



A Lightweight CNN-Transformer Hybrid Network for Efficient Cancer Detection Using Ultrasound Images

Thanh-An Pham^{1,2}, Van-Dung Hoang³ , Van-Tuong-Lan Le⁴ ,
and Dung Nguyen¹ 

¹ University of Sciences, Hue University, Hue City, Vietnam
{phamthanhan, nguyendung}@hueuni.edu.vn

² Ho Chi Minh University of Banking, Ho Chi Minh City, Vietnam

³ HCMC University of Technology and Education, Ho Chi Minh City, Vietnam
dunghv@hcmute.edu.vn

⁴ Hue University, Hue City, Vietnam
lvtlan@hueuni.edu.vn

Abstract. Ultrasound is a practical, non-invasive clinical method, which is the first choice for screening and detecting many diseases in current medical examination and treatment, especially breast cancer and thyroid cancer. In this study, a novel model based on stacking attention blocks and depth-wise convolutional blocks is proposed. This hybrid deep learning architecture enables highly accurate detection of breast and thyroid cancer, outperforming state-of-the-art models. The model employs window attention and coordinate attention mechanisms together with large depth-wise convolution kernels to reduce the number of parameters and expand the receptive field. The depth-wise convolutional blocks extract local features, while the attention blocks capture long-range dependencies; their combination produces feature representations rich in both texture and contextual information, thereby improving detection performance. Two datasets of breast cancer, BUSI, and thyroid cancer, TN5000 are selected experimentally to evaluate the proposed model and state-of-the-art models. Our method achieves 96.13% accuracy on the BUSI dataset and 88.99% on the TN5000 set, outperforming the SOTA models. The proposed model has 14.5 million parameters and 5.02 Gflops, which is optimal in both the number of parameters and computational performance compared to state-of-the-art models.

Keywords: Lightweight · Transformer · CNN · Ultrasound images · Hybrid model

1 Introduction

Breast cancer and thyroid cancer are among the most prevalent malignancies worldwide, significantly impacting global health. Early detection of these cancers is critical for improving patient outcomes and survival rates. Ultrasound

imaging has emerged as a preferred diagnostic modality due to its non-invasive nature, cost-effectiveness, real-time imaging capabilities, and absence of ionizing radiation. However, accurate interpretation of ultrasound images remains challenging due to image artifacts, speckle noise, low contrast, and the requirement for expert radiologists to identify subtle malignant features.

In recent years, deep learning techniques have shown substantial potential for analyzing medical images. Convolutional Neural Networks (CNNs) such as ResNet [1], DenseNet [2], and MobileNets [3] are particularly effective at capturing hierarchical local features through layered convolution operations. On the other hand, Transformer-based models, initially developed for sequence modeling and later adapted to visual tasks with Vision Transformer (ViT) [4], excel at modeling long-range relationships in the data using self-attention. Recently, hybrid networks combining CNNs and Transformer [5] modules have gained attention for their ability to exploit both fine-grained local features and broader contextual information.

Despite these advances, current models still encounter challenges in clinical applications. CNN-only architectures often have difficulty capturing global context, while Transformer-based networks typically require high computational resources and extensive datasets for training. Models like Swin Transformer [6] and MaxViT [7] provide strong performance but are often too large and computationally expensive for practical use in resource-limited settings. Additionally, many state-of-the-art architectures have not been specifically designed to accommodate the unique characteristics of ultrasound images, which differ significantly from natural images commonly used in benchmarks such as ImageNet.

To address these limitations, we introduce StackAttention, a lightweight hybrid CNN-Transformer model specifically tailored for efficient cancer detection in ultrasound images. Our approach integrates window attention and coordinate attention mechanisms with large depth-wise convolutional kernels, offering a balance between computational efficiency and detection accuracy. The contributions of this work are threefold: (1) we propose a novel stacking attention design that effectively combines local feature extraction and global context modeling; (2) we achieve state-of-the-art performance on breast cancer and thyroid cancer detection tasks using only 14 million parameters and 5 GFLOPs; and (3) we demonstrate that our lightweight architecture maintains high accuracy while remaining feasible for deployment in resource-constrained clinical environments.

2 Related Work

Deep learning has revolutionized medical image analysis, showing particular success in ultrasound-based diagnostics. Early applications to breast ultrasound classification demonstrated significant improvements over traditional machine learning methods. Yap et al. [8] developed an automated breast lesion detection system using convolutional networks combined with hand-crafted features, achieving promising results but requiring careful feature engineering. More recent works have focused on end-to-end learning approaches that eliminate the need

for manual feature design. Similarly, for thyroid nodule classification, several studies have explored deep network architectures. Chi et al. [9] applied AlexNet and GoogLeNet to thyroid ultrasound images, demonstrating the feasibility of deep learning for this task. Ma et al. [10] proposed a cascaded framework that first segments thyroid nodules before classification, showing improved performance through this two-stage approach. However, these methods primarily rely on standard architectures without specifically addressing the unique challenges of ultrasound imaging, such as speckle noise and variable image quality.

Over the past decade, convolution-based networks have been the dominant architecture for medical image classification. ResNet [1] introduced residual connections that enabled training of very deep networks, addressing the vanishing gradient problem, while DenseNet [2] improved feature propagation through dense connections, achieving better parameter efficiency. These architectures have been widely adopted due to their ability to learn hierarchical representations from raw pixel data. EfficientNet [11] proposed compound scaling to balance network depth, width, and resolution, achieving state-of-the-art results with fewer parameters. MobileNet [3] and MobileNetV2 [12] introduced depth-wise separable convolutions to reduce computational cost, making them suitable for mobile and edge devices. Despite their efficiency, these architectures are inherently limited in capturing long-range dependencies due to the local receptive fields, which can be crucial for understanding global tissue patterns in ultrasound images.

To address the limitations of convolution-only models, attention-based architectures have been explored. The Vision Transformer [4] adapted the Transformer from natural language processing to computer vision by treating images as sequences of patches, showing that attention-based models can match or exceed convolution-based performance when trained on large datasets. However, this approach requires substantial computational resources and massive amounts of training data, limiting its applicability for medical imaging where datasets are often relatively small. Several variants have been proposed to improve efficiency and performance. Swin Transformer [6] introduced hierarchical feature maps and shifted windows for efficient attention computation. DeiT [13] incorporated distillation techniques to train Transformers more efficiently with less data, while MaxViT [7] combined multi-axis attention with convolutional layers to capture both local and global features effectively. Although these models perform strongly, their large size and computational demands make them impractical for deployment in resource-constrained clinical environments.

Recent research has further explored hybrid architectures that combine convolution and attention mechanisms. TransUNet [14] integrated attention encoders with convolution decoders for medical image segmentation, showing improved performance over pure convolution networks. CoAtNet [15] systematically studied the design space of hybrid networks, demonstrating that a proper combination of convolution and attention layers can achieve superior efficiency and accuracy. ConvNeXt [16] modernized convolution networks with design choices inspired by attention models, achieving competitive performance

with simpler architectures, and MobileViT [17] introduced lightweight attention modules suitable for mobile devices. Despite these advances, most hybrid architectures were developed and evaluated on natural image datasets, and their effectiveness for ultrasound image analysis remains underexplored.

Attention mechanisms have become essential components in modern deep learning architectures. Squeeze-and-Excitation Networks [18] introduced channel attention to recalibrate feature responses adaptively. Coordinate Attention [19] extended this concept by preserving spatial information along both horizontal and vertical directions, providing more precise positional encoding. CBAM [20] combined channel and spatial attention sequentially to refine feature maps. Window-based attention mechanisms, such as those used in Swin Transformer [6] and the Neighborhood Attention Transformer [21], have also been explored to reduce computational complexity while maintaining modeling capacity. These mechanisms are particularly promising for medical imaging, where both local texture details and broader anatomical context are important.

In parallel, recent studies have revisited large convolution kernels to improve feature extraction. RepLKNet [22] demonstrated that large kernel convolutions, up to 31×31 , can capture long-range dependencies efficiently through depth-wise separable design. ConvNeXt [16] used 7×7 depth-wise convolutions as part of its modernized architecture, and SLaK [23] further explored very large sparse kernels for efficient feature extraction. Large kernel convolutions offer a simple and efficient alternative to attention mechanisms for expanding receptive fields. However, their application to medical ultrasound images, where both fine-grained texture patterns and broader tissue structures are clinically relevant, remains underexplored.

Finally, deployment in clinical practice requires models that are not only accurate but also computationally efficient. Lightweight architectures such as MobileNetV3 [24], EfficientNetV2 [25], and GhostNet [26] have been proposed to reduce parameter counts and computation while maintaining performance. For ultrasound imaging, some specialized lightweight models exist, but many either sacrifice accuracy for efficiency or remain too costly for practical deployment. There is a clear need for models that can balance accuracy and efficiency to support real-time diagnostic use.

In summary, while deep learning has made substantial progress in medical image analysis, significant challenges remain. Convolution-based architectures struggle with global context, attention-based models are often too resource-intensive, and existing hybrid designs have not been specifically optimized for the unique characteristics of ultrasound images. To address these limitations, we propose *StackAttention*, a lightweight hybrid architecture that combines window attention, coordinate attention, and large depth-wise convolutions. Unlike previous approaches, this design specifically targets ultrasound image analysis while remaining computationally efficient, achieving competitive performance with only 14 million parameters and 5 GFLOPs, making it suitable for deployment in resource-constrained clinical environments where early cancer detection can significantly impact patient outcomes.

3 Methodology

Convolutional Neural Networks (CNNs) demonstrate strong capability in capturing local patterns such as edges, textures, and fine-grained structures, yet their inherently local receptive fields limit their ability to model long-range contextual dependencies. Conversely, Vision Transformer (ViT)-based architectures [4] effectively capture global interactions through self-attention mechanisms but lack inductive biases like locality and translation equivariance, making them less sample-efficient and more prone to overfitting on small medical datasets. Although ViTs achieve state-of-the-art performance on large-scale corpora, their high computational cost has motivated refinements such as Swin Transformer [6] and MaxViT [7] to improve scalability and efficiency. Building on these insights, the proposed model is designed to integrate the complementary strengths of CNNs and Transformers while mitigating their respective weaknesses through a tightly coupled hybrid architecture.

As illustrated in Fig. 1, the proposed architecture adopts a hierarchical design inspired by MaxViT [7], Swin Transformer [6], and ResNet [1]. At the stem, a Coordinate Attention (CA) module [19] is incorporated to enhance early feature encoding. A 7×7 convolution with stride 4 reduces spatial resolution, followed by a 3×3 convolution for refinement. The CA mechanism embeds directional positional information, effectively compensating for the Transformer’s tendency to process image patches as unordered tokens, thereby restoring crucial spatial structure from the earliest stage.

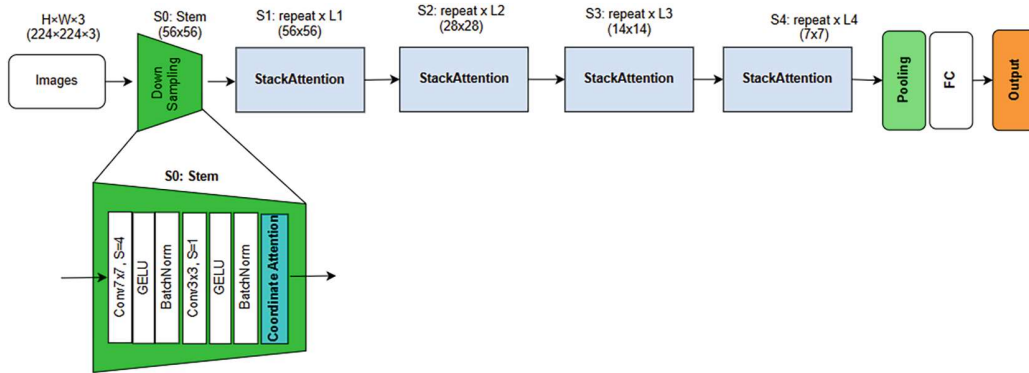


Fig. 1. Overall architecture of the proposed model.

To expand receptive fields while maintaining parameter efficiency, the architecture introduces a StackDepthwise-Conv block, which utilizes large depthwise kernels (5×5 and 7×7) with dilation $d = 3$. Unlike standard convolutions, these depthwise operations increase spatial coverage at low computational cost, enabling the model to integrate information across window boundaries—an essential enhancement for overcoming the locality constraint of window-based attention used in Swin Transformer.

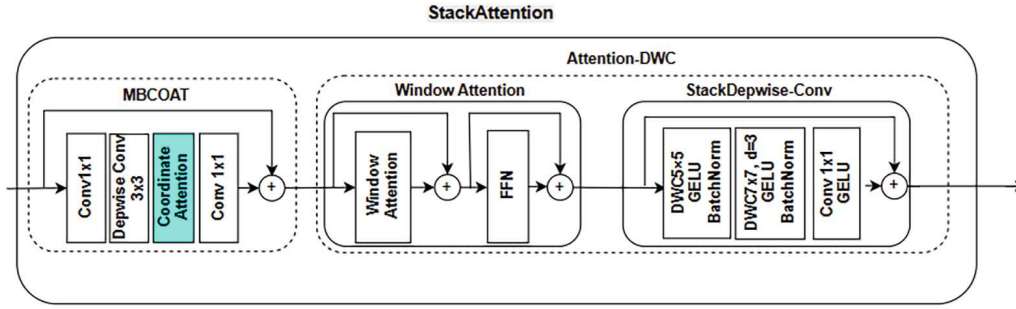


Fig. 2. The detailed structure of the StackAttention block.

For deeper stages, the model incorporates an enhanced MBCoAT block, derived from MBConv [12]. Here, the SE module [18] is replaced with Coordinate Attention to introduce spatially-aware channel recalibration prior to attention computation. Window Attention [6] is then applied, enabling efficient global dependency modeling through non-overlapping windows. The outputs of the Window Attention module are subsequently fused with the long-range receptive fields produced by the StackDepthwise-Conv block. This fusion strategy forms a synergistic hybrid mechanism in which: (1) CNN-based depthwise convolutions provide expanded spatial perception and structured locality; (2) Transformer-based window attention captures global interactions and long-range dependencies.

By coupling these two complementary pathways, the proposed architecture achieves more expressive feature representations, improves cross-region information flow, and enhances robustness in breast and thyroid cancer detection.

The detailed structure of the StackAttention block is described in Fig. 2. The combination of coordinate attention and window-based self-attention is an alternative to using positional coding as in standard ViT. This combined mechanism considers the spatial relationship of neighboring features in the feature map.

4 Materials

In this study, two publicly available ultrasound datasets were selected for evaluation: the BUSI dataset [27] for breast cancer and the TN5000 dataset [28] for thyroid cancer. Both datasets were collected and annotated according to strict protocols by reputable organizations, ensuring high-quality labels. The BUSI dataset contains a total of 780 breast ultrasound images, comprising 437 benign cases, 210 malignant cases, and 133 normal images. Some sample images from the BUSI dataset are shown in Fig. 3.

The TN5000 dataset consists of 5,000 thyroid ultrasound images, categorized into benign and malignant classes. Some representative images from this dataset are shown in Fig. 4.

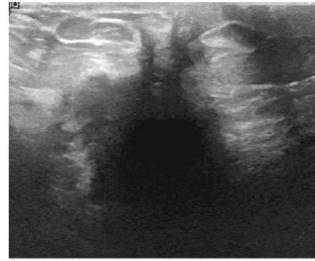
For the TN5000 dataset, an 80:20 split between training and validation sets was used. Table 2 summarizes the number of images per class and their distribution across training and validation sets. As shown, 4,001 images were allocated

Table 1. Statistics of the BUSI dataset.

Class	Number of Images	Train	Validation
Benign	437	350	87
Malignant	210	168	42
Normal	133	107	26
Total	780	625	155



(a) Benign



(b) Malignant



(c) Normal

Fig. 3. Ultrasound Images of breast tumor from BUSI**Table 2.** Statistics of the TN5000 dataset.

Class	Number of Images	Train	Validation
Benign	1,426	1,141	285
Malignant	3,574	2,860	714
Total	5,000	4,001	999

for training and 999 for validation, ensuring a representative sample from each category (Table 1).

The training and validation sets were randomly partitioned from the original public datasets. All images were resized to 224×224 pixels prior to being input into the models for training. In this study, the evaluation metrics—Accuracy, Recall, Precision, Specificity (SPE), and F1-score—are computed from the confusion matrix based on the numbers of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN).

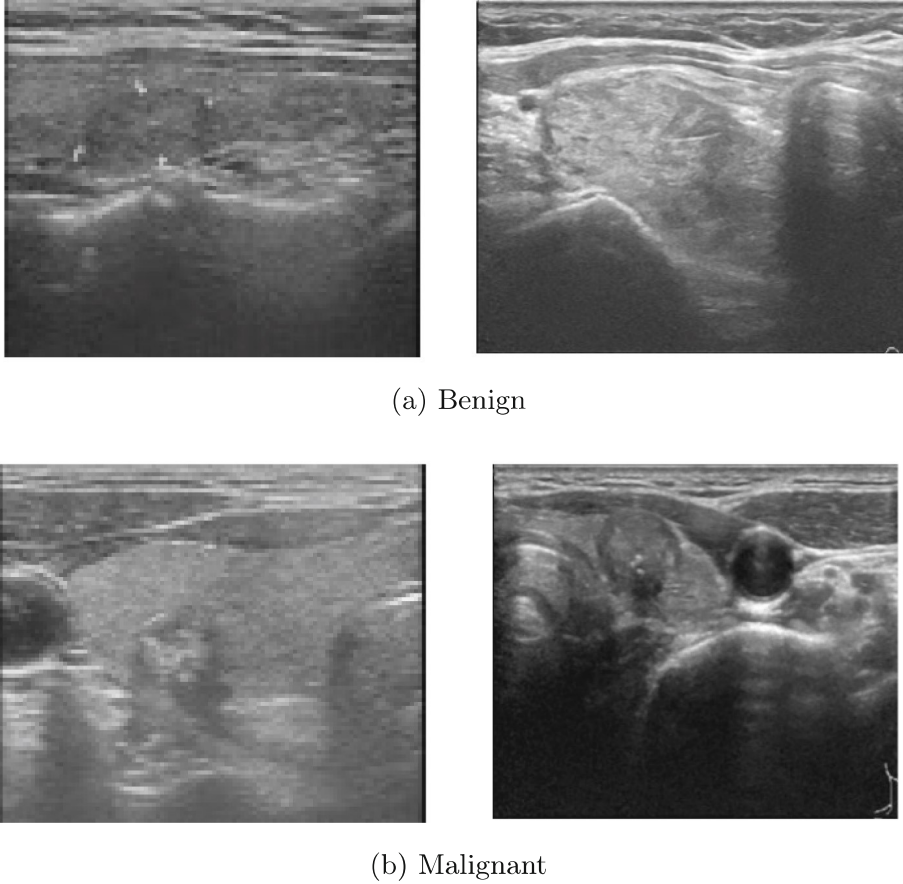


Fig. 4. Ultrasound Images of thyroid nodules from the TN5000 Dataset.

5 Experimental Results and Discussion

In this work, the proposed methodology employs a transfer learning approach, leveraging various pre-trained models on ImageNet-1K and fine-tuning them on the target research datasets. State-of-the-art pre-trained models from the Torchvision library with weights trained on ImageNet-1K were utilized, including ResNet [1], DenseNet [2], ViT [4], Swin-T [6], and MaxViT-T [7]. The proposed StackAttention model, however, was trained from scratch on the ImageNet-1K dataset [29] for 100 epochs, and the best-performing model from this pre-training phase was then used for subsequent fine-tuning on the target datasets (Table 3).

The experimental results on both the BUSI and TN5000 datasets demonstrate the effectiveness of the proposed StackAttention model compared to state-of-the-art CNN and Transformer architectures. On the BUSI dataset (Table 4), our model achieves the highest accuracy of 96.13%, surpassing Swin-T (95.74%) and MaxViT-T (95.20%). It also achieves superior precision (94.92%), specificity (98.50%), and F1-score (94.32%), indicating a strong ability to correctly identify both positive and negative cases while minimizing false positives and false negatives. This improved performance can be attributed to the hybrid architecture,

Table 3. Experimental parameters

Experimental Parameters	Values
Window Size	7
Num_heads	32
Expansion layer of each MLP (α)	4
Number of layers [S1, S2, S3, S4]	[2, 2, 5, 2]
Channel dimensions of [S1, S2, S3, S4]	[64, 128, 256, 512]
Number of Epochs	300
Loss Function	Cross-entropy loss
Batch size	64
Learning rate	0.00002
Experimental Environment	Ubuntu 20.4, GPU L40 (48 GB VRAM), PyTorch 2.4.0+cu121

Table 4. Performance of models on the BUSI dataset (%)

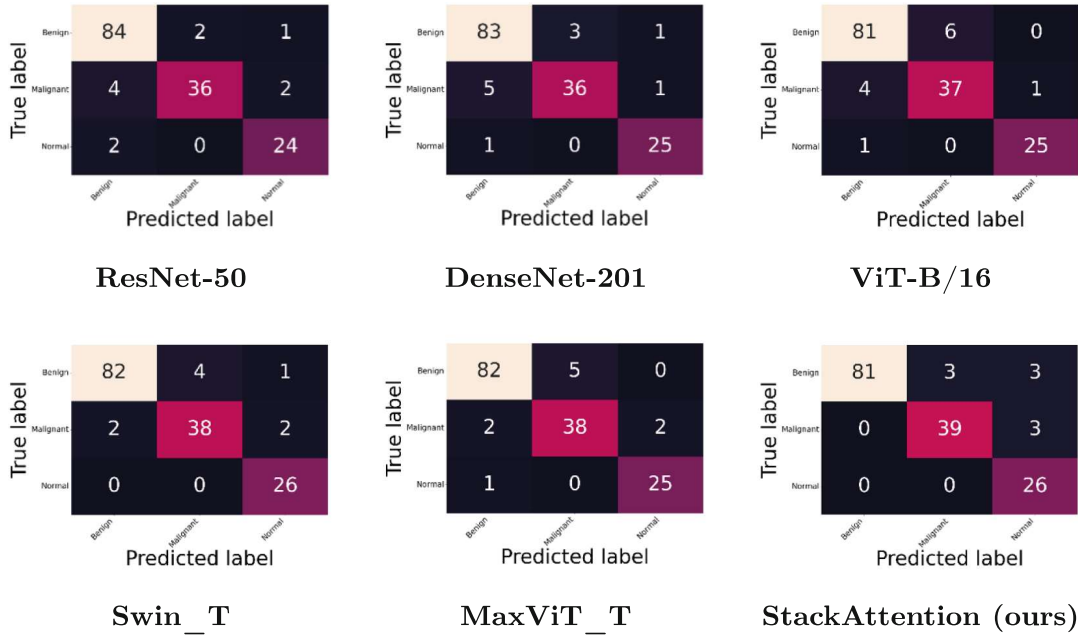
Model	Accuracy	Precision	Recall	Specificity	F1-Score
ResNet-50	94.48	92.97	92.90	94.19	92.85
DenseNet-201	94.48	92.89	92.90	94.07	92.85
ViT_B/16	93.88	92.31	92.26	94.30	92.28
MaxViT_T	95.20	93.63	93.55	96.06	93.57
Swin_T	95.74	94.35	94.19	97.00	94.21
Ours	96.13	94.92	94.19	98.50	94.32

Table 5. Performance of different models on the TN5000 dataset (%).

Model	Accuracy	Precision	Recall	Specificity	F1-Score
ResNet-50	86.79	86.68	86.79	73.43	86.17
DenseNet-201	87.49	87.31	87.49	75.82	87.03
ViT_B/16	88.49	88.60	88.49	75.80	87.94
MaxViT_T	88.29	88.08	88.29	79.30	88.05
Swin_T	88.49	88.33	88.49	81.49	88.38
Ours	88.99	89.01	88.99	77.47	88.55

which effectively combines convolutional operations for local feature extraction with attention mechanisms for capturing long-range dependencies. The use of window attention and coordinate attention allows the model to focus on clinically relevant regions, while large depth-wise convolutional kernels expand the receptive field, enhancing contextual understanding of breast tissue structures in ultrasound images.

On the TN5000 thyroid dataset (Table 5), StackAttention similarly outperforms the compared models with an accuracy of 88.99%, a precision of 89.01%, and an F1-score of 88.55%. While Swin-T achieves higher specificity (81.49%)

**Fig. 5.** Confusion matrix of models on the BUSI dataset.**Table 6.** Parameters and Gflops of the models

Model	Input size	Parameters(M)	GFLOPs(G)
ResNet-50	224×224	23.51	8.26
DenseNet-201	224×224	18.00	8.77
MaxViT_T	224×224	30.40	10.91
Swin_T	224×224	27.50	5.91
ViT_B/16	224×224	85.80	22.57
Ours	224×224	14.50	5.02

than our model (77.47%), StackAttention maintains a better balance between recall and precision, which is crucial for minimizing both false negatives and false positives in clinical diagnosis. The combination of CNN and Transformer modules allows the model to extract fine-grained texture features while simultaneously capturing global context, which is particularly beneficial for ultrasound images that often suffer from low contrast and speckle noise.

Overall, these results confirm that the proposed StackAttention architecture not only improves detection performance over existing models but also maintains computational efficiency, with only 14.5 million parameters and 5.02 GFLOPs. The architecture's ability to integrate local and global features while focusing on important regions explains its superior performance and demonstrates its potential for practical deployment in clinical ultrasound-based cancer screening (Figs. 5, 6 and Table 6).

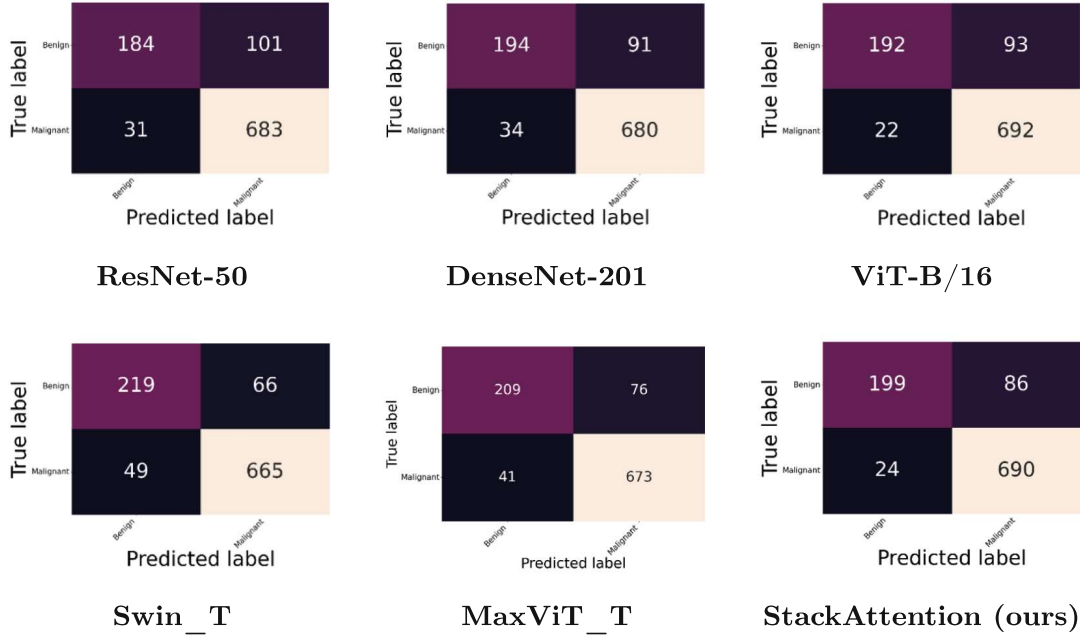


Fig. 6. Confusion matrix of models on the TN5000 dataset.

6 Conclusions

In this study, we proposed a novel hybrid deep learning architecture that effectively integrates stacking attention blocks with depth-wise convolutional blocks for the detection of breast and thyroid cancer from ultrasound images. By combining convolutional operations with attention mechanisms, including window attention and coordinate attention, and employing large convolution kernels, the model not only captures both local and global contextual information but also maintains a lightweight structure with only 14.5 million parameters and 5.02 GFLOPs. Experimental results on the BUSI and TN5000 datasets demonstrate that the proposed model achieves superior accuracy compared to state-of-the-art convolutional and Transformer-based architectures, with 96.13% accuracy on BUSI and 88.99% on TN5000. These results highlight the potential of the model for practical deployment in clinical settings, offering an optimal balance between detection performance and computational efficiency. Overall, the proposed approach provides a promising solution for accurate and efficient ultrasound-based cancer detection, which can contribute to improved early diagnosis and patient outcomes.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
2. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
3. Howard, A.G.: MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv preprint [arXiv:1704.04861](https://arxiv.org/abs/1704.04861) (2017)
4. Dosovitskiy, A.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929) (2020)
5. Vaswani, A.: Attention is all you need. Adv. Neural Inf. Process. Syst. (2017)
6. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
7. Tu, Z., et al.: MaxViT: multi-axis vision transformer. In: European Conference on Computer Vision, pp. 459–479. Springer (2022)
8. Yap, M.H., et al.: Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J. Biomed. Health Inform. **22**(4), 1218–1226 (2017)
9. Chi, J., Walia, E., Babyn, P., Wang, J., Groot, G., Eramian, M.: Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. J. Digit. Imaging **30**(4), 477–486 (2017)
10. Ma, J., Fa, W., Zhu, J., Dong, X., Kong, D.: A pre-trained convolutional neural network based method for thyroid nodule diagnosis. Ultrasonics **73**, 221–230 (2017)
11. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
12. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
13. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357 (2021)
14. Chen, J., et al.: TransUNet: transformers make strong encoders for medical image segmentation. arXiv preprint [arXiv:2102.04306](https://arxiv.org/abs/2102.04306) (2021)
15. Dai, Z., Liu, H., Le, Q.V., Tan, M.: CoatNet: marrying convolution and attention for all data sizes. Adv. Neural. Inf. Process. Syst. **34**, 3965–3977 (2021)
16. Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986 (2022)
17. Mehta, S., Rastegari, M.: MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer. In: International Conference on Learning Representations (2022)
18. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
19. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13713–13722 (2021)

20. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
21. Hassani, A., Walton, S., Li, J., Li, S., Shi, H.: Neighborhood attention transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6185–6194 (2023)
22. Ding, X., Zhang, X., Han, J., Ding, G.: Scaling up your kernels to 31x31: revisiting large kernel design in CNNs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11963–11975 (2022)
23. Liu, S., et al.: More convnets in the 2020s: scaling up kernels beyond 51x51 using sparsity. arXiv preprint [arXiv:2207.03620](https://arxiv.org/abs/2207.03620) (2022)
24. Howard, A., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1314–1324 (2019)
25. Tan, M., Le, Q.: Efficientnetv2: smaller models and faster training. In: International Conference on Machine Learning, pp. 10096–10106. PMLR (2021)
26. Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., Xu, C.: GhostNet: more features from cheap operations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1580–1589 (2020)
27. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. *Data Brief* **28**, 104863 (2020)
28. Hou, X., et al.: An ultrasonography of thyroid nodules dataset with pathological diagnosis annotation for deep learning. *Sci. Data* **11**(1), 1272 (2024)
29. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Li, F.-F.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)