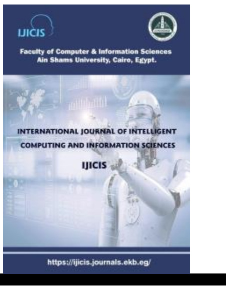
IJICIS, Vol.25, No.3, 93-106
DOI: 10.21608/ijicis.2025.414585.1422

International Journal of Intelligent Computing and Information Sciences

https://ijicis.journals.ekb.eg/



COMPARATIVE ANALYSIS OF LIGHTWEIGHT DEEP LEARNING MODELS FOR CLASSIFICATION OF MEDICAL IMAGES

Omar S. EL-Assiouti*

Scientific Computing,
Faculty of Computer and Information Sciences, Ain Shams University
Cairo, Egypt
omarsherif@cis.asu.edu.eg

Ghada Hamed

Scientific Computing,
Faculty of Computer and Information Sciences, Ain Shams University
Cairo, Egypt
ghadahamed@cis.asu.edu.eg

Dina Khattab

Scientific Computing,
Faculty of Computer and Information Sciences, Ain Shams University
Cairo, Egypt
dina.khattab@cis.asu.edu.eg

Hala M. Ebied

Scientific Computing,
Faculty of Computer and Information Sciences, Ain Shams University
Cairo, Egypt
halam@cis.asu.edu.eg

Received 2025-08-16; Revised 2025-08-25; Accepted 2025-08-25

Abstract: *Accurate medical imaging analysis is crucial for clinical decision-making and effective diagnosis. While deep learning has shown impressive results in different vision tasks, including medical image classification tasks, many of these models are designed to be computationally intensive and come with large number of parameters and high computational cost, making them impractical for deployment on resource-constrained and edge devices. Recent advances have introduced efficient lightweight models that can achieve comparable results, while being resource efficient and suitable for mobile and embedded applications. In this paper, we perform a comprehensive comparison of recent state-of-the art lightweight models that fall under three different categories, including Convolutional neural networks (CNNs), Vision Transformers (ViTs), and hybrid approaches that combine the strengths of both paradigms. These models are evaluated on multiple medical imaging tasks. Specifically, we conduct experiments using HAM-10000 skin lesion dataset and brain tumor dataset for skin and brain cancer classification tasks, respectively. On the brain tumor dataset, MobileNetV4, DeiT-Ti, and FastViT-T12 achieved the highest accuracy of 99.92%, while on the HAM-10000 dataset, DeiT-Ti and MobileViT-V2 obtained the best accuracy of 92.75%.*

Keywords: *Medical Imaging, Vision Transformers (ViTs), Convolutional Neural Networks (CNNs), Hybrid Models, lightweight models*

1. Introduction

*__Corresponding Author__: Omar S. EL-Assiouti

Scientific Computing Department, Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt

Email address: omarsherif@cis.asu,edu.eg

Medical imaging plays an important role in capturing detailed visual information about internal body structures, lesions and abnormal conditions, enabling clinicians and doctors to make accurate decisions and develop personalized treatment planning. Over the past years, artificial intelligence has been widely employed in medical tasks to enhance the diagnosis and assist in the decision-making process [1]. Especially, the evolution of deep learning has shown remarkable success in automating and improving performance across different medical tasks, including medical image classification [2], segmentation [3], detection [4], etc.

These deep learning models are typically composed of multiple layers and a large number of parameters to capture complex and fine-grained features with high precision, resulting in accurate predictive capabilities. However, the high computational demands, memory requirements, and latency associated with such models limit their applicability in real-time and resource-constrained environments. To address these limitations, recent research has focused on the development of lightweight efficient deep learning models [5] that offers robust performance while significantly reducing the number of parameters and computational overhead, making them more suitable in resource-constraint environments. In this paper, we focus on experimenting multiple lightweight models on different medical imaging tasks such as, brain cancer and skin lesion classification tasks. For instance, we utilize the brain tumor dataset [6], which includes four categories: non-tumorous and three distinct tumor types, and the HAM-10000 dataset [7] which comprises seven different classes of skin lesion.

Recent advances in deep learning have introduced various architectural paradigms, primarily falling into these three categories: Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and Hybrid architectures. CNNs have been the traditional backbones for different computer vision tasks, due to their strong spatial features extraction capabilities and inductive biases, but they lack global processing capabilities. On the other hand, ViTs are designed to capture global relationships and long-range dependencies more effectively, but they typically demand more computational cost and larger datasets to achieve better or comparable results to CNNs. Hybrid architectures combine the benefits of both CNNs and ViTs, achieving a balance between performance and efficiency.

The main purpose of this paper is to provide a comparative study between different state-of-the-art models that belong to different categories, including CNNs, ViTs, and hybrid models by evaluating them on diverse medical image classification tasks. We evaluate their performance under two training settings: from scratch and with transfer learning.

The remainder of this paper is structured as follows: Section 2 reviews different categories of lightweight models, including CNNs, ViTs and hybrid architectures. Section 3 provides a detailed exploration of different state-of-the-art models belonging to these categories. Section 4 provides details about the datasets, implementation details and experimental results. Finally, Section 5 discusses the conclusion and future work of our study.

2. Related Work

2.1. Convolutional Neural Networks

CNNs have been widely employed in different computer vision tasks due to their capabilities in capturing efficient hierarchal representations through stacked convolution and pooling layers. Convolutional layers are designed to capture local features through trainable filters, while pooling layers reduce the spatial resolution, which enhances the scalability of the model. This hierarchical structure makes CNNs highly effective for downstream tasks such as image classification, segmentation, and object detection, etc. Over the years, several studies have introduced different CNN models, achieving state-of-the-art results on different benchmarks [8], [9], [10], [11]. However, many of these models are computationally intensive and require large memory requirements. To address this, other studies [12], [13], [14], [15], [16], [17] have focused on introducing efficient lightweight CNN models with lower computational requirements while maintaining robust performance, making them a preferred choice on mobile and resource-constrained devices.

Several studies have introduced different lightweight CNN models for medical image classification tasks [18], [19], [20], [21]. For instance, [21] proposed a custom lightweight CNN for brain tumor classification, which extracts features from augmented, skull-stripped brain MRI images to classify them as normal or abnormal. TurkerNet [19] presented a lightweight CNN model for skin cancer classification, which integrates MBConv4, squeeze-excitation (SE) blocks, residual connections, and decomposition blocks.

2.2. Vision Transformers

Vision Transformers (ViTs) [22] have been introduced recently to the vision community, inspired by its outstanding success in natural language processing tasks [23] and their ability to capture global context through self-attention mechanism. Moreover, self-attention enables the model to capture long-range dependencies, as each patch token attends to every other token in the image, enabling the model to have comprehensive understanding of sematic patterns. However, self-attention comes with high computational complexity and typically requires large-scale datasets to train effectively from scratch. Recent research has explored strategies such as distillation, transfer learning, extensive data augmentations, efficient attention mechanisms, hierarchical pooling and other approaches to enhance the generalization of ViTs [24], [25], [26].

Several studies introduced different lightweight variants of ViTs in medical classification tasks [27], [28], [29]. For example, SkinDistilViT [27] proposed a lightweight vision transformer network, which employs a cascading distillation process for enhanced skin lesion classification. MaxCerVixT [29] proposed a new lightweight vision transformer model for detecting cervical cancer, and it is based on MaxViT architecture, but replaces MBConv block with ConvNextV2 block. Additionally, it replaces MLP blocks with GRN-based MLPs.

2.3. Hybrid Approaches

Hybrid approaches are designed to combine the benefits of both CNNs and ViTs. CNNs excels at processing local features and incorporate inductive biases which improves its generalization, especially on limited datasets. However, their reliance on local context restricts their ability to capture global context and long-range dependencies, which limits their ability to understand complex relationships. Moreover, CNNs require fixed input size, making them less flexible when dealing with images of varying resolutions. On the other hand, ViTs are effective at capturing global context and long-range dependencies through self-attention mechanisms, but it lacks the inductive biases inherent in CNNs, making them more data-hungry and less effective when trained from scratch on smaller datasets. Hybrid models [30], [31], [32] combine the benefits of both worlds to alleviate the problems discussed in each approach, making it the preferred choice for many downstream tasks. Recent studies [33], [34], [35], [36], [37], [38] have also introduced resource-efficient hybrid models that are suitable for edge devices.

Recently, several works incorporated lightweight hybrid network designs for medical classification tasks [39], [40]. Specifically, NeXtBrain [39] introduced a novel hybrid model for brain tumor classification, they mainly propose a novel convolution and transformer blocks named the NeXt Convolutional Block (NCB) and the NeXt Transformer Block (NTB), to effectively combine local and global features, and thus improving the classification results. In [40], they introduced a lightweight model based on improved hybrid MobileViT network. They introduced an enhanced feature representation (EFR) module to obtain a richer feature representation and a cosine similarity downsampling (CSD) module to reduce information loss and to enhance the model capability to capture the key features.

2.4. Discussion

The primary goal of this study is to compare the raw performance of state-of-the-art lightweight models from three architectural families — Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and Hybrid models — across datasets from different medical domains. Unlike many prior works, we do not focus on applying task-specific preprocessing or extensive augmentation and balancing techniques to optimize performance for a particular dataset or task. Our main focus is on evaluating lightweight architectures for their efficiency, which enables their practical use in real-time scenarios and on resource-constrained devices.

3. Model Exploration

3.1. CNN Architectures

MobileNetV2 [13], introduced in 2018, is a widely utilized lightweight model that builds upon the MobileNets family [12]. These models are widely recognized in the vision community due to their efficiency, lightweight design, and their applicability for deployment on resource-constrained devices. MobileNetV2 is based on the inverted residual structure and replaces the standard convolution operations with depthwise separable convolutions to reduce computation and model size. The depthwise separable convolution is composed of two separate layers: a depthwise convolution followed by a pointwise convolution. The depthwise convolution performs lightweight filtering by applying one convolutional filter per each input channel, and the pointwise convolution is simply a 1x1 convolution, which effectively combines the outputs across channels. This design significantly decreases the number of parameters while maintaining competitive performance compared to other state-of-the-art models. MobileNetV2 has been also widely employed as a backbone in different vision tasks such as object detection, image segmentation, few shot learning, and generative models due to its efficient design. MobileNetV2 model is given Figure 1.

MobileNetV4 [15] is the most recent version of the MobileNet series, designed to preserve the family’s focus on efficient and lightweight models for deployment on edge devices. The architecture introduces two novel building blocks named Universal Inverted Bottleneck (UIB) and Extra Depthwise (ExtraDW) Inverted Bottleneck block. The UIB extends the classical Inverted Bottleneck block proposed in MobileNetV2 [13], by incorporating useful design benefits from modern architectures, including

ConvNext block [11] and the Feedforward Network (FFN) block structure utilized in ViT [22]. The ExtraDW block introduces additional depthwise convolution layers to further improve spatial feature extraction, enhancing the model's features representation with minimal added cost. They also offer distillation in their methodology to further enhance the model accuracy and efficiency. MobileNetV4 surpassed its previous family versions and other state-of-the-art lightweight models on different benchmarks. Additionally, an efficient multi-query attention (MQA) module is also employed in their medium and large variants to further boost performance, making them hybrid-like approaches, but this work evaluates the small pure CNN-variant as it is more lightweight and efficient.

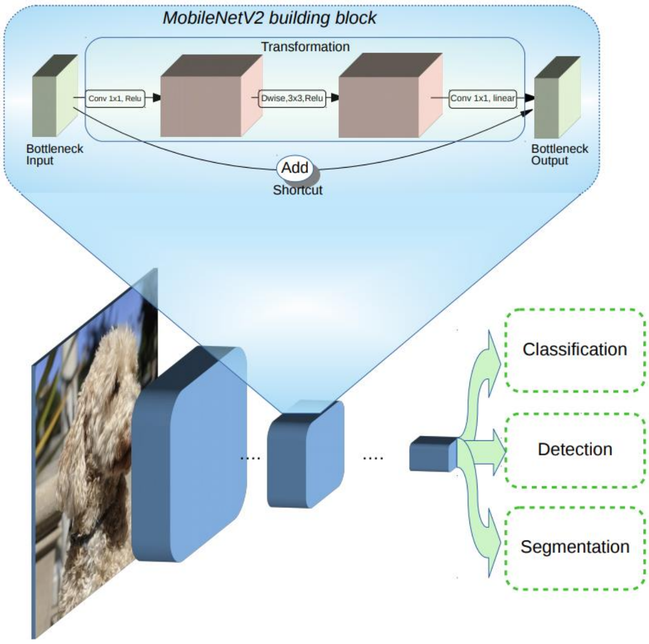


Figure. 1: MobileNetV2 model [13].

3.2. ViT Architectures

Data-efficient Image Transformer (DeiT) [24] is the first work to introduce knowledge distillation in Vision Transformers (ViTs) [22]. The authors propose a teacher-student paradigm, by augmenting the ViT architecture with an additional distillation token, alongside the original image patch tokens and the classification (CLS) token. This distillation token interacts with other tokens through self-attention and is trained to match the output logits of a pre-trained CNN-based teacher model, while the CLS token is optimized to match the ground truth labels, as in the original ViT. In their experiments, the authors utilize RegNet [8] as the teacher CNN model. Traditional ViTs typically require pretraining on massive datasets (e.g., JFT-300M dataset [41] which contains 300 million images) to achieve CNN level performance, which excel at generalizing on limited data due to its strong inductive biases. In contrast, DeiT achieves competitive generalization while being trained only on the ImageNet dataset [42], without relying on external large-scale data, thanks to their proposed distillation strategy. In our study, we experiment the DeiT-Ti variant due to its compact design and efficiency. DeiT model is given in Figure 2.

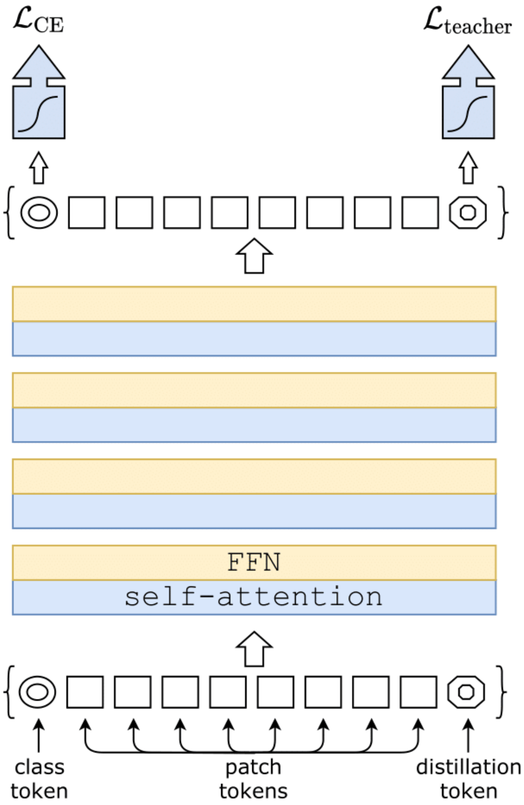
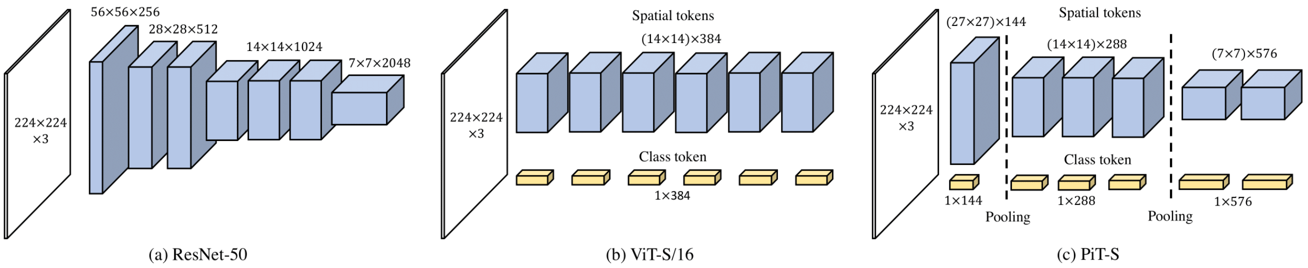


Figure. 2: DeiT model [24].

Pooling-based Vision Transformer (PiT) [25] is a pure transformer network that is built upon ViT original design, while incorporating the successful design benefits of convolutional neural networks (CNNs). CNNs are designed basically to increase feature depth and reduce spatial resolution through pooling or strided convolutions, and thus capturing hierarchical representation effectively. In contrast, standard vision transformers rely mainly on self-attention operations which maintain a fixed size resolution across all layers, potentially limiting model scalability. PiT addresses this by introducing a hierarchical structure using pooling layers between transformer stages, which downsample token embeddings similar to how CNNs reduces feature map resolution. The authors show that utilizing resnet-style [9] dimensional settings enhances the performance of ViT, while keeping the model lightweight and scalable. In our work, we experiment with the tiny variant of PiT, which is called PiT-Ti. As shown in Figure 3, PiT employs a hierarchical architecture similar to ResNet and avoids the fixed-resolution constraint of traditional ViTs.



(a) ResNet-50 (b) ViT-S/16 (c) PiT-S

Figure. 3: Architectural comparison between the PiT model, ResNet, and ViT [25].

3.3. Hybrid Architectures

MobileViT [34] aims to build lightweight and mobile friendly networks for different mobile vision tasks by combining the benefits of both convolution and transformers into a unified network. Their architecture comprises two main blocks, including MobileNetV2 block [13], which focuses on local feature extraction and their proposed MobileViT block, which integrates both convolutional and transformer layers. In a MobileViT block, the input is first processed by convolutions, then unfolded into non-overlapping patches and passed through a transformer encoder to model global representations, and finally, the transformed features are folded back to the original spatial resolution and fused with the block input via additional convolutional layers. This hybrid design allows the network to learn local and global representations. MobileViTV2 [35] is built upon MobileViT and introduces a more efficient design, by replacing the multi-headed self-attention layer in the MobileViT block with an efficient separable self-attention method. This proposed method operates with linear complexity. Specifically, it reduces the attention complexity from $\boldsymbol{o(k^2)\ to\ o(k)}$ only. In our study, we evaluate MobileViTV2 on the experimented datasets due to its improved efficiency compared to MobileViT. MobileViT is illustrated in Figure 4.

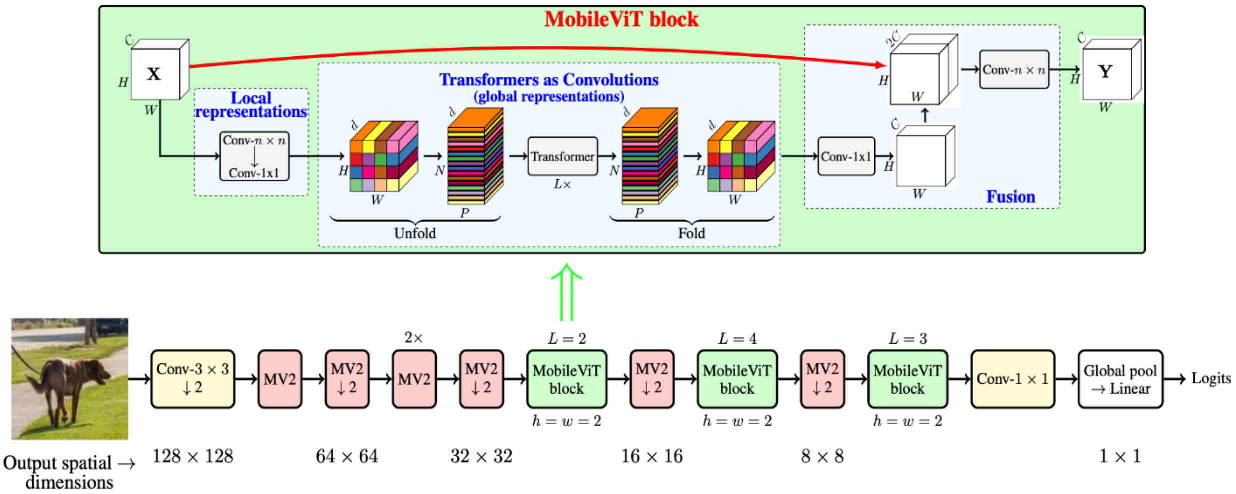


Figure. 4: MobileViT model [34].

FastViT [36] is a fast hybrid Vision Transformer architecture that achieves a state-of-the-art latency and accuracy tradeoff. Its core innovation is the RepMixer block, an efficient convolution-based token mixer that employs structural reparameterization similar to ConvMixer [43]. This structural reparameterization enhances model capacity and reduces memory access by reparametrizing the skip connections during inference. For training, FastViT employs linear train-time overparameterization, where standard dense $\boldsymbol{kxk}$ convolutions are replaced with factorized versions to improve performance. The architecture strategically avoids self-attention in its initial layers, where features maintain high resolution. Instead, it utilizes large kernel convolutions as an efficient alternative in the early stages of the network, while self-attention mechanism is applied only in the latest stage that operates on low-resolution features. FastViT has achieved comparable performance to state-of-the-art models like CMT [44], ConvNext [11], EfficientNet [17] and other models, while being much faster on mobile and edge devices at inference.

Figure 5 illustrates the FastViT architecture and the application of reparameterization during inference to fuse linear blocks.

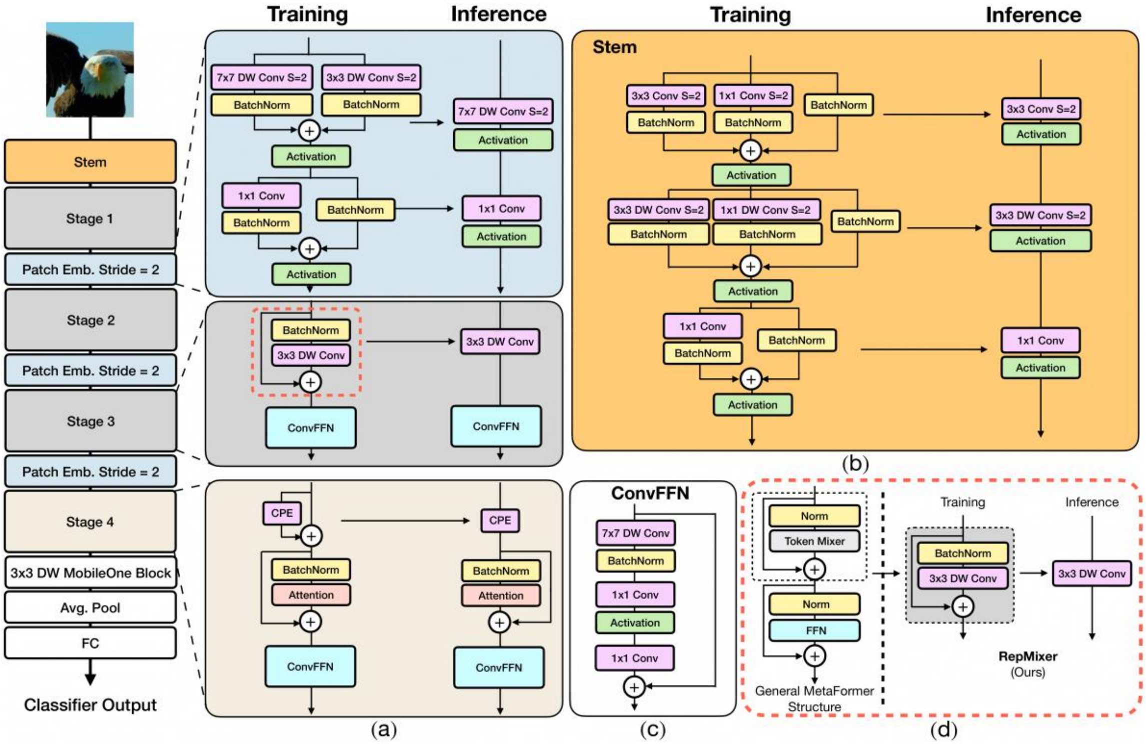


Figure. 5: FastViT model [36].

4. Experiments

4.1. Datasets

The brain tumor dataset [6] comprises 7023 MRI images captured from different patients. This dataset is a collection of three different datasets including, Figshare, Br35H, and SARTAJ datasets. The dataset images are categorized into four different types, including no-tumor, meningioma, glioma, and pituitary. The training and testing sets comprise 5712 and 1311 samples, respectively. Table 1 presents the distribution of the samples utilized in each of the training and testing sets. Figure 6 shows a random sample from each of the four different brain tumor categories.

Table 1: The distribution of training and testing samples for brain tumor dataset.

Class	# Training samples	# Testing samples
No tumor	1595	405
pituitary	1457	300
meningioma	1339	306
glioma	1321	300

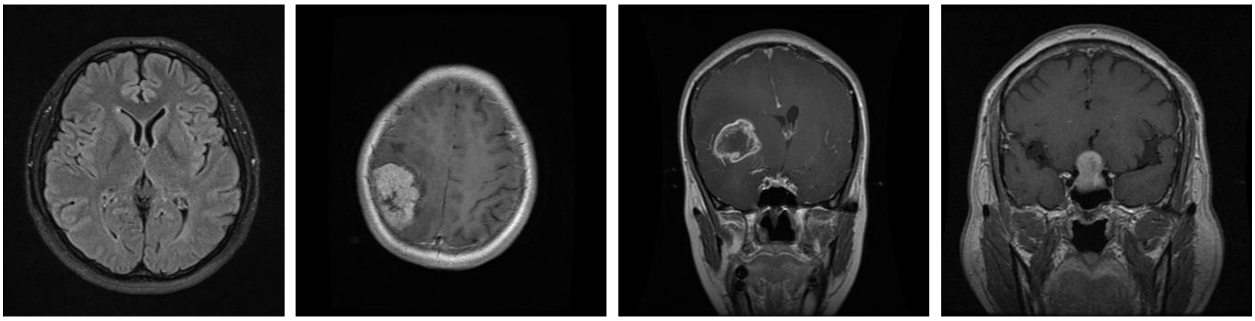


Figure. 6: Random samples of the brain tumor dataset.

The skin lesion HAM-1000 dataset [7] contains 10015 total dermatoscopic images. Each image is of size 450 x 600 pixels, and the dataset is widely used for training and evaluating models in skin cancer classification tasks. The dataset consists of seven different skin lesion categories: Keratosis and Intra-Epithelial Carcinoma (AKIEC), Melanoma (MEL), Dermatofibroma (DF), Melanocytic Nevi (NV), Vascular lesions (VASC), Benign Keratosis (BKL), and Basal Cell Carcinoma (BCC). The dataset distribution is imbalanced which represents a challenge in medical imaging tasks. The samples range from 6705 for the largest class (NV) to only 115 sample for the smallest one (DF). In this paper, we follow the same data split utilized in [38], [45], where the training set consists of 9187 samples and the testing set consists of 828 samples. Figure 2 represents random samples from the HAM-10000 dataset.

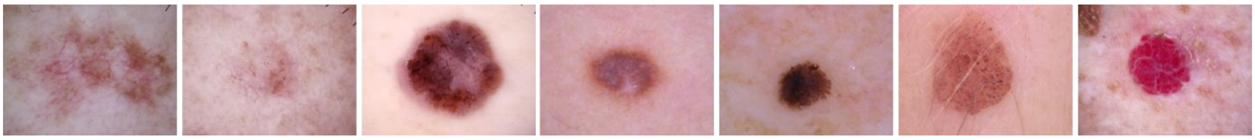


Figure. 7: Random samples of the HAM-10000 dataset.

4.2. Implementation Details

We utilize an image resolution of 224x224 for both datasets. All models are initialized with ImageNet-pretrained weights and fine-tuned for 200 epochs using the PyTorch framework on an NVIDIA L4 GPU. We utilize a batch size of 64 during training and we employ basic augmentation techniques such as horizontal flipping and random rotation. For optimization, the Adam optimizer [46] was utilized for CNN-based models with a learning rate of 5e-4. Conversely, for Vision Transformers and Hybrid models, the AdamW optimizer [47] was employed, with a cosine learning rate scheduler, an initial learning rate of 5e-5, a warmup learning rate of 2e-5 for the first 5 epochs, and a weight decay of 0.05.

4.3. Experimental Results

In this section, we evaluate different lightweight models that are categorized into three different families, including CNNs, ViTs, and Hybrid models, on two different medical tasks. All models are tested under two configurations: trained from scratch and with transfer learning. In the transfer learning setting, the models are initialized with ImageNet-pretrained weights [42] to study the effect of transfer learning on the state-of-the-art models. Utilizing pretrained models enhances the performance specifically on HAM-10000 due to the limited and unbalanced nature of this dataset. The comparison between models' computational cost is provided in Table 2, while the comparison between models' performance is summarized in Table 3.

Comparison between CNN models. In our experiments, we evaluate two lightweight pure CNN models, including MobileNetV2 [13] and MobileNetV4 [15] on the two medical datasets under the two training configurations. When trained from scratch, both MobileNetV2 and MobileNetV4 achieved 99.62% accuracy on the brain tumor dataset, whereas on the HAM-10000 dataset, MobileNetV2 achieved 90.10%, outperforming MobileNetV4 by 1.69%. With transfer learning, MobileNetV4 achieved 99.92% on brain tumor dataset, surpassing MobileNetV2 by 0.07% and on the HAM-10000 dataset, MobileNetV2 achieved 92.51%, surpassing MobileNetV4 by 0.36%. The detailed comparison for both settings is given in Table 3.

Table 2: Comparison of the experimented models' computational requirements.

Category	Model	# Parameters	FLOPs (G)
Convolutional Neural Networks (CNNs)	MobileNetV2 [13]	2.2M	0.3G
	MobileNetV4 [15]	2.5M	0.2G
Vision Transformers (ViTs)	PiT-Ti [25]	4.6M	0.7G
	DeiT-Ti [24]	5.5M	1.1G
Hybrid models	FastViT-T12 [36]	6.5M	1.1G
	MobileViTV2 [35]	4.4M	1.4G

Table 3: Comparison of the performance of the evaluated models on both datasets.

Category	Model	Transfer learning	Brain Tumor Accuracy	HAM-10000 Accuracy
Convolutional Neural Networks (CNNs)	MobileNetV2 [13]	✗	99.62%	90.10%
		✓	99.85%	92.51%
	MobileNetV4 [15]	✗	99.62%	88.41%
		✓	**99.92%**	92.15%
Vision Transformers (ViTs)	PiT-Ti [25]	✗	99.31%	87.08%
		✓	99.85%	92.64%
	DeiT-Ti [24]	✗	98.47%	87.20%
		✓	**99.92%**	**92.75%**
Hybrid models	FastViT-T12 [36]	✗	99.54%	87.92%
		✓	**99.92%**	92.39%
	MobileViTV2 [35]	✗	99.62%	88.16%
		✓	99.85%	**92.75%**

Comparison between ViT models. Similarly, we evaluate two lightweight pure ViT models, including PiT-Ti [25] and DeiT-Ti [24] on the same datasets under both training configurations. When training from scratch, PiT-Ti achieved 99.31% on brain tumor dataset, surpassing DeiT-Ti by 0.84%, whereas on the HAM-10000 dataset, DeiT-Ti achieved 87.20%, surpassing PiT-Ti by 0.12%. With transfer learning, DeiT-Ti outperformed PiT-Ti on both datasets, achieving 99.92% and 92.75% on brain tumor and HAM-10000 datasets respectively, surpassing PiT-Ti by 0.07% and 0.11%.

Comparison between Hybrid models. Finally, we evaluate two lightweight hybrid models, including FastViT-T12 [36] and MobileViTV2 [35] on both datasets under the same configurations. When training from scratch, MobileViTV2 outperformed FastViT-T12 on both datasets, achieving 99.62% and 88.16% on brain tumor and HAM-10000 datasets, respectively surpassing FastViT-T12 by 0.08% and 0.24% on both datasets. With transfer learning, FastViT-T12 achieved 99.92% on brain tumor dataset, surpassing MobileViTV2 by 0.07%, while MobileViTV2 achieved 92.75% on HAM-10000 dataset, surpassing FastViT-T12 by 0.36%.

Due to the limited size of datasets, all models demonstrated significant performance improvement in the transfer learning setting, as observed in Table 3, especially on the HAM-10000 dataset. Specifically, the Vision Transformers (ViTs)-based models achieved the lowest performance on both datasets compared to CNN and Hybrid models when trained from scratch. This is attributed to their high data requirements, due to their large model capacity and reliance on global feature extraction with the self-attention mechanism, which fails to generalize well when trained from scratch on small datasets. However, with transfer learning, the generalization ability of all models — particularly the ViT-based ones — improved significantly.

5. Conclusion and Future work

In this paper, we present a comparative analysis for different state-of-the-art lightweight models that lies into three different categories, including CNNs, ViTs, and Hybrid models. We experiment these models on two different medical tasks, including brain tumor and skin cancer classification, and in two different settings, with and without transfer learning. Specifically, we utilize brain tumor and HAM-10000 datasets for each task, respectively. This study main focus is to evaluate multiple efficient and lightweight models that can be deployed on resource constraint environments and real-time use-cases.
For brain tumor dataset, MobileNetV4, DeiT-Ti, and FastViT-T12 achieved the best testing accuracy of 99.92%. Regarding the HAM-10000 dataset, MobileViTV2 and DeiT-Ti achieved the best testing accuracy of 92.75%. In this study, we did not apply any task-specific preprocessing, augmentation, or oversampling strategies, as our primary focus was to compare the raw performance of lightweight models across different medical tasks.

In future work, we can experiment more advanced preprocessing techniques such as extensive augmentation, and oversampling techniques, to improve performance on the class-imbalanced HAM-10000 dataset. Additionally, we may explore powerful techniques such as knowledge distillation, ensemble learning, and mixture of experts (MOEs) to further enhance model accuracy and generalization.

References

[1] L. Pinto-Coelho, “How Artificial Intelligence Is Shaping Medical Imaging Technology: A Survey of Innovations and Applications,” *Bioengineering (Basel)*, vol. 10, no. 12, p. 1435, Dec. 2023, doi: 10.3390/bioengineering10121435.

[2] L. Dao and N. Q. Ly, “Recent Advances in Medical Image Classification,” *ijacsa*, vol. 15, no. 7, 2024, doi: 10.14569/ijacsa.2024.0150727.

[3] F. Bougourzi and A. Hadid, “Recent Advances in Medical Imaging Segmentation: A Survey,” May 14, 2025, *arXiv*: arXiv:2505.09274. doi: 10.48550/arXiv.2505.09274.

[4] C. Albuquerque, R. Henriques, and M. Castelli, “Deep learning-based object detection algorithms in medical imaging: Systematic review,” *Heliyon*, vol. 11, no. 1, p. e41137, Jan. 2025, doi: 10.1016/j.heliyon.2024.e41137.

[5] H.-I. Liu *et al.*, “Lightweight Deep Learning for Resource-Constrained Environments: A Survey,” Apr. 12, 2024, *arXiv*: arXiv:2404.07236. doi: 10.48550/arXiv.2404.07236.

[6] “Brain Tumor MRI Dataset.” Accessed: Apr. 08, 2024. [Online]. Available: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

[7] P. Tschandl, C. Rosendahl, and H. Kittler, “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,” *Sci Data*, vol. 5, no. 1, p. 180161, Aug. 2018, doi: 10.1038/sdata.2018.161.

[8] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollar, “Designing Network Design Spaces,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA: IEEE, Jun. 2020, pp. 10425–10433. doi: 10.1109/CVPR42600.2020.01044.

[9] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.

[11] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 11966–11976. doi: 10.1109/CVPR52688.2022.01167.

[12] A. G. Howard *et al.*, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” Apr. 16, 2017, *arXiv*: arXiv:1704.04861. Accessed: Mar. 31, 2024. [Online]. Available: http://arxiv.org/abs/1704.04861

[13] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT: IEEE, Jun. 2018, pp. 4510–4520. doi: 10.1109/CVPR.2018.00474.

[14] A. Howard *et al.*, “Searching for MobileNetV3,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South): IEEE, Oct. 2019, pp. 1314–1324. doi: 10.1109/ICCV.2019.00140.

[15] D. Qin *et al.*, “MobileNetV4: Universal Models for the Mobile Ecosystem,” in *Computer Vision – ECCV 2024*, vol. 15098, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., in Lecture Notes in Computer Science, vol. 15098. , Cham: Springer Nature Switzerland, 2025, pp. 78–96. doi: 10.1007/978-3-031-73661-2_5.

[16] X. Zhang, X. Zhou, M. Lin, and J. Sun, “ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT: IEEE, Jun. 2018, pp. 6848–6856. doi: 10.1109/CVPR.2018.00716.

[17] M. Tan and Q. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” in *Proceedings of the 36th International Conference on Machine Learning*, PMLR, May 2019, pp. 6105–6114. Accessed: Mar. 31, 2024. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html

[18] S. Holagannavar, S. Ranjanagi, A. Mastiholi, Vinod, and P. Patil, “A Novel Light Weight CNN Model for Brain Tumor Classification,” in *2024 5th International Conference for Emerging Technology (INCET)*, May 2024, pp. 1–6. doi: 10.1109/INCET61516.2024.10593001.

[19] T. Tuncer, P. D. Barua, I. Tuncer, S. Dogan, and U. R. Acharya, “A lightweight deep convolutional neural network model for skin cancer image classification,” *Applied Soft Computing*, vol. 162, p. 111794, Sep. 2024, doi: 10.1016/j.asoc.2024.111794.

[20] S. A. Opee, A. A. Eva, A. T. Noor, S. M. Hasan, and M. F. Mridha, "ELW-CNN: An extremely lightweight convolutional neural network for enhancing interoperability in colon and lung cancer identification using explainable AI," *Healthcare Technology Letters*, vol. 12, no. 1, p. e12122, 2025, doi: 10.1049/htl2.12122.

[21] K. R. Reddy and R. Dhuli, "A Novel Lightweight CNN Architecture for the Diagnosis of Brain Tumors Using MR Images," *Diagnostics (Basel)*, vol. 13, no. 2, p. 312, Jan. 2023, doi: 10.3390/diagnostics13020312.

[22] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," presented at the International Conference on Learning Representations, Oct. 2020. Accessed: Mar. 31, 2024. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[23] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Mar. 31, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[24] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, Jul. 2021, pp. 10347–10357. Accessed: Mar. 31, 2024. [Online]. Available: https://proceedings.mlr.press/v139/touvron21a.html

[25] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. J. Oh, "Rethinking Spatial Dimensions of Vision Transformers".

[26] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer: Self-Attention with Linear Complexity," Jun. 14, 2020, *arXiv*: arXiv:2006.04768. doi: 10.48550/arXiv.2006.04768.

[27] V.-C. Lungu-Stan, D.-C. Cercel, and F. Pop, "SkinDistilViT: Lightweight Vision Transformer for Skin Lesion Classification," in *Artificial Neural Networks and Machine Learning – ICANN 2023*, L. Iliadis, A. Papaleonidas, P. Angelov, and C. Jayne, Eds., Cham: Springer Nature Switzerland, 2023, pp. 268–280. doi: 10.1007/978-3-031-44207-0_23.

[28] F. J. P. Montalbo, L. R. T. Hernandez, L. P. Palad, R. C. Castillo, A. S. Alon, and A. L. P. De Ocampo, "Performance Analysis of Lightweight Vision Transformers and Deep Convolutional Neural Networks in Detecting Brain Tumors in MRI Scans: An Empirical Approach," in *Proceedings of the 2023 8th International Conference on Biomedical Imaging, Signal Processing*, in ICBSP '23. New York, NY, USA: Association for Computing Machinery, Jan. 2024, pp. 17–25. doi: 10.1145/3634875.3634878.

[29] I. Pacal, "MaxCerVixT: A novel lightweight vision transformer-based Approach for precise cervical cancer detection," *Knowledge-Based Systems*, vol. 289, p. 111482, Apr. 2024, doi: 10.1016/j.knosys.2024.111482.

[30] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoAtNet: Marrying Convolution and Attention for All Data Sizes," presented at the Advances in Neural Information Processing Systems, Nov. 2021. Accessed: Mar. 31, 2024. [Online]. Available: https://openreview.net/forum?id=dUk5Foj5CLf

[31] C. Yang *et al.*, "MOAT: Alternating Mobile Convolution and Attention Brings Strong Vision Models," presented at the The Eleventh International Conference on Learning Representations, Sep. 2022. Accessed: Mar. 31, 2024. [Online]. Available: https://openreview.net/forum?id=H0HGljkxQFN

[32] Z. Tu *et al.*, "MaxViT: Multi-axis Vision Transformer," in *Computer Vision – ECCV 2022*, vol. 13684, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., in Lecture Notes in Computer Science, vol. 13684. , Cham: Springer Nature Switzerland, 2022, pp. 459–479. doi: 10.1007/978-3-031-20053-3_27.

[33] H. S. EL-Assiouti, H. El-Saadawy, M. N. Al-Berry, and M. F. Tolba, "CTRL-F: Pairing convolution with transformer for image classification via multi-level feature cross-attention and

representation learning fusion," *Engineering Applications of Artificial Intelligence*, vol. 156, p. 111076, Sep. 2025, doi: 10.1016/j.engappai.2025.111076.

[34] S. Mehta and M. Rastegari, "MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer," presented at the International Conference on Learning Representations, Oct. 2021. Accessed: Mar. 31, 2024. [Online]. Available: https://openreview.net/forum?id=vh-0sUt8HlG

[35] S. Mehta and M. Rastegari, "Separable Self-attention for Mobile Vision Transformers," Jun. 06, 2022, *arXiv*: arXiv:2206.02680. Accessed: Mar. 31, 2024. [Online]. Available: http://arxiv.org/abs/2206.02680

[36] P. K. Anasosalu Vasu, J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan, "FastViT: A Fast Hybrid Vision Transformer using Structural Reparameterization," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France: IEEE, Oct. 2023, pp. 5762–5772. doi: 10.1109/ICCV51070.2023.00532.

[37] Y. Chen *et al.*, "Mobile-Former: Bridging MobileNet and Transformer," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, Jun. 2022, pp. 5260–5269. doi: 10.1109/CVPR52688.2022.00520.

[38] O. S. EL-Assiouti, G. Hamed, D. Khattab, and H. M. Ebied, "HDKD: Hybrid data-efficient knowledge distillation network for medical image classification," *Engineering Applications of Artificial Intelligence*, vol. 138, p. 109430, Dec. 2024, doi: 10.1016/j.engappai.2024.109430.

[39] I. Pacal, O. Akhan, R. T. Deveci, and M. Deveci, "NeXtBrain: Combining local and global feature learning for brain tumor classification," *Brain Research*, vol. 1863, p. 149762, Sep. 2025, doi: 10.1016/j.brainres.2025.149762.

[40] F. He, R. Wu, X. Zeng, H. Song, G. Li, and Z. Wei, "Skin lesion classification network based on improved MobileViT," *Engineering Applications of Artificial Intelligence*, vol. 159, p. 111726, Nov. 2025, doi: 10.1016/j.engappai.2025.111726.

[41] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice: IEEE, Oct. 2017, pp. 843–852. doi: 10.1109/ICCV.2017.97.

[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.

[43] A. Trockman and J. Z. Kolter, "Patches Are All You Need?," Jan. 24, 2022, *arXiv*: arXiv:2201.09792. doi: 10.48550/arXiv.2201.09792.

[44] J. Guo *et al.*, "CMT: Convolutional Neural Networks Meet Vision Transformers," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, Jun. 2022, pp. 12165–12175. doi: 10.1109/CVPR52688.2022.01186.

[45] S. K. Datta, M. A. Shaikh, S. N. Srihari, and M. Gao, "Soft-Attention Improves Skin Cancer Classification Performance," Jun. 04, 2021, *arXiv*: arXiv:2105.03358. doi: 10.48550/arXiv.2105.03358.

[46] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. Accessed: Mar. 31, 2024. [Online]. Available: http://arxiv.org/abs/1412.6980

[47] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019. Accessed: Mar. 31, 2024. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7