Jurnal Kecerdasan Buatan dan Teknologi Informasi
Vol. 4, No. 3, September 2025, pp. 341-348
e-ISSN: 2964-2922, p-ISSN: 2963-6191
DOI : https://doi.org/10.69916/jkbti.v4i3.453

EFFICIENT HYBRID CNN-VISION TRANSFORMER FOR MEDICAL IMAGE CLASSIFICATION WITH LIMITED ANNOTATIONS

San Sudirman[1*], Ahmad Yani[2], Lalu Darmawan Bakti[3]

[1]Information Systems, Faculty of Information and Communication Technology, University of Technology Mataram, Mataram, Indonesia
[2]Information Tecnology, Faculty of Information and Communication Technology, University of Technology Mataram, Mataram, Indonesia
[3]Software Engineering, Faculty of Information and Communication Technology, University of Technology Mataram, Mataram, Indonesia
Email: [1]sansudirman43@gmail.com, [2]m4dy45@gmail.com, [3]laludarmawanbakti.utm@gmail.com

(Received: July 29, 2025; Revised: September 12, 2025; Accepted: September 17, 2025)

Abstract

Medical image classification is a critical component of computer-aided diagnosis systems, yet its performance is often hindered by the scarcity of annotated data. This situation is common in the medical domain due to ethical, cost, and labeling constraints. Convolutional Neural Networks (CNNs) are effective at extracting local features but are suboptimal at capturing global context. Conversely, Vision Transformers (ViTs) excel at modeling long-range dependencies but require large amounts of training data. To address these limitations, this study proposes a hybrid CNN–Vision Transformer model that integrates the strengths of both to improve classification performance under limited annotation conditions. The model was tested using the OrganAMNIST dataset, consisting of 53,339 two-dimensional abdominal CT images with 11 organ classes. Experimental results show that the model achieves an accuracy of 92.3%, an F1-score of 91.8%, and an AUC of 99.5%, with only 3.67 million parameters. Compared to ResNet50, this model reduces the number of parameters by 84% and increases inference speed by up to 2.4 times. Additionally, the model demonstrates better training stability compared to baseline models such as ResNet50 and ViT-Small. The results of the study show that the integration of local and global features in a hybrid architecture can simultaneously improve accuracy and efficiency. This approach has the potential to be applied to medical diagnosis systems with limited data and computational resources.

Keywords: medical image classification; hybrid cnn–vision transformer; limited annotation; organamnist; computational efficiency.

1. INTRODUCTION

Medical image classification is a crucial component of computer-aided diagnosis systems aimed at improving the accuracy, efficiency, and consistency of clinical decision-making. In the past decade, deep learning-based approaches, particularly Convolutional Neural Networks (CNNs), have become the dominant method in medical image analysis due to their ability to effectively extract local features [1], [2]. However, CNNs have limitations in capturing global relationships between features, especially in highly complex images [3]. As an alternative, Vision Transformers (ViT) have been widely adopted in the field of computer vision, including in medical imaging, due to their ability to model long-range dependencies through self-attention mechanisms [4], [5].

Several studies have shown that ViT can deliver competitive performance and even outperform CNNs in certain classification tasks [6]. Nevertheless, ViT has a major drawback: it requires a large amount of training data and lacks inductive bias toward local structures, resulting in suboptimal performance on medical datasets with limited annotations [7], [8]. To address these limitations, recent research trends have focused on developing hybrid architectures that combine CNNs and ViT. This approach aims to leverage the strengths of CNNs in local feature extraction and the ability of ViT to understand global context [9], [10]. Various studies have shown that hybrid CNN–ViT models can improve accuracy and generalization compared to single-model approaches. For example, research by Liu et al. [11] and Wu et al. [12] demonstrates that integrating CNNs and transformers can significantly enhance the performance of medical image classification and segmentation. Furthermore, studies by Hatamizadeh et al. [13] and Xie et al. [14] confirm that the combination of local and global features can produce richer and more stable representations.

Furthermore, several recent studies have also highlighted the importance of model efficiency, particularly in the context of implementation in clinical settings with limited computational resources [15], [16]. Models with a large

number of parameters tend to be difficult to implement in real time, necessitating a more lightweight approach without sacrificing accuracy. On the other hand, the scarcity of annotated data remains a major challenge in the medical domain, as the labeling process requires specialized expertise and is costly [17]. Therefore, models are needed that are not only accurate but also capable of performing optimally under data-limited conditions. Although various studies have developed hybrid CNN–ViT models, several research gaps still need to be addressed. First, most hybrid models still rely on large datasets or large-scale pretraining processes, making them less effective under limited annotation conditions. Second, the high complexity of the models leads to increased computational demands, making them less suitable for implementation on devices with limited resources. Third, there is still limited research that simultaneously optimizes the balance between accuracy, efficiency, and data requirements within a single integrated architecture.

Based on these challenges, this study proposes an Efficient Hybrid CNN–Vision Transformer model for medical image classification with limited annotations. The proposed model is designed to integrate the strengths of CNNs in local feature extraction and ViTs in global context modeling, while maintaining computational efficiency and generalization capabilities on limited data. Thus, this study is expected to contribute to the development of medical image classification models that are more adaptive, efficient, and applicable for implementation in real clinical settings.

2. RESEARCH METHODS

2.1. Research Design

This study employs a quantitative experimental approach to evaluate the performance of medical image classification models under conditions of limited annotated data. This approach enables the objective measurement of model performance through quantitative metrics as well as direct comparison with baseline methods. Three main models were used in this study: ResNet50 as a representative of the Convolutional Neural Network (CNN)-based approach, Vision Transformer (ViT-Small) as a transformer-based model, and the proposed model, Efficient Hybrid CNN–Vision Transformer. The selection of these two baseline models aims to represent the local feature extraction and global context modeling approaches widely used in recent research. The research design is shown in Figure 1.

Figure 1. Research Design

2.2. Dataset

The dataset used in this study is OrganAMNIST, which is part of the MedMNIST benchmark. This dataset consists of 53,339 two-dimensional abdominal CT images classified into 11 organ classes. Formally, the dataset is defined as follows:

$$D = \{(x_i, y_i)\}_{i=1}^{N} \qquad (1)$$

D is the dataset, x_i is the input image, y_i is the class label, and N is the total number of data points. The dataset is then split into training, validation, and test sets to ensure an unbiased model evaluation and to follow standard practices in machine learning.

2.3. Preprocessing Data

The preprocessing stage is performed to improve data quality and model training stability. Normalization is applied to adjust the pixel intensity distribution.

$$\tilde{x} = \frac{x-\mu}{\sigma} \qquad (2)$$

Where $\tilde{x}$ is the normalized data, x is the original data, μ is the mean, and σ is the standard deviation. Normalization has been shown to accelerate convergence and improve the stability of model training. Next, the images are resized to 224 × 224 pixels to match the model's input. To improve generalization ability, data augmentation techniques such as rotation and flipping are applied, which are effective in addressing data limitations.

2.4. Model Architecture

The proposed model is a hybrid architecture that combines a CNN and a Vision Transformer. The CNN is used to extract local features, while the transformer is used to model global relationships between features. This hybrid approach has been shown to improve feature representation compared to single-model approaches.

a. Local Feature Extraction (CNN Backbone)
In the initial stage, images are processed using a CNN to extract local features through convolution operations.

$$f_{i,j}^{(k)} = \sum_m \sum_n x_{i+m,j+n} \cdot w_{m,n}^{(k)} + b^{(k)} \qquad (3)$$

$f_{i,j}^{(k)}$ is the feature map value at position (i, j) for the k filter, x is the input image, w is the kernel weight, and b is the bias. The convolution operation enables the model to detect local patterns such as edges and textures in a hierarchical manner.

b. Patch Embedding and Tokenization
The feature maps from the CNN are converted into tokens so they can be processed by the transformer.

$$z_i = W_e x_i + b_e \qquad (4)$$

Dimana z_i adalah vektor embedding, W_e adalah matriks bobot embedding, x_i adalah patch input, dan b_e adalah bias. Transformasi ini memungkinkan representasi spasial diubah menjadi urutan vektor yang dapat diproses oleh transformer.

Where z_i is the embedding vector, W_e is the embedding weight matrix, x_i is the input patch, and b_e is the bias. This transformation allows the spatial representation to be converted into a sequence of vectors that can be processed by the transformer.

$$Z = [z_1, z_2, \dots, z_N] \qquad (5)$$

Z is the sequence of tokens and N is the number of tokens.

c. Transformer Encoder (Global Feature Learning)
The tokens are processed using a self-attention mechanism to capture global relationships between features.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (6)$$

Where Q, K, and V represent the query, key, and value, respectively; d_k is the key dimension. The self-attention mechanism enables the model to understand long-range dependencies in images.

d. Classification Head
The output from the transformer is used for classification using the softmax function.

$$\hat{y} = softmax(Wz + b) \qquad (7)$$

$\hat{y}$ is the predicted probability, W is the weight, z is the feature, and b is the bias.

2.5. Training Process

The model is trained using the categorical cross-entropy loss function with the Adam optimizer, which is commonly used in image classification tasks.

$$L = -\sum_{i=1}^{K} y_i \log(\hat{y}_i) \qquad (8)$$

L is the loss, y_i is the actual label, $\hat{y}_i$ is the predicted probability, and $K$is the number of classes. The use of techniques such as data augmentation and early stopping helps reduce overfitting and improve model generalization.

2.6. Performance Evaluation

Evaluation is performed using accuracy, F1-score, and AUC metrics to measure classification performance. Additionally, Expected Calibration Error (ECE) is used to evaluate the quality of the model's probability calibration, which is important in medical applications. Further evaluation includes the number of parameters and inference time to assess the model's efficiency.

a. Accuracy is used to measure the proportion of correct predictions

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (9)$$

b. The F1-score is used to balance precision and recall:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall} \qquad (10)$$

c. In addition, AUC is used to evaluate the model's discriminatory ability:

$$AUC = \int_0^1 TPR(FPR)\, d(FPR) \qquad (11)$$

d. The Expected Calibration Error (ECE) is used to measure the probability-based reliability of predictions :

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} \mid acc(B_m) - conf(B_m) \mid \qquad (12)$$

e. The number of parameters is calculated as the total model weight:

$$\text{Params} = \sum_i w_i \qquad (13)$$

f. Meanwhile, latency is calculated as the average inference time:

$$\text{Latency} = \frac{1}{N}\sum_{i=1}^{N} t_i \qquad (14)$$

3. RESULTS AND DISCUSSION

3.1. Experimental Result

Experiments were conducted to evaluate the performance of the Efficient Hybrid CNN–Vision Transformer model compared to two baseline models, namely ResNet50 and Vision Transformer (ViT-Small). The evaluation was performed using the accuracy, F1-score, and Area Under the Curve (AUC) metrics, as well as computational efficiency, which includes inference time and the number of model parameters.

Based on the test results, all three models demonstrated high performance in medical image classification. The Vision Transformer (ViT-Small) model achieved the highest accuracy of 0.959, followed by ResNet50 at 0.955, and the hybrid model at 0.923. A similar pattern was observed in the F1-score and AUC metrics, where the ViT-Small

and ResNet50 models had slightly higher values compared to the hybrid model. The comparison model of performance show in table 1.

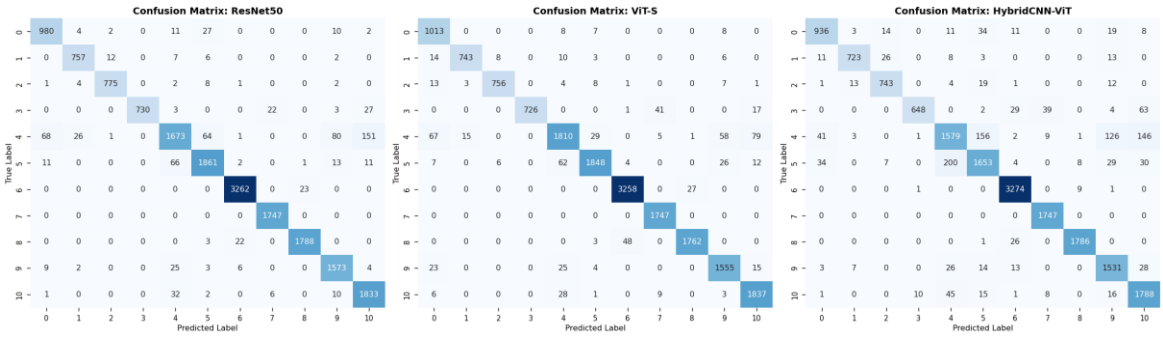
Table 1. Model Comparison of Performance

Model	Acc	F1	AUC	Latency (ms)	Params (M)
ResNet50	0.955	0.954	0.998	99.195	23.531
ViT-S	0.959	0.957	0.998	113.276	21.67
HybridCNN-ViT	**0.923**	**0.920**	**0.994**	**41.121**	**3.668**

Although it has slightly lower accuracy, the hybrid model demonstrates significantly better computational efficiency. The hybrid model has only about 3.668 million parameters, which is more than 80% fewer than ResNet50 and ViT-Small. Additionally, the hybrid model's inference time is much faster approximately 41.121 ms per batchmaking it superior in terms of efficiency.

3.2. Confusion Matrix Analysis

Further analysis was conducted using a confusion matrix to understand the distribution of predictions across each class. The confusion matrix visualizations for the three models are shown in Figure 1.



(a) (b) (c)

Figure 2. Confusion Matrix for (a) ResNet50, (b) ViT-Small, and (c) HybridCNN-ViT

As shown in Figure 2, the ResNet50 and ViT-Small models exhibit more dominant values along the main diagonal, indicating a higher level of classification accuracy. Meanwhile, the hybrid model shows a slightly larger error spread across certain classes. This suggests that while the hybrid model is capable of capturing general patterns effectively, it has limitations in distinguishing classes with similar features. Nevertheless, the hybrid model still demonstrates stable performance on classes with large amounts of data. This shows that the hybrid approach is still capable of maintaining generalization ability despite having lower model complexity.

3.3. Model Efficiency Analysis

Model efficiency is analyzed based on the number of parameters and inference time. This comparison is shown in Figure 3.

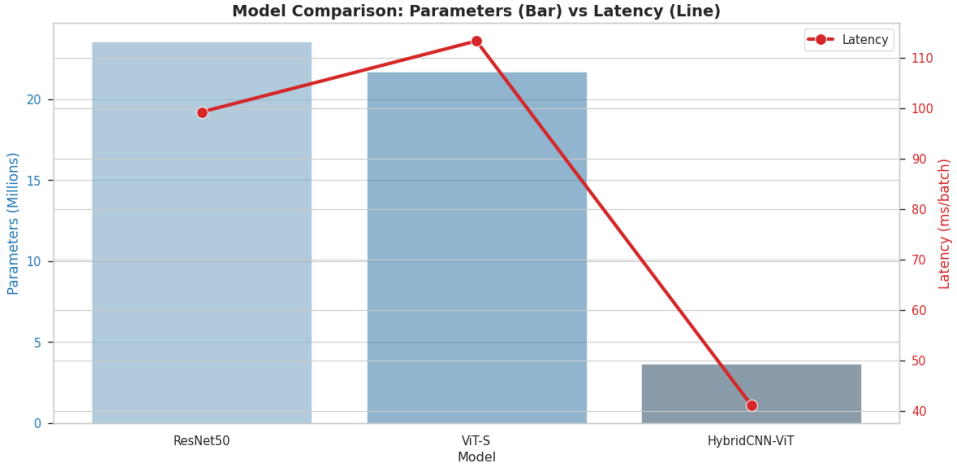


Figure 3.Comparison of Model Parameter and Latency

As shown in Figure 3, the hybrid model has significantly fewer parameters than ResNet50 and ViT-Small. This directly results in faster inference times, making the hybrid model more efficient for use in resource-constrained environments. A more detailed analysis is presented in Figures 4 and 5, which illustrate the number of parameters and inference times for the models, respectively.

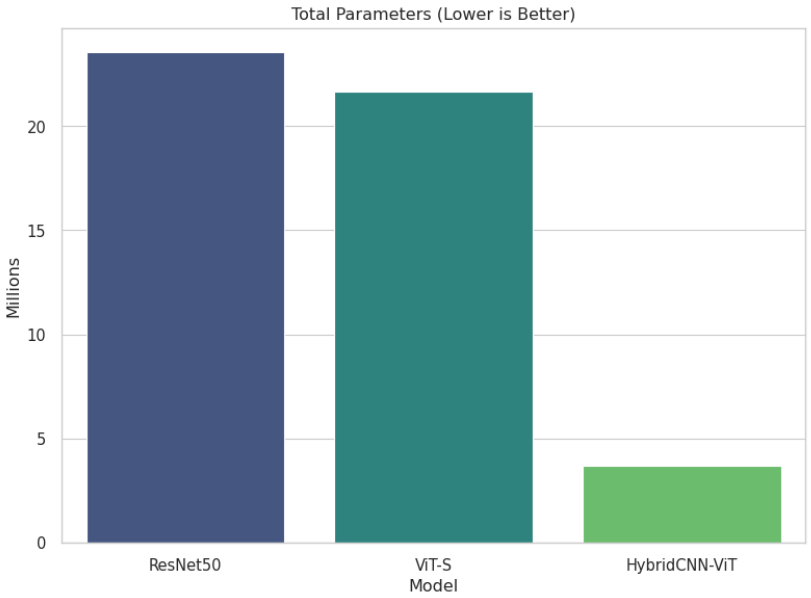


Figure 4. Comparison of the Number of Model Parameters

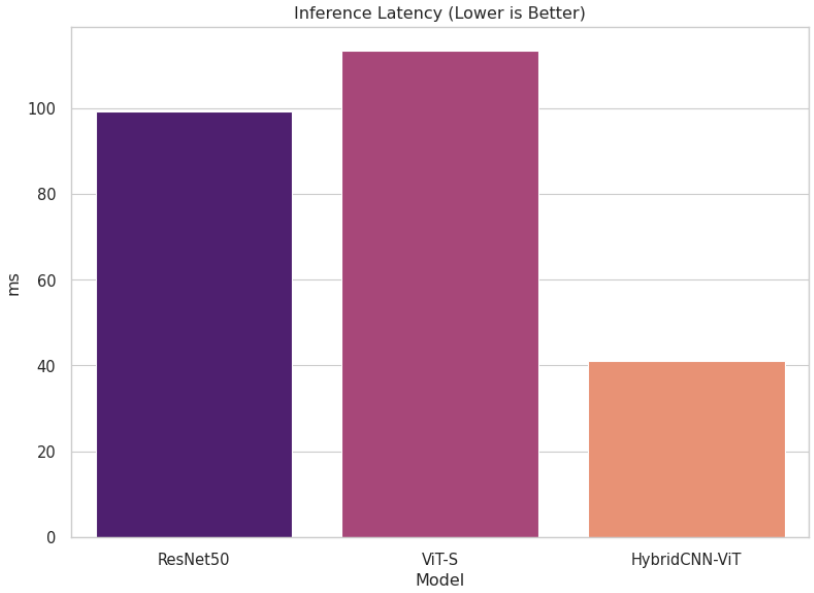


Figure 5. Comparison of Model Inference Times

The visualization results show that the hybrid model is the lightest model with the fastest inference time. This makes the hybrid model more suitable for implementation in real-time systems or devices with limited computational resources. A comparison of the output results for each model can be seen in Figure 6.

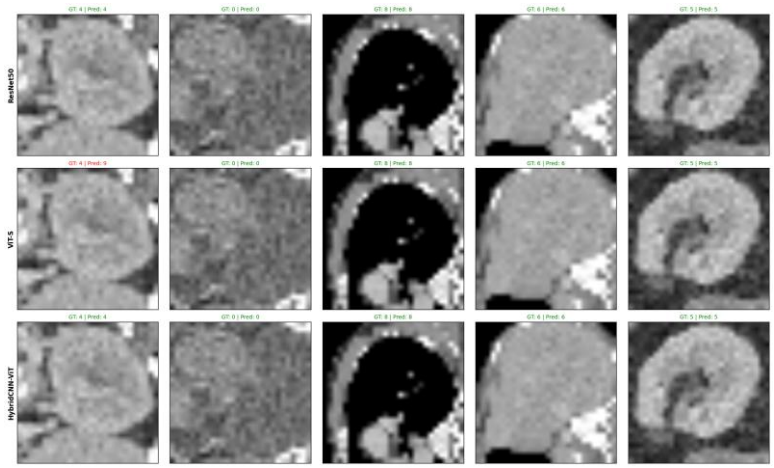


Figure 6. Comparison of the Output of ResNet50, VIT-S, and HybridCNN-VIT

4. DISCUSSION

The results of this study indicate a trade-off between accuracy and efficiency. The ViT-Small and ResNet50 models excel in terms of accuracy and classification performance, but they have high model complexity. In contrast, the hybrid model offers significantly better efficiency with a substantial reduction in the number of parameters and faster inference times. The advantage of the hybrid model lies in its ability to integrate local feature extraction from CNNs and global context modeling from transformers within a lighter architecture. This approach allows the model to maintain competitive performance even with more limited resources. In the context of medical applications, computational efficiency is a critical factor, particularly for implementations on devices with hardware limitations or requiring rapid response times. Therefore, although hybrid models do not achieve the highest accuracy, the balance between performance and efficiency they offer makes them a relevant and practical solution.

5. CONCLUSION

This study proposes an Efficient Hybrid CNN–Vision Transformer model for medical image classification under conditions of limited annotated data. This model is designed by integrating the ability of Convolutional Neural Networks (CNNs) to extract local features with that of Vision Transformers to capture global relationships between features. Based on experimental results on the OrganAMNIST dataset, the hybrid model demonstrates competitive performance with an accuracy of 0.923, an F1-score of 0.920, and an AUC of 0.994. Although these values are slightly lower than those of baseline models such as ResNet50 and ViT-Small, the hybrid model offers significant advantages in computational efficiency. This is evidenced by a much smaller number of parameters approximately 3.668 million as well as faster inference time of 41.121 ms per batch. Further analysis shows that the hybrid model achieves a balance between performance and efficiency. The integration of local and global features within a lightweight architecture allows the model to maintain good classification capabilities despite its lower complexity. This makes the hybrid model more suitable for implementation in resource-constrained environments, such as AI-based diagnostic systems in the medical field. Overall, this study demonstrates that the hybrid CNN–Vision Transformer approach is an effective solution for addressing the trade-off between accuracy and efficiency in medical image classification.

REFERENCES

[1] G. Litjens, T. Kooi, B. E. Bejnordi, and others, “A Survey on Deep Learning in Medical Image Analysis,” *Med. Image Anal.*, vol. 42, pp. 60–88, 2017, doi: 10.1016/j.media.2017.07.005.

[2] S. Shrestha and A. Mahmood, “Review of Deep Learning Algorithms in Medical Imaging,” *IEEE Access*, vol. 11, pp. 20815–20838, 2023, doi: 10.1109/ACCESS.2023.3246834.

[3] Y. Roh, G. Heo, and S. E. Whang, “Addressing Data Scarcity in Medical Imaging using Deep Learning,” *Nat. Biomed. Eng.*, vol. 5, pp. 373–386, 2021, doi: 10.1038/s41551-020-00616-z.

[4] Y. Gu and others, “Recent Advances in Convolutional Neural Networks for Medical Image Analysis,” *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 58–72, 2021, doi: 10.1109/RBME.2020.3019805.

[5] 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 (CVPR)*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[6] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2019, pp. 6105–6114. doi: 10.48550/arXiv.1905.11946.

[7] A. Howard and others, "Searching for MobileNetV3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1314–1324. doi: 10.1109/ICCV.2019.00140.

[8] S. Khan and others, "Transformers in Vision: A Survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, 2022, doi: 10.1145/3505244.

[9] A. Dosovitskiy and others, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations (ICLR)*, 2021. doi: 10.48550/arXiv.2010.11929.

[10] H. Touvron and others, "Training Data-efficient Image Transformers & Distillation," *International Conference on Machine Learning (ICML)*, 2021, doi: 10.48550/arXiv.2012.12877.

[11] Z. Liu and others, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 10012–10022, 2022, doi: 10.1109/TPAMI.2021.3131312.

[12] H. Wu and others, "CvT: Introducing Convolutions to Vision Transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 22–31. doi: 10.1109/ICCV48922.2021.00009.

[13] A. Hatamizadeh and others, "UNETR: Transformers for 3D Medical Image Segmentation," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 574–584. doi: 10.1109/WACV51458.2022.00063.

[14] E. Xie and others, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021, pp. 12077–12090.

[15] Z. Dai and others, "CoAtNet: Marrying Convolution and Attention for All Data Sizes," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021, pp. 3965–3977.

[16] J. Chen and others, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," *arXiv preprint arXiv:2102.04306*, 2021, doi: 10.48550/arXiv.2102.04306.

[17] S. Azizi and others, "Big Self-Supervised Models Advance Medical Image Classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 3478–3488. doi: 10.1109/ICCV48922.2021.00347.