Breast Ultrasound Tumor Classification with Deep Learning

Ceyda Babuz, Muhammed Dilber
Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi, 34220 Istanbul, Türkiye
{ceyda.babuz, muhammed.dilber}@std.yildiz.edu.tr

***Özetçe* —Bu çalışma, meme ultrason görüntülerinin (BUSI ve BUS-UCLM) sınıflandırılmasında geleneksel CNN mimarileri ile modern Vision Transformer (ViT) modellerinin performansını incelemektedir. Araştırma kapsamında, transfer öğrenme ve alan uyarlaması teknikleri kullanılarak 11 farklı derin öğrenme modeli eğitilmiştir. Deneysel sonuçlar, Vision Transformer modellerinin tıbbi görüntülerin karmaşık doku yapılarını anlamlandırmada CNN modellerine göre daha başarılı olduğunu ortaya koymuştur. Özellikle CaiT modeli, %91,30 doğruluk ve %93,10 kesinlik oranlarıyla en yüksek performansı sergilemiş olup, bu tür modellerin klinik karar verme süreçlerinde yüksek güvenilirlikli bir kaynak olabileceği gösterilmiştir.**

Anahtar Kelimeler*—*Meme Kanseri, Ultrason Sınıflandırma, Vision Transformer, CaiT, Derin Öğrenme,Evrişimli Sinir Ağları.

***Abstract*—This study investigates the performance of traditional Convolutional Neural Network (CNN) architectures versus modern Vision Transformer (ViT) models in the classification of breast ultrasound images (BUSI and BUS-UCLM). Within the scope of the research, 11 different deep learning models were trained using transfer learning and domain adaptation techniques. Experimental results demonstrate that Vision Transformer models are more successful than CNN models in interpreting the complex tissue structures of medical images. In particular, the CaiT model exhibited the highest performance with an accuracy rate of 91.30% and a precision rate of 93.10%, indicating that such models can serve as a highly reliable resource in clinical decision-making processes.**

Keywords*—*Breast Cancer, Breast Ultrasound Classification, Vision Transformer, CaiT, Deep Learning, Convolutional Neural Networks.

I. Introduction

Breast cancer is the most frequently diagnosed cancer among women and remains a major global health challenge. Early detection is critical for survival, yet ultrasound imaging often suffers from low contrast and speckle noise. These factors complicate manual diagnosis and increase the need for computer-aided classification systems [1]. Convolutional Neural Networks (CNNs) have long been the standard for this task due to their ability to extract local hierarchical patterns. Specifically, architectures such as Inception and EfficientNet have demonstrated significant success in medical imaging by utilizing deep feature extraction capabilities[2].

Despite their success, CNNs struggle to capture the global context and long-range pixel dependencies within an image. To address these limitations, Vision Transformer (ViT) models have introduced attention mechanisms that evaluate both local and global features simultaneously [3]. Advanced variations of these models have shown great promise in medical tasks where data-efficient learning is required. Furthermore, architectures like CaiT enhance the interaction between classification tokens and image patches, offering high potential for distinguishing complex textures in ultrasound scans. These modern approaches aim to provide more robust features compared to traditional methods.

This study provides a comprehensive comparative analysis of traditional CNNs and modern Vision Transformers for breast ultrasound classification. Using the BUSI and BUS-UCLM datasets, transfer learning and domain adaptation strategies are implemented to evaluate model robustness. The experimental results demonstrate that Transformer-based models, specifically the CaiT architecture, outperform traditional CNNs in interpreting complex tumor structures. This research contributes to the field by identifying reliable architectural strategies for medical classification tasks involving limited datasets.

II. Related Work

Recent literature primarily focuses on binary classification of breast lesions into benign and malignant categories. Studies utilize a wide range of datasets, varying from large-scale repositories to more constrained image sets, where data augmentation—such as multi-angle rotations and scaling—is systematically applied to enhance model generalization. Over the last five years, deep learning has become the dominant paradigm, evolving from traditional CNN architectures like VGG, and ResNet to contemporary transformer-based models. While early studies often rely on direct classification or ROI-based cropping via segmentation, recent research trends have moved toward more advanced methodologies. This includes the optimization of architectures like DenseNet121 and Xception , alongside the implementation of advanced methodologies to enhance diagnostic precision. In addition, recent advancements integrate transfer learning ensembles—merging features from ResNet-50 and InceptionV3 —and utilize modern Vision Transformer architectures like CaiT. These models incorporate internal attention mechanisms to ensure they prioritize critical morphological details, ultimately providing

superior performance in interpreting complex medical tissue structures.

Several researchers have moved beyond simple classification by integrating segmentation and attention mechanisms to focus on specific Regions of Interest. Some of them combined different imaging modalities or extracted deep features from multiple architectures with some optimization strategies. The reviewed articles are summarized as follows:

- Hossain et al. [4] introduced a cascaded hybrid system using a U-Net-based model for precise tumor segmentation, followed by a fine-tuned CNN for classification to minimize background noise interference.
- He et al. [5] advanced this further with a multi-task learning approach , which simultaneously optimizes segmentation and classification through a "Gate Unit" and deformable spatial attention to handle morphological variations.
- Cruz-Ramos et al.[1] developed a hybrid system merging ultrasound (BUSI) and mammography (mini-DDSM) data, employing hand-crafted features (HOG, LBP) alongside DenseNet and optimizing the fusion via genetic algorithms to reach 97.6
- Atrey et al. [6] similarly utilized multi-modality fusion by combining ResNet-18 deep features from both ultrasound and mammograms, followed by an SVM classifier to achieve 99.2% accuracy.
- Caliskan et al. [7] focused on model ensembles, aggregating deep features from VGG, ResNet, and Inception into a single vector, which leveraged complementary structural information to achieve a 99.57% success rate.
- Jabeen [8] et al. implemented a four-stage framework using VGG19, ResNet50, and InceptionV3, notably incorporating probabilistic selection to identify the most discriminative features from the fused vector before final classification.
- Liu et al. [9] integrated attention mechanisms directly into CNN blocks to improve the model's focus on critical morphological details like tumor boundaries.
- Kormpos et al. [10] conducted a comprehensive benchmark, concluding that while lightweight models like MobileNetV2 are optimal for efficiency, DenseNet and ResNet remain the preferred choices for maximum diagnostic precision.

III. Materials and Methods

The proposed methodology for breast ultrasound classification follows a structured pipeline, integrating advanced data preprocessing, multi-architectural deep learning models, and domain adaptation strategies. The overall technical workflow of the study is illustrated in Figure 1, providing a high-level overview of the progression from data preparation to final performance evaluation.

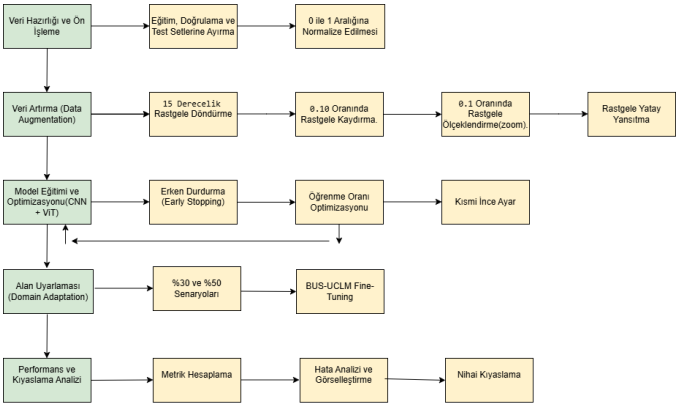

Figure 1 Overall technical workflow of the proposed system including data preprocessing, augmentation, model training (CNN and ViT), and domain adaptation phases.

A. Dataset Description and Augmentation

In this study, two primary datasets were used. These are BUSI (Breast Ultrasound Images) and BUS-UCLM. The BUSI dataset contains 780 images. These images are categorized as Benign, Malignant, and Normal. A multi-stage data augmentation strategy was implemented to prevent model bias. This strategy included random rotation, scaling, and horizontal flipping. Through these methods, the dataset was expanded to 7,031 images.

The BUS-UCLM dataset includes 683 original images from 38 patients. This dataset was used to test the domain adaptation capabilities of the models. Data augmentation was applied during the training phase. The total number of images in this set reached 2,170. Two different training scenarios were designed. These scenarios used 30% and 50% splits of the data. This approach allowed us to observe how models generalize knowledge across different sources.

B. Architectural Frameworks

The study evaluates two main categories of deep learning models for classification. The first group consists of six different Convolutional Neural Networks (CNNs). ResNet-50 is tested, which uses skip connections to solve the vanishing gradient problem in deep networks [4]. InceptionV3 was included for its ability to use multiple filter sizes in the same layer to capture different image features [10]. DenseNet is also analyzed, where each layer is directly connected to every other layer to ensure maximum information flow [11]. EfficientNet was chosen for its balanced scaling of depth, width, and resolution [12]. Additionally, NASNet's automatically designed cells [12] and Xception's depthwise separable convolutions [13] were evaluated. These CNN architectures are highly effective at extracting local patterns and textures from medical images.

The second group of models focuses on Vision Transformers (ViTs), which represent a more modern approach in image analysis [3]. Unlike CNNs, these models use a self-attention mechanism to understand the relationship between different parts of an image. The input image is divided into small, fixed-size patches. These patches

are treated as a sequence of tokens, similar to words in a sentence. This structure allows the model to capture global context and long-range dependencies. By looking at the entire image at once, Transformers can identify complex pathological structures that CNNs might miss.

Several advanced variations of the Vision Transformer were analyzed in this research. DeiT was used for its data-efficient training strategy, which is ideal for smaller medical datasets. Swin Transformer was included for its hierarchical structure that uses shifted windows to compute attention. We also evaluated BEiT, which uses a masked image modeling approach for pre-training. Finally, the CaiT (Class-Attention in Image Transformers) model was specifically selected. This model optimizes the interaction between the image patches and the classification token. This feature is particularly useful for distinguishing subtle differences between benign and malignant tissues in ultrasound scans.

C. Transfer Learning and Domain Adaptation

Transfer learning was utilized to improve training efficiency and model accuracy on limited medical datasets. Instead of random initialization, models were initialized with pre-trained weights from the ImageNet dataset [14]. This allowed the architectures to leverage universal visual features like edges and textures before fine-tuning on ultrasound images. During this process, early layers were frozen to preserve general knowledge, while the final classification layers were updated to distinguish between benign, malignant, and normal tumor classes. This strategy effectively reduces the risk of overfitting and significantly shortens the training duration.

Domain adaptation was also implemented to ensure model robustness across different clinical imaging sources. Medical images often vary significantly depending on the device used and patient demographics. This process aimed to align the features of the source dataset (BUSI) with the target dataset (BUS-UCLM) to maintain high performance in new environments [11]. We designed specific training scenarios with 30% and 50% data splits to evaluate the model's generalization capability. By bridging the gap between different clinical sources, the study demonstrates a reliable methodological approach for real-world medical diagnostic applications.

IV. Experimental Results and Analysis

A. Performance Analysis

In this phase, the CNN architectures were initially trained and validated on the BUSI dataset across various learning rates to establish a performance baseline. Following hyperparameter optimization involving three distinct learning rate configurations, the models' efficacy on the test set was evaluated based on accuracy, F1-score, precision, and recall metrics, with the top-performing weights being preserved for further analysis. Subsequently, these optimized models underwent a secondary evaluation phase where they were fine-tuned on diverse datasets through a Domain Adaptation approach to assess their cross-institutional generalizability.

1) Experimental Results on BUSI Dataset: The performance of six different CNN architectures was evaluated on the BUSI dataset across three learning rates (10^{-3}, 10^{-4}, and 10^{-5}) using transfer learning111111111. The detailed analysis for each model is as follows:

EfficientNet demonstrated a balanced performance, peaking at a 10^{-4} learning rate with an accuracy of 86.96% and precision of 90.57%. Confusion matrix analysis reveals that the model correctly identified 27 out of 31 malignant cases at this rate, proving its high diagnostic reliability.

Table 1 Performance metrics of EfficientNet based on learning rate

Model/Output	Accuracy	Recall	Precision	F1-Score
EfficientNet (LR=10^{-3})	0.8609	0.8609	0.9000	0.88
EfficientNet (LR=10^{-4})	0.8696	0.8348	0.9057	0.8688
EfficientNet (LR=10^{-5})	0.7913	0.6696	0.8370	0.744

This model maintained a consistent accuracy of 87.83% across all tested learning rates. The 10^{-3} rate was identified as the optimum operating point due to its superior recall (86.96%), which is essential for minimizing false negatives in clinical breast cancer detection.

Table 2 Performance metrics of InceptionV3 based on learning rate

Model/Output	Accuracy	Recall	Precision	F1-Score
InceptionV3 (LR=10^{-3})	0,8783	0,8696	0,885	0,8772
InceptionV3 (LR=10^{-4})	0,8783	0,8174	0,9216	0,8664
InceptionV3 (LR=10^{-5})	0,8783	0,7478	0,9247	0,8269

For DenseNet, the 10^{-3} learning rate yielded the most significant results, achieving 83.48% accuracy and 85.15% precision. Performance notably declined at lower learning rates, likely due to underfitting or convergence issues during the 7,031-image training process.

Table 3 Performance metrics of DenseNet based on learning rate

Model/Output	Accuracy	Recall	Precision	F1-Score
DenseNet (LR=10^{-3})	0,8348	0,7892	0,8515	0,8192
DenseNet (LR=10^{-4})	0,8174	0,7617	0,8483	0,8027
DenseNet (LR=10^{-5})	0,7565	0,5914	0,7871	0,6754

ResNet-50 exhibited stable behavior, with its most reliable performance observed at the 10^{-5} learning rate, reaching an accuracy of 80.00%. This indicates that a more gradual weight update allows the model to better capture the complex morphological details within the medical imagery.

Table 4 Performance metrics of Resnet-50 based on learning rate

Model/Output	Accuracy	Recall	Precision	F1-Score
ResNet-50 (LR=10^{-3})	0,7826	0,7826	0,7843	0,7745
ResNet-50 (LR=10^{-4})	0,7417	0,7417	0,7461	0,7319
ResNet-50 (LR=10^{-5})	0,8	0,798	0,7991	0,7943

This architecture achieved its peak performance at 10^{-3} with 85.22% accuracy and a high recall of 87.79%. Its strength in identifying true positive cases makes it a valuable candidate for screening, although performance diminished as the learning rate decreased.

Table 5 Performance metrics of NasNetLarge based on learning rate

Model/Output	Accuracy	Recall	Precision	F1-Score
NasNetLarge (LR=10^{-3})	0,8522	0,8779	0,8380	0,8529
NasNetLarge (LR=10^{-4})	0,8348	0,8564	0,8154	0,8325
NasNetLarge (LR=10^{-5})	0,8	0,8111	0,7839	0,7296

Xception followed a linear trend where accuracy decreased alongside the learning rate, peaking at 80.00% with 10^{-3}. Despite variations in overall accuracy, the model maintained a respectable recall for malignant cases, highlighting its resilience in identifying critical pathologies.

Table 6 Performance metrics of Xception based on learning rate

Model/Output	Accuracy	Recall	Precision	F1-Score
Xception (LR=10^{-3})	0,8	0,8178	0,7683	0,787
Xception (LR=10^{-4})	0,7739	0,7889	0,7463	0,7581
Xception (LR=10^{-5})	0,7217	0,7637	0,7126	0,7173

2) Experimental Results on BUS-UCLM Dataset: The results of transfer learning experiments conducted on the BUS-UCLM dataset are presented to evaluate the models' performance on a secondary dataset and to compare the outcomes obtained with and without the application of domain adaptation approaches.

Table 7 Performance evaluation of CNN models on the BUS-UCLM dataset

Model	Accuracy	Recall	Precision	F1-Score	Learning Rate
EfficientNet	0.7788	0.7019	0.8202	0.7565	10^{-4}
DenseNet	0.7500	0.7212	0.8152	0.7653	10^{-3}
InceptionV3	0.8269	0.8077	0.8571	0.8317	10^{-3}
ResNet	0.7802	0.7800	0.7840	0.7345	10^{-5}
NasNet	0.7596	0.7596	0.7476	0.7489	10^{-3}
Xception	0.8462	0.8462	0.8474	0.8464	10^{-3}

An assessment of the BUS-UCLM dataset revealed that the Xception model demonstrated superior performance with an 84.62% accuracy rate, exhibiting a balanced and reliable capacity for differentiating between healthy and pathological tissues. InceptionV3 closely followed with an 82.69% success rate. A noteworthy observation was the ResNet architecture, which attained its optimal performance through a significantly more gradual and sensitive learning trajectory compared to its counterparts, highlighting that distinct structural frameworks require varying temporal dynamics for feature convergence. The distribution of the training data played a pivotal role in these outcomes. Although the overall performance on the BUS-UCLM dataset was slightly lower than that achieved on the BUSI dataset,the results remains robust and validate the models' cross-institutional generalizability.

3) Domain Adaptation Results: In this stage of the study, a domain adaptation approach was utilized to evaluate the models' ability to generalize across different data sources. Models previously trained on the BUSI dataset were adapted to the BUS-UCLM dataset, which served as the target domain. The primary goal was to observe how knowledge learned from BUSI images contributes to the classification of BUS-UCLM images compared to standard transfer learning. Experiments were conducted under two scenarios based on the target dataset utilization: a 30% training split and a 50% training split.

Table 8 Domain adaptation results with 30% target training data

Model	Accuracy	Recall	Precision	F1-Score
EfficientNet	0.8839	0.8839	0.8871	0.8845
ResNet-50	0.8433	0.8290	0.8315	0.8302
DenseNet	0.7825	0.7644	0.8042	0.7830
InceptionV3	0.7397	0.7166	0.7539	0.7348
Xception	0.7185	0.6954	0.7022	0.6988
NasNet	0.6763	0.6408	0.6477	0.6442

The results for the 30% training data scenario, presented in Table 8, evaluate the models' ability to adapt to a new domain with limited target samples. In this scenario, EfficientNet and ResNet-50 stood out as the most efficient architectures for feature transfer, achieving high accuracy rates despite the constrained data. This highlights their robustness in adapting pre-learned features to new clinical environments.

Table 9 Domain adaptation results with 50% target training data

Model	Accuracy	Recall	Precision	F1-Score
EfficientNet	0.9260	0.9310	0.8870	0.9030
ResNet-50	0.8675	0.8410	0.8745	0.8537
DenseNet	0.8387	0.8306	0.8502	0.8403
InceptionV3	0.7949	0.7811	0.8014	0.7911
Xception	0.7166	0.6980	0.7245	0.7092
NasNet	0.6993	0.6574	0.6799	0.6648

As shown in Table 9, the 50% training data scenario led to significant performance improvements across all

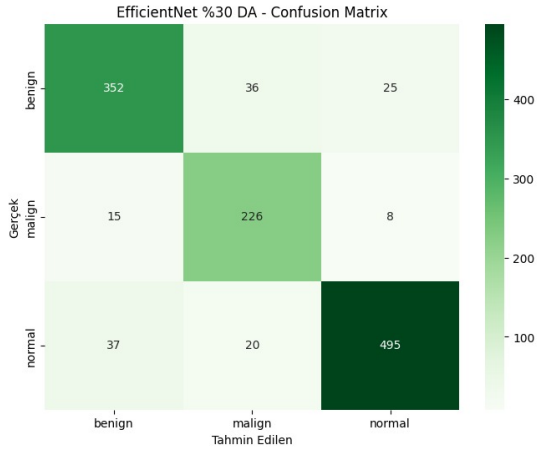

Figure 2 Confusion Matrix of EfficientNet Domain Adaptation with 50% Target Training Data

models. EfficientNet reached a peak accuracy of 92.60%, proving its superior generalization capacity when provided with slightly larger target datasets. Overall, the domain adaptation strategy outperformed traditional transfer learning, confirming that ultrasound-specific features learned from one institution remain highly relevant and transferable to different clinical datasets.

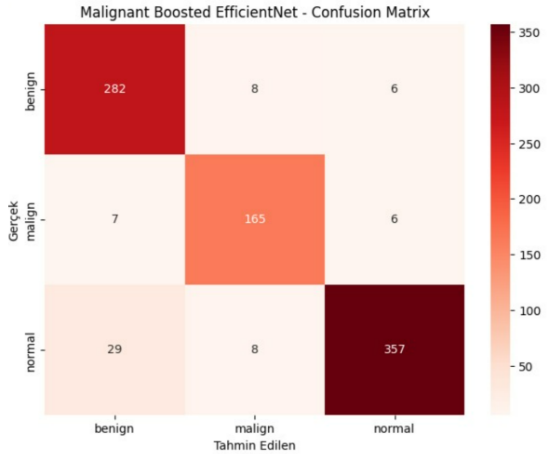

Figure 3 Confusion Matrix of EfficientNet Domain Adaptation with 50% Target Training Data

4) Performance of Transformer-Based Architectures: Vision Transformers (ViT) represent a significant shift from traditional CNNs by utilizing self-attention mechanisms originally designed for natural language processing. Unlike CNNs, which process images through local receptive fields, ViTs treat an image as a sequence of patches, allowing the model to capture global dependencies and long-range relationships between different regions of an ultrasound scan. In this section, the performance of the standard ViT model and four advanced variations (DeiT-Tiny, Swin Transformer, CaiT, and BEiT) were evaluated on the BUSI dataset to determine their effectiveness in breast cancer diagnosis.

Table 10 Performance comparison of Transformer-based models on BUSI dataset

Model	Accuracy	Recall	Precision	F1-Score
CaiT	**0.9130**	**0.8901**	**0.9310**	**0.9084**
BEiT	0.8957	0.8754	0.9266	0.8953
Swin Transformer	0.8870	0.8567	0.8947	0.8737
ViT (Standard)	0.8783	0.8685	0.8653	0.8655
DeiT-Tiny	0.8696	0.8465	0.8949	0.8677

As demonstrated in Table 10, Transformer-based architectures yielded highly successful results in detecting breast cancer. The CaiT model achieved the highest performance with an accuracy of 91.30% and a precision of 93.10%, highlighting its superior capability in distinguishing between healthy tissue and pathological structures. Furthermore, models such as BEiT and Swin Transformer also demonstrated robust outcomes, likely due to their ability to capture complex textural details and tissue boundaries within ultrasound images. Overall, the success of these architectures, particularly in surpassing the 90% accuracy threshold, suggests that the global context provided by self-attention mechanisms plays a crucial role in enhancing the reliability of medical diagnoses.

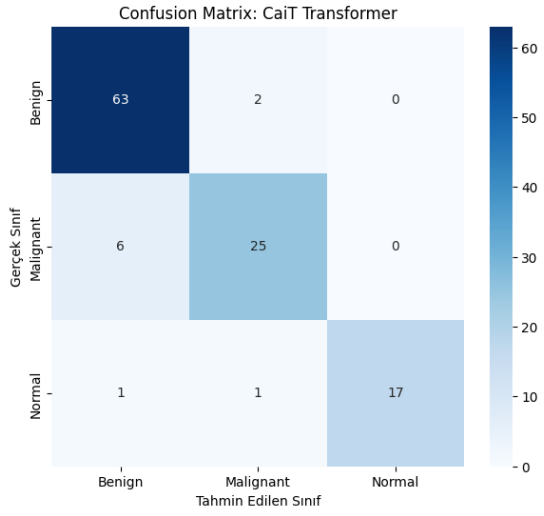

Figure 4 Confusion Matrix of CaiT Model

B. *Comparative Analysis*

The comparative results demonstrate that ResNet-50 achieved the greatest performance gains from the application of the domain adaptation strategy, increasing its accuracy from 78.02% in standard transfer learning to 84.33% with only 30% training data; this proves its ability to efficiently adapt medical tissue characteristics learned from the source domain. Similarly, EfficientNet demonstrated superior generalization by reaching a peak accuracy of 92.60% under the 50% adaptation scenario, significantly outperforming its baseline results. However, its slightly lower performance in the malignant category suggests that even with domain adaptation, data volume remains a critical factor for certain classes.

In contrast, models like Xception and NasNetLarge faced challenges during the adaptation process. Although Xception was the top performer in standard transfer learning (84.62%), its accuracy dropped to around 71% when adapted, indicating a misalignment between the features extracted from the source domain and the specific clinical attributes of the target dataset. DenseNet and InceptionV3 showed more stable transitions, with DenseNet proving the advantage of domain adaptation by exceeding its baseline 75.00% accuracy even in the restricted 30% data scenario. Overall, these findings confirm that while adaptation strategies enhance the performance of architectures like ResNet and EfficientNet by leveraging pre-learned ultrasound features, the complexity of models like NasNet requires more sensitive tuning to maintain cross-institutional reliability.

Beyond the CNN architectures, Transformer-based models demonstrated exceptional precision in distinguishing tumorous structures from healthy tissue. In particular, the CaiT model achieved a remarkable accuracy of 91.30% and a precision rate of 93.10% on the BUSI dataset, outperforming most standard CNN configurations. While models like BEiT and Swin Transformer also surpassed the 88% accuracy threshold, their success is attributed to the self-attention mechanism's ability to capture global dependencies across the ultrasound image. When compared to CNNs, these Transformer-based architectures offer a more holistic understanding of the tissue boundaries, providing a powerful alternative for clinical diagnostics where identifying subtle architectural distortions is critical.

V. Conclusion and Discussion

This study explores breast tumor classification using ultrasound imagery by benchmarking six CNN architectures against five Vision Transformer models through transfer learning and domain adaptation. Initial results on the BUSI dataset via ImageNet transfer learning identified InceptionV3 (87.83%) and EfficientNet (86.96%) as the strongest CNN performers, noted for their balanced feature extraction despite varying lesion types. However, Vision Transformers consistently outperformed CNNs across nearly all metrics. Especially, the CaiT model achieved the highest accuracy at 91.30%, demonstrating a superior capacity for identifying malignant lesions thanks to its robust self-attention mechanism. The research also highlights the critical impact of data distribution and training strategies. Through domain adaptation from BUSI to the BUS-UCLM dataset, most models—particularly ResNet—showed a significant performance boost, exceeding standard transfer learning results by approximately 10%. While some models struggled with limited training data (the 30% scenario), performance scaled effectively as data volume increased, underscoring the benefits of cross-domain knowledge transfer in medical imaging.

In conclusion, CaiT stands out for high-precision diagnostic tasks, while EfficientNet proves highly adaptable to diverse datasets. Despite limitations such as malignant class scarcity affecting CNN sensitivity, these findings suggest that such models can serve as vital decision-support tools for clinicians. Future work will investigate hybrid CNN-Transformer architectures and integrate multi-modal patient data—including age and genetics—to enhance generalization across larger, more balanced clinical datasets.

References

[1] C. Cruz-Ramos, O. García-Avila, J.-A. Almaraz-Damian, V. Ponomaryov, R. Reyes-Reyes, and S. Sadovnychiy, "Benign and malignant breast tumor classification in ultrasound and mammography images via fusion of deep learning and handcraft features," *Entropy*, vol. 25, no. 7, 2023. [Online]. Available: https://www.mdpi.com/1099-4300/25/7/991

[2] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, vol. 97, pp. 6105–6114, 2019. [Online]. Available: https://arxiv.org/abs/1905.11946

[3] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[4] S. Hossain, S. Azam, S. Montaha, A. Karim, S. S. Chowa, C. Mondol, M. Zahid Hasan, and M. Jonkman, "Automated breast tumor ultrasound image segmentation with hybrid unet and classification using fine-tuned cnn model," *Heliyon*, vol. 9, no. 11, p. e21369, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405844023085778

[5] Q. He, Q. Yang, H. Su, and Y. Wang, "Multi-task learning for segmentation and classification of breast tumors from ultrasound images," *Computers in Biology and Medicine*, vol. 173, p. 108319, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482524004037

[6] K. Atrey, B. K. Singh, and N. K. Bodhey, "Integration of ultrasound and mammogram for multimodal classification of breast cancer using hybrid residual neural network and machine learning," *Image and Vision Computing*, vol. 145, p. 104987, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S026288562400091X

[7] A. CALISKAN, F. F. ATES, and M. TOGACAR, "Ultrason tabanlı meme kanseri görüntülerinin derin Öğrenme yaklaşımları ile sınıflandırılması," *Fırat Üniversitesi Fen Bilimleri Dergisi*, vol. 34, no. 2, pp. 179–187, 2022.

[8] K. Jabeen, M. A. Khan, M. Alhaisoni, U. Tariq, Y.-D. Zhang, A. Hamza, A. Mickus, and R. Damaševičius, "Breast cancer classification from ultrasound images using probability-based optimal deep learning feature fusion," *Sensors*, vol. 22, no. 3, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/3/807

[9] B. Liu, S. Liu, Z. Cao, J. Zhang, X. Pu, and J. Yu, "Accurate classification of benign and malignant breast tumors in ultrasound imaging with an enhanced deep learning model," *Frontiers in Bioengineering and Biotechnology*, vol. 13, 06 2025.

[10] C. Kormpos, F. Zantalis, S. Katsoulis, and G. Koulouras, "Evaluating deep learning architectures for breast tumor classification and ultrasound image detection using transfer learning," *Big Data and Cognitive Computing*, vol. 9, p. 111, 04 2025.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[12] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning Transferable Architectures for Scalable Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8697–8710. [Online]. Available: https://arxiv.org/abs/1707.07012

[13] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251–1258. [Online]. Available: https://arxiv.org/abs/1610.02357

[14] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *Advances in neural information processing systems*, vol. 27, 2014.