justifies the use of multimodal reasoning.

D. Modality Ablation

We adopted a drop-based ablation approach to quantify the relative importance of each modality. For each sample, we computed the model's predicted probability for the malignant class using all validation data across 5 folds. We repeated the process after zeroing out either the image or the clinical features. The magnitude of the change in probability is treated as the modality's contribution. We normalized these changes to obtain relative contribution scores per sample. Figure 3 illustrates that BUS contributes 63% $\pm$ 0.09 in the class prediction and clinical data contributes to 37% $\pm$ 0.09. The box plot boundaries represent the 25th and 75th percentiles.

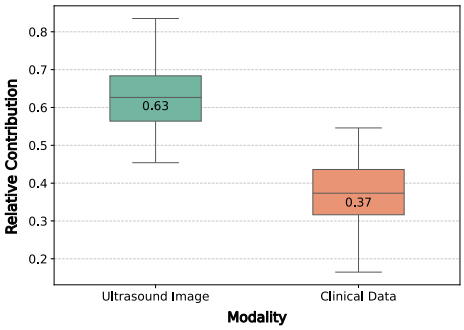

Fig. 3. Modality ablation analysis showing relative contributions of ultrasound images and clinical data in the multimodal ConvNeXt model. The box boundaries represent the interquartile range (IQR). Ultrasound imaging contributed more strongly (mean 0.63 $\pm$ 0.09) than clinical features (mean 0.37 $\pm$ 0.09), while both provided complementary information for classification.

E. Performance comparison between unimodal and multimodal approaches

We investigated the contribution of each modality by evaluating unimodal and multimodal models using our best-performing image encoder (ConvNeXt) and a multilayer perceptron (MLP) for clinical data. Table II summarizes the classification performance across different input configurations. The multimodal model performed best, with an accuracy of 0.9098, F1-score of 0.6720, and AUC of 0.9427, demonstrating the benefit of combining data sources. The image-only ConvNeXt also performed well (AUC 0.8919), while the clinical-only MLP was much weaker with a F1-score of 0.33, showing that clinical features alone have limited discriminative power. Overall, these findings confirm the value of multimodal fusion for improving diagnostic accuracy and robustness.

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT INPUT MODALITIES

Modality	Accuracy	F1-Score	AUC	95% CI (AUC)
Only clinical data	0.8140	0.3300	0.7846	[0.6561, 0.9132]
Only BUS image	0.8906	0.6138	0.8919	[0.8259, 0.9578]
Multimodal	**0.9098**	**0.6720**	**0.9427**	**[0.8842, 1.0000]**

Figure 4 shows ROC curves of the two best-performing models, namely ConvNeXt and CCT, in both unimodal and multimodal configurations. The multimodal ConvNeXt model achieved the highest AUC of 0.9427 (95% CI: [0.8842, 1.0000]), followed closely by multimodal CCT with an AUC of 0.9406 (95% CI: [0.8633, 1.0000]), both of which surpassed their unimodal counterparts. At the EER point, multimodal ConvNeXt achieved the highest score (0.8799), followed by multimodal CCT (0.8519), both surpassing unimodal ConvNeXt (0.8348) and CCT (0.8098). This reiterates the value of multimodal fusion, especially in cases where image features alone may be insufficient for confident classification.

IV. DISCUSSION

While unimodal models based solely on ultrasound imaging achieved reasonably strong performance, the multimodal ap-

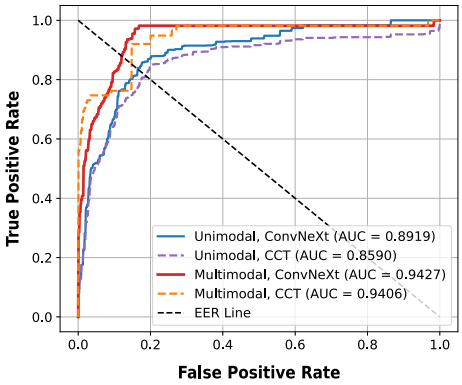

Fig. 4. Receiver operating characteristic (ROC) curves of ConvNeXt and CCT in unimodal and multimodal settings. Multimodal models outperformed unimodal counterparts (ConvNeXt: 0.9427 vs. 0.8919; CCT: 0.9406 vs. 0.8590), confirming the benefit of integrating clinical data. The multimodal ConvNeXt achieved the highest AUC (0.9427) and EER (0.8799), followed by multimodal CCT (AUC 0.9406, EER 0.8519), with unimodal ConvNeXt (0.8348) and CCT (0.8098) performing lower.

proach consistently outperformed them across all key metrics. Although the modality ablation study indicates a greater contribution from BUS images, the clinical data still plays a significant complementary role. This aligns with clinical practice, where both imaging and patient history are considered.

Moreover, Score-CAM visualization provides a foundation for clinical interpretability. Misclassified cases tended to exhibit weaker or diffused attention maps, indicating that the image encoder struggled with ambiguous visual patterns, which further justifies the use of auxiliary modalities.

Despite promising results, this study has a few limitations. The phyllodes dataset is relatively small for malignant and borderline cases, which limits the statistical strength of the findings. Larger and more diverse multi-center datasets are needed before the clinical trial to further refine and test the AI models. Overall, this work shows that multimodal deep learning frameworks can improve phyllodes tumor diagnosis, even when working with small and imbalanced datasets which is a common challenge in rare diseases.

V. Conclusion

We proposed a multimodal deep learning framework that integrates ultrasound image features with structured clinical data for phyllodes tumor classification. The multimodal architecture consistently outperformed unimodal baselines across multiple image encoders, improving diagnostic accuracy and robustness while potentially reducing unnecessary surgical biopsies. A key challenge was the strong class imbalance, with relatively few borderline and malignant cases, which we mitigated using class-aware sampling and subject-stratified cross-validation.

Future work will aim to expand the dataset, incorporate radiologist annotations, and explore transformer-based encoders with attention mechanisms to enhance interpretability and performance. Prospective validation in real-world clinical settings will also be critical to assess the system's generalizability and its utility in guiding biopsy decisions.

References

[1] T. Soleimani, D. Euhus, O. Sogunro, L. Cope, M. Janjua, M. Vasigh, and L. K. Jacobs, "De-escalating indications for excision when breast core needle biopsy returns fibroepithelial lesion—not further characterized," *Breast Cancer Research and Treatment*, vol. 207, no. 3, pp. 561–568, 2024.

[2] G. Lissidini, A. Mulè, A. Santoro, G. Papa, L. Nicosia, E. Cassano, A. A. Ashoor, P. Veronesi, L. Pantanowitz, J. L. Hornick *et al.*, "Malignant phyllodes tumor of the breast: a systematic review," *Pathologica*, vol. 114, no. 2, p. 111, 2022.

[3] Y. Yan, Y. Liu, Y. Wang, T. Jiang, J. Xie, Y. Zhou, X. Liu, M. Yan, Q. Zheng, H. Xu *et al.*, "Hierarchical diagnosis of breast phyllodes tumors enabled by deep learning of ultrasound images: a retrospective multi-center study," *Cancer Imaging*, vol. 25, no. 1, pp. 1–13, 2025.

[4] D. Mousa-Doust, C. K. Dingee, L. Chen, A. Bazzarelli, U. Kuusk, J.-S. Pao, R. Warburton, and E. C. McKevitt, "Excision of breast fibroepithelial lesions: when is it still necessary?—a 10-year review of a regional centre," *Breast Cancer Research and Treatment*, vol. 194, no. 2, pp. 307–314, 2022.

[5] Y. Yan, Y. Liu, J. Yao, L. Sui, C. Chen, T. Jiang, X. Liu, Y. Wang, D. Ou, J. Chen *et al.*, "Deep learning-assisted distinguishing breast phyllodes tumours from fibroadenomas based on ultrasound images: a diagnostic study," *British Journal of Radiology*, vol. 97, no. 1163, pp. 1816–1825, 2024.

[6] A. Pfob, C. Sidey-Gibbons, R. G. Barr, V. Duda, Z. Alwafai, C. Balleyguier, D.-A. Clevert, S. Fastner, C. Gomez, M. Goncalo *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*, vol. 32, no. 6, pp. 4101–4115, 2022.

[7] C. Ben Rabah, A. Sattar, A. Ibrahim, and A. Serag, "A multimodal deep learning model for the classification of breast cancer subtypes," *Diagnostics*, vol. 15, no. 8, p. 995, 2025.

[8] S. Muhtadi and C. M. Gallippi, "Breast tumor diagnosis via multimodal deep learning using ultrasound b-mode and nakagami images," *Journal of Medical Imaging*, vol. 12, no. S2, pp. S22 009–S22 009, 2025.

[9] J. Chen, T. Pan, Z. Zhu, L. Liu, N. Zhao, X. Feng, W. Zhang, Y. Wu, C. Cai, X. Luo *et al.*, "A deep learning-based multimodal medical imaging model for breast cancer screening," *Scientific Reports*, vol. 15, no. 1, p. 14696, 2025.

[10] P. Xu, J. Zhao, M. Wan, Q. Song, Q. Su, and D. Wang, "Classification of multi-feature fusion ultrasound images of breast tumor within category 4 using convolutional neural networks," *Medical Physics*, vol. 51, no. 6, pp. 4243–4257, 2024.

[11] L. Shen, Z. Lin, and Q. Huang, "Relay backpropagation for effective learning of deep convolutional neural networks," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*. Springer, 2016, pp. 467–482.

[12] M. Varaganti, A. Vankayalapati, N. Awad, G. R. Dion, and L. J. Brattain, "T2id-cas: Diffusion model and class aware sampling to mitigate class imbalance in neck ultrasound anatomical landmark detection," *arXiv preprint arXiv:2504.21231*, 2025.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[16] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," *https://arxiv.org/abs/2104.05704*, 2021.

[17] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

[18] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 24–25.