Leveraging Modern Vision Architectures and Transfer Learning for Medical Image Analysis

Do Tran Dang Khoa*, Thai Nguyen Tri†
FPT University, Ho Chi Minh City, Vietnam
*First Author, Main Contributor. Email: dotrandangkhoa1105@gmail.com
†Co-Author

***Abstract*—Medical image analysis is an integral field that plays a role in modern healthcare by enabling early detection, diagnosis, and treatment planning. Traditional methods for image analysis are often labor-intensive and time-consuming, thus limiting their scalability and responsiveness. In contrast, deep learning techniques have demonstrated their efficacy in medical image analysis by automatically featuring extraction and improving pattern recognition in visual data. Besides, the development of modern architectures—such as ResNet, DenseNet, Vision Transformers (ViTs), and ConvNeXt—which has the ability to train and learn on large-scale datasets, has further enhanced the potential of transfer learning. These models can acquire rich, complex representations that can be effectively adapted to the nuances of medical imaging tasks, offering promising potential for more accuracy and efficiency.**

Throughout this study, we investigate the adaptation of modern vision architectures for medical image classification tasks using transfer learning. Our approach leverages pre-trained models on large-scale datasets. It fine-tunes them on multiple small-scale medical imaging datasets, where each dataset corresponds to a distinct imaging modality, such as CT scans, histopathological images, or MRI. This allows us to assess the model's ability to extract meaningful features and generalize to domain-specific challenges. The experiments show that the results are highly impressive, achieving an average accuracy of 97.74% without requiring deep modifications to model architectures and without extensive hyperparameter tuning, and only apply a simple transfer learning strategy based on scheduling the learning rate and applying early stopping with dual monitoring. We hope that these findings underscore the potential of modern vision architectures and transfer learning in medical imaging, offering valuable insights into their optimal selection and adaptation for real-world healthcare applications.

I. Introduction

A. Medical Image Analysis

Medical image analysis has become a crucial research domain with profound implications for diagnostic accuracy and patient care [11]. The increasing availability of imaging methods such as CT scans, MRI, and X-rays (see Figure 1) has led to a growing demand for reliable image classification methods [4]. However, traditional manual analysis faces significant challenges. Not only is it time-consuming, but it also relies heavily on the expertise of highly trained specialists, whose availability is often limited due to workforce constraints. The process requires a deep understanding of complex imaging data, frequently necessitating years of specialized training and ongoing professional development to stay updated with evolving diagnostic criteria. These inherent limitations not only slow the diagnostic process but also increase the likelihood of inconsistency in patient care. Moreover, medical image classification presents an additional challenge, requiring the careful selection of methods and techniques to effectively leverage image processing and pattern recognition outputs while ensuring that classification results align with medical expert knowledge [6].

Artificial intelligence addresses these issues by emerging as a transformative solution. At its core, machine learning—a dynamic branch of AI—has the potential to eliminate the need for explicit programming by learning from vast amounts of historical data. Puttagunta and Ravi (2021) [10] show that deep learning enables algorithms to discern subtle patterns and relationships within complex medical imaging datasets, which serve as a crucial foundation for clinical decision-making, and demonstrates that these algorithms have achieved significant success in analyzing medical images through classification, detection, and segmentation tasks. CNN-based models, in particular, have shown exceptional performance in medical image classification and detection tasks, making these models the most widely adopted approach. These sophisticated networks could potentially extract hierarchical features automatically, capturing both low-level textures and high-level semantic representations, which in turn might lead to improved classification accuracy. By leveraging such powerful techniques in medical imaging, machine learning not only enhances diagnostic precision but also accelerates the diagnostic process. Deep learning-based automation can assist radiologists by streamlining workflows, reducing human errors, and improving diagnostic accuracy.

B. Modern Vision Architectures

In recent years, these architectures: ResNet [2], DenseNet [3], Vision Transformer [1] (ViT), and ConvNeXt [5] are some of the most modern architecture models. Which are being involved by an increased number of parameters, enhanced data extraction capabilities, and rising accuracy. These contemporary models, especially by using their deep architectures and exceptional performance on large-scale data, can help them possess an extraordinary capacity to learn and represent complex features from raw inputs. Their ability to extract rich, meaningful representations establishes a robust foundation for developing advanced AI systems in diverse domains, thereby offering even greater potential for various applications.

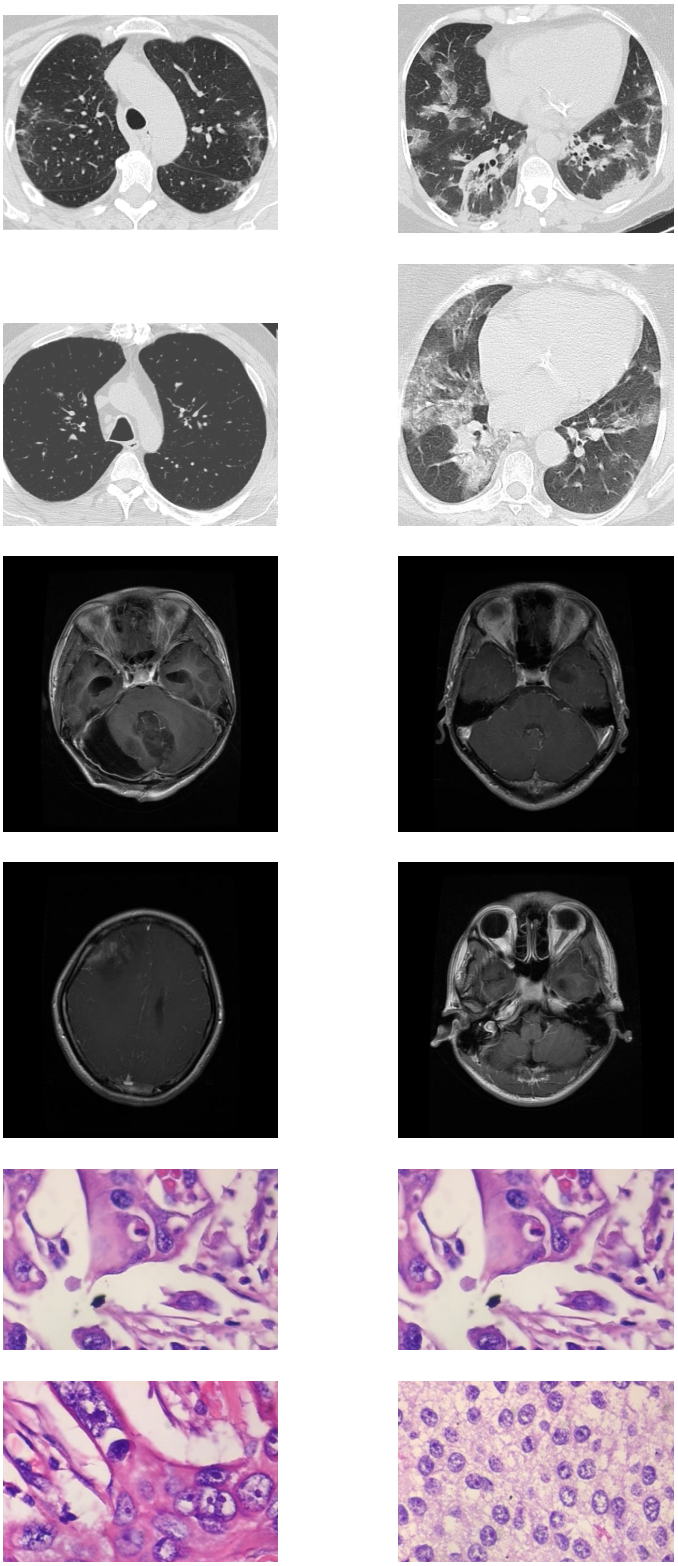


Fig. 1: Medical Images: (Top) SARS-COV-2 CT-scan images, (Middle) Brain tumor MRI images, (Bottom) Breast cancer images.

ResNet [2], with its deep residual learning framework. It effectively solves the vanishing gradient problem, enabling the training of extremely deep networks that can capture complex hierarchical features in medical images. DenseNet [3], by introducing dense connections between layers, maximizes feature reuse and gradient flow, improving both efficiency and accuracy, particularly in tasks requiring detailed texture analysis, such as tumor detection. At the same time, Vision Transformers [1](ViTs) is lean on self-learning mechanisms to capture long-range dependencies and spatial relationships in medical images, making them very effective for large-scale medical imaging datasets. ConvNeXt [5], a modernized CNN architecture inspired by Transformer models, integrates design improvements such as depthwise convolutions and normalization techniques, achieving performance on par with ViTs while maintaining the efficiency of convolutional networks. These advanced architectures collectively push the boundaries of automated medical image analysis, enabling more precise feature extraction, higher diagnostic accuracy, and greater adaptability across diverse imaging modalities.

However, even with these promising advancements, the application of such complex models in medical imaging is not easy due to challenges in managing the unique characteristics of medical data. One major issue is data scarcity. Unlike natural image datasets that often contain millions of labeled examples, medical datasets are typically small and highly specialized—often ranging from a few hundred to several thousand images. This limitation can lead to overfitting and restricts the model's ability to generalize, a problem more complicated by prevalent class imbalances where certain conditions are significantly underrepresented. Another problem is training cost, which arises from the substantial computational resources required to develop state-of-the-art deep learning models for medical image analysis. Modern architectures are highly computation, requiring high-performance hardware to be effective with processing data. To earn the optimal performance, training these models necessitates dedicated infrastructure, including high-end GPUs or TPUs, usually running for extended periods to converge effectively. This computational burden results in high operational costs, both in terms of energy consumption and hardware maintenance, which can be prohibitive for smaller research institutions or healthcare facilities with limited budgets.

C. Addressing the Challenges with Modern Vision Architectures and Transfer Learning

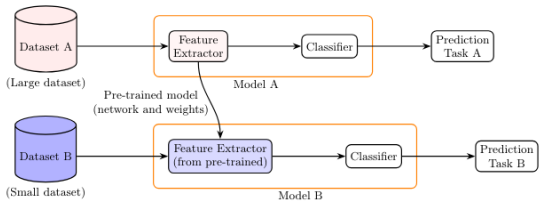


Fig. 2: Transfer Learning

To fully harness the capabilities of modern vision architectures with small and medium datasets, transfer learning has become the best choice as a powerful and widely adopted paradigm for adapting pre-trained models to domain-specific tasks. This approach is particularly advantageous in scenarios where collecting large-scale labeled datasets is impractical or resource-intensive. Vision architecture models initially trained on extensive datasets, such as ImageNet, acquire rich, hierarchical feature representations that generalize well across various tasks, forming a strong foundation for downstream applications. Fine-tuning these pre-trained models on smaller, domain-specific datasets refines their learned representations while preserving fundamental feature extraction capabilities. This process enables special things in the target domain while retaining generalizable knowledge from broader datasets. Consequently, transfer learning enhances model generalization, mitigates overfitting with small and medium datasets, and accelerates convergence by leveraging previously learned feature hierarchies. This is very important in medical image analysis, where annotated data is often scarce and costly to obtain, and transfer learning has proven invaluable. By leveraging prior knowledge from the bone of main models, transfer learning can better distinguish subtle variations in medical images that are critical. In medical image analysis, where annotated data is often scarce and costly to obtain, transfer learning has proven invaluable. Moreover, transfer learning almost reduces computational costs and training time by avoiding the need to train models from scratch. Given the vast resources required to develop deep networks from the ground up, leveraging pre-trained models allows researchers and practitioners to achieve state-of-the-art performance more efficiently, facilitating broader adoption of AI-driven solutions. In our study, we systematically evaluate and adapt modern vision architectures for medical image classification using transfer learning. Our approach leverages pre-trained models on large-scale datasets and fine-tunes them on multiple small-scale medical imaging datasets, each corresponding to a distinct imaging modality, such as CT scans, histopathological images, or MRI. By utilizing well-established architectures like DenseNet, ResNet, ConNeXt, and ViT, we effectively transfer high-level feature representations from general vision datasets to specialized medical imaging domains. Without extensive architectural modifications or major interventions in model structures, our approach demonstrates strong adaptability across diverse medical imaging datasets. Despite variations in dataset size, image resolution, and imaging modality, we achieve consistently high performance without relying on self-supervised fine-tuning or complex model-specific adjustments. Rather than introducing task-specific modifications or redesigning layers for each dataset, we preserve the integrity of the pre-trained architectures and capitalize on their inherent feature-extraction capabilities. Moreover, instead of exhaustive hyperparameter tuning for each model, we adopt a straightforward yet effective transfer learning strategy. By scheduling learning rate adjustments to ensure stable convergence and applying early stopping with dual monitoring to prevent overfitting, we enhance training efficiency while maintaining computational feasibility. The results highlight the robustness of this approach, achieving an impressive average accuracy of 97.74% across all datasets, demonstrating that transfer learning is a highly effective solution for data-limited medical imaging tasks without the need for extensive manual optimization. These findings underscore the potential of transfer learning in medical image analysis, paving the way for its broader application in real-world diagnostic systems, where efficient, accurate, and scalable AI-driven solutions can significantly enhance clinical decision-making.

II. Related Work

As introduced, the application of deep learning in medical image classification has attained significant attention in recent years. The rapid advancement of computational power and the availability of large-scale annotated medical datasets have facilitated the widespread adoption of convolutional neural networks (CNNs) and transfer learning techniques in this domain. Consequently, some of the studies have found the optimization of deep learning models for medical image analysis, particularly in the classification of histopathological images for breast cancer diagnosis.

A notable study by Musa Adamu Wakili et al. investigated the classification of breast cancer histopathological images using DenseNet and transfer learning techniques[7]. The authors suggested a novel model, **DenTnet**, which was built by the DenseNet architecture while addressing limitations related to computational efficiency and overfitting. By leveraging transfer learning, DenTnet was able to enhance feature extraction from histopathological images, ultimately achieving a classification accuracy of 99.28% on the BreaKHis dataset. This result represents a better performance over the baseline DenseNet model, which achieved an accuracy of only 80%. This research emphasized the effectiveness of combining transfer learning with CNNs to optimize classification performance in medical imaging.

Sadia Showkat et al. explored the classification of COVID-19 from chest X-ray images using ResNet and transfer learning techniques. By implementing batch normalization and freezing selected layers, the model earned an accuracy of 95%, along with a precision of 95. 65%, a specificity of 92. 74%, and a sensitivity of 95. 9%. This study underscores the effectiveness of transfer learning in medical image analysis, particularly in adapting deep learning architectures for new imaging modalities and disease classifications[12].

Furthermore, Omid Nejati Manzari et al. proposed a transfer learning-based approach utilizing a Vision Transformer (ViT) as the backbone model. Their study demonstrated that using the fine-tuned model could increase those accuracy results to 84%, 85. 1%, and 84. 2% in 12 datasets. One of the most important techniques was introduced as a novel patch moment changer augmentation, which fulfills the diversity and affinity of training. This enhancement strategy further highlights the potential for innovative preprocessing techniques to

improve the generalization and classification performance of the models[9].

To conclude, predecessor research has demonstrated that modern vision models, particularly when combined with transfer learning, can increase the accuracy and reliability of medical image classification. However, challenges such as dataset availability, computational cost, and overfitting remain critical areas for further investigation. Based on these findings, the purpose of this project is to have a deep exploration and a refinement of deep learning methodologies to improve the accuracy and efficiency of medical image classification models.

III. Methodology

A. Base Model

In what follows, let

$$x \in \mathbb{R}^n$$

denote the input (e.g., an image), and let

$$f_{\text{base}} : \mathbb{R}^n \to \mathbb{R}^d$$

denote a pre-trained base model that extracts features according to

$$F = f_{\text{base}}(x).$$

The function f_{base} can be viewed as a composition of L sequential transformation blocks:

$$f_{\text{base}}(x) = \Phi_L \circ \Phi_{L-1} \circ \cdots \circ \Phi_1(x).$$

Each transformation Φ_l represents a "block" in the network.

Residual Learning Framework (ResNet) [2]

ResNet introduces deep architectures by utilizing residual connections, which mitigate the vanishing gradient problem. Each residual block follows the formulation:

$$\Phi_l(x) = x + F_l(x),$$

where $F_l(x)$ represents a transformation (typically including convolution, normalization, and activation):

$$F_l(x) = W_{l,2}\sigma(W_{l,1}x).$$

This residual connection ensures stable training for very deep networks and is well-suited for extracting hierarchical features in medical images.

Dense Connectivity (DenseNet) [3]

DenseNet enhances feature propagation by introducing direct connections from any layer to all subsequent layers. Instead of simple residual addition, each layer receives concatenated feature maps from all preceding layers:

$$\tilde{x}_l = [x_0, x_1, \ldots, x_{l-1}].$$

The transformation in a Dense Block is given by:

$$\Phi_l(\tilde{x}_l) = H_l(\tilde{x}_l),$$

where $H_l(\cdot)$ includes batch normalization, ReLU activation, and a convolution operation. This architecture is particularly effective in capturing fine-grained medical image details.

Self-Attention Mechanism (Vision Transformer) [1]

ViT processes images as sequences of patches and applies self-attention to model long-range dependencies. The input is divided into N patches, each embedded as:

$$z_0^i = \text{Evec}(x_i), \quad i = 1, \ldots, N.$$

The transformer encoder applies a series of attention-based transformations:

- Multi-Head Self-Attention (MSA):

$$\tilde{z}_l = \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1}.$$

- Feed-Forward Network (MLP):

$$z_l = \text{MLP}(\text{LN}(\tilde{z}_l)) + \tilde{z}_l.$$

The final feature vector is extracted from the classification token in the last layer. This allows ViT to capture rich global relationships in anatomical structures.

Modern CNN with Transformer Principles (ConNeXt) [5]

ConNeXt refines traditional CNN architectures by incorporating transformer-style design elements while retaining convolutional efficiency. The residual block is restructured as:

$$\Phi_l(x) = x + F_l(x),$$

but with a more advanced transformation:

$$F_l(x) = W_{l,2}\,\phi(\text{LN}(W_{l,1}x)),$$

where $\text{LN}(\cdot)$ is layer normalization and $\phi(\cdot)$ is a modern activation function (e.g., GELU). This improves performance on high-resolution medical images while maintaining efficiency.

Algorithm 1: Adaptive Training with Early Stopping

Input: Number of epochs N, Monitoring threshold T, Patience p, stopping patience P, ϕ reduction factor, α Initial learning rate α

Output: Optimal parameters θ^*

Initialize key metrics:

$A_{\text{best\_train}} \leftarrow 0$

$L_{\text{best\_train\_val}} \leftarrow \infty$

$L_{\text{best\_val}} \leftarrow \infty$

$c \leftarrow 0$ $s \leftarrow 0$

for $n \leftarrow 1$ **to** N **do**

  $L_{\text{train}}^n \leftarrow$ compute_training_loss(n)

  $A_{\text{train}}^n \leftarrow$ compute_training_accuracy(n)

  $L_{\text{val}}^n \leftarrow$ compute_validation_loss(n)

  $A_{\text{val}}^n \leftarrow$ compute_validation_accuracy(n)

  if $A_{\text{train}}^n < T$ **then**

    // Monitoring Training Accuracy

    $A_{\text{best\_train}} \leftarrow \max(A_{\text{best\_train}}, A_{\text{train}}^n)$

    if $A_{\text{train}}^n > A_{\text{best\_train}}$ **then**

      $A_{\text{best\_train}} \leftarrow A_{\text{train}}^n$

      $c \leftarrow 0$

      if $L_{\text{val}}^n < L_{\text{best\_train\_val}}$ **then**

        $L_{\text{best\_train\_val}} \leftarrow L_{\text{val}}^n$

        $\theta^* \leftarrow \theta_n$

      end

    else

      $c \leftarrow c + 1$

      if $c \geq p$ **then**

        $\alpha \leftarrow \alpha \times \phi$

        $c \leftarrow 0$

        $s \leftarrow s + 1$

      end

    end

  else

    // Monitoring Validation Loss

    $L_{\text{best\_val}} \leftarrow \min(L_{\text{best\_val}}, L_{\text{val}}^n)$

    if $L_{\text{val}}^n < L_{\text{best\_val}}$ **then**

      $L_{\text{best\_val}} \leftarrow L_{\text{val}}^n$

      $c \leftarrow 0$

      $\theta^* \leftarrow \theta_n$

    else

      $c \leftarrow c + 1$

      if $c \geq p$ **then**

        $\alpha \leftarrow \alpha \times \phi$

        $c \leftarrow 0$

        $s \leftarrow s + 1$

      end

    end

  end

  if $s \geq P$ **then**

    $\theta \leftarrow \theta^*$

    break

  end

end

B. Fine-tuning Procedure

Our fine-tuning strategy is designed to efficiently adapt modern vision architectures, pre-trained on large-scale datasets, for medical image classification. By leveraging transfer learning, we refine these pre-trained models to capture domain-specific features while preserving the general visual representations learned during pre-training.

The fine-tuning process involves initializing the model with pre-trained weights, employing adaptive learning rate scheduling, and implementing early stopping mechanisms to ensure stable convergence. The optimization procedure minimizes a cross-entropy loss function, which is well-suited for multi-class classification tasks. Additionally, we continuously monitor training and validation metrics to guide the learning rate adjustment and prevent overfitting.

*a) **Model Initialization***

The model architecture consists of a pre-trained base model for feature extraction, followed by a global pooling operation and a linear classification layer.

1.Base Model Structure

Let x denote the input (e.g., an image) and f_{base} denote the pre-trained base model. The base model extracts features as follows:

$$F = f_{\text{base}}(x)$$

Depending on the architecture, F can have one of the following shapes:

- For Vision Transformers (ViT):

$$F \in \mathbb{R}^{B \times T \times D}$$

where B is the batch size, T is the number of tokens , D is the embedding dimension.

- For Convolutional Neural Networks (CNNs, e.g., ResNet):

$$F \in \mathbb{R}^{B \times C \times H \times W}$$

where C is the number of channels, H and W are the height and width of the feature map.

2. Global Pooling Operation

To obtain a single feature vector for each sample, a global pooling operation is applied:

- For ViT (3D feature tensor):

$$z = \frac{1}{T} \sum_{t=1}^{T} F(t)$$

where $F(t)$ is the feature vector corresponding to the t-th token.

- For CNNs (4D feature tensor):

$$z = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{ij}$$

where F_{ij} is the feature vector in the spatial location (i, j).

In both cases, the resulting vector $z \in \mathbb{R}^d$ has dimension d, which is equal to the embedding dimension or the number of channels.

3. Classification Layer

After pooling, the feature vector z is passed through a linear classifier:

$$y = Wz + b$$

where $W \in \mathbb{R}^{K \times d}$ is the weight matrix, $b \in \mathbb{R}^K$ is the bias vector, K is the number of classes. The output $y \in \mathbb{R}^K$ represents the logits, which can then be transformed into class probabilities using the softmax function.

In the subsequent discussion, we will refer to the model as a mathematical function $f(x; \theta)$, where parameters θ represents the set of parameters that govern its behavior and determine its output based on the given input x .

*b) **Adaptive Learning Rate Scheduling***

To adapt the pretrained model to the target domain, fine-tuning is performed using cross-entropy loss, which is well-suited for multi-class classification tasks in medical imaging. Given a sample (x, y), the loss function is defined as:

$$\ell(f(x;\theta), y) = -\log \frac{\exp(f_y(x;\theta))}{\sum_k \exp(f_k(x;\theta))}$$

This loss function ensures that the model effectively learns to distinguish between different classes by maximizing the likelihood of the correct label.

The model parameters θ are updated over mini-batches from the training set using a gradient-based optimizer with a learning rate α. The parameter update for a given mini-batch is expressed as:

$$\theta \leftarrow \theta - \alpha \cdot \nabla_\theta \, \ell\big(f(x;\theta), y\big),$$

where the gradient ∇_θ is computed over the current mini-batch.

To effectively adjust the learning rate during training, we employ a dual monitoring strategy that tracks both training accuracy and validation loss. Specifically, we define two monitoring phases: one based on training accuracy and another based on validation loss. This dual-phase monitoring mechanism provides a robust framework for balancing model performance and mitigating the risk of overfitting.

Assume the training set is given by:

$$D_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{M}$$

and the validation set by:

$$D_{\text{val}} = \{(x_j, y_j)\}_{j=1}^{M'}$$

The following definitions and mechanisms are introduced:

α_n: the learning rate at epoch n (with initial $\alpha_0 = \alpha_{\text{init}}$).

T: a threshold used to decide which metric to monitor.

c: a counter for the number of consecutive epochs without significant improvement.

s: a counter for the number of times the learning rate has been reduced.

p: the number of allowed epochs without improvement (i.e., the "learning rate patience").

P: the maximum allowed number of learning rate reductions (i.e., the "stopping patience").

ϕ: the multiplicative factor used to reduce the learning rate, such that $\alpha_{\text{new}} = \alpha \times \phi$.

At the beginning of the training process, we initialize the following key metrics:

$$A_{\text{best}}^{\text{train}} = 0, \quad L_{\text{best(train)}}^{\text{val}} = \infty, \quad L_{\text{best}}^{\text{val}} = \infty.$$

Here, $A_{\text{best}}^{\text{train}}$ represents the highest training accuracy observed in all epochs, serving as a reference to assess model improvement. The term $L_{\text{best(train)}}^{\text{val}}$ denotes the lowest validation loss recorded during the phase in which training accuracy is the monitoring criterion, while $L_{\text{best}}^{\text{val}}$ refers to the lowest validation loss observed when validation loss itself is the monitoring metric.

By continually tracking the current training accuracy A_n^{train} and validation loss L_n^{val}, we update $A_{\text{train}}^{\text{best}}$ as the highest observed training accuracy and $L_{\text{val}}^{\text{best}}$ as the lowest recorded validation loss. Building on this, in each epoch n (with $n = 1, 2, \ldots, N$, where N denotes the total number of epochs), we compute key metrics to monitor the training process. Specifically, we calculate :

Training Loss:

$$L_n^{\text{train}} = \frac{1}{M} \sum_{i=1}^{M} \ell\big(f(x_i; \theta_n), y_i\big)$$

Training Accuracy:

$$A_n^{\text{train}} = \frac{1}{M} \sum_{i=1}^{M} \mathbf{1}\Big\{ \operatorname{argmax}_k \ f_k(x_i; \theta_n) = y_i \Big\}$$

Validation Loss:

$$L_n^{\text{val}} = \frac{1}{M'} \sum_{j=1}^{M'} \ell\big(f(x_j; \theta_n), y_j\big)$$

Validation Accuracy:

$$A_n^{\text{val}} = \frac{1}{M'} \sum_{j=1}^{M'} \mathbf{1}\Big\{ \operatorname{argmax}_k \ f_k(x_j; \theta_n) = y_j \Big\}$$

Then, , we adapt the learning rate based on the monitored metrics using the following update rules:

a. When $A_n^{\text{train}} < T$ (Monitoring Training Accuracy)

The best training accuracy from previous epochs is:

$$A_{\text{best}}^{\text{train}} = \max\{A_1^{\text{train}}, \ldots, A_{n-1}^{\text{train}}\}$$

- If $A_n^{\text{train}} > A_{\text{best}}^{\text{train}}$, update:

$$A_{\text{best}}^{\text{train}} \leftarrow A_n^{\text{train}}, \quad c \leftarrow 0.$$

If the current validation loss satisfies $L_n^{\text{val}} < L_{\text{best(train)}}^{\text{val}}$, update:

$$L_{\text{best(train)}}^{\text{val}} \leftarrow L_n^{\text{val}}, \quad \theta^* \leftarrow \theta_n.$$

- Otherwise ($A_n^{\text{train}} < A_{\text{best}}^{\text{train}}$):

Increment $c \leftarrow c + 1$.

If $c \geq p$, we reduce the learning rate:

$$\alpha_{n+1} = \alpha_n \times \phi,$$

then reset $c \leftarrow 0$, and increment $s \leftarrow s + 1$.

b. When $A_n^{\text{train}} \geq T$ (Monitoring Validation Loss)

The best validation loss from previous epochs is:

$$L_{\text{best}}^{\text{val}} = \min\{L_1^{\text{val}}, \ldots, L_{n-1}^{\text{val}}\}$$

- If the current validation loss satisfies $L_n^{\text{val}} < L_{\text{best}}^{\text{val}}$, update:

$$L_{\text{best}}^{\text{val}} \leftarrow L_n^{\text{val}}, \quad c \leftarrow 0, \quad \theta^* \leftarrow \theta_n.$$

- Otherwise ($L_n^{\text{val}} \geq L_{\text{best}}^{\text{val}}$):

Increment $c \leftarrow c + 1$.

If $c \geq p$, we reduce the learning rate:

$$\alpha_{n+1} = \alpha_n \times \phi,$$

then reset $c \leftarrow 0$, and increment $s \leftarrow s + 1$.

c. Early Stopping: If $s \geq P$, terminate training and restore best checkpoint:

$$\theta \leftarrow \theta^*, \quad k^* = \operatorname{argmin}_k L_k^{\text{val}}.$$

This adaptive strategy enables efficient convergence by dynamically adjusting the learning rate based on observed improvements, preventing overfitting while ensuring optimal performance.

IV. Experiments

A. Experiment Setup

1) Datasets

To rigorously evaluate the transfer learning framework across diverse medical imaging tasks, three distinct datasets were employed. The selection criteria were guided by the need to encompass variability in imaging modalities, disease characteristics, and data scale , all dataset shown in Table I).

Datasets	Soares et al. [13]	Spanhol et al. [14]	Nickparvar et al. [8]
Disease	SARS-CoV-2	BreaKHis	Brain Tumor
Number of Images	2,482	1,693	7,023
Number of Classes	2	2	4
Imaging Modality	CT Scan	Microscopic Biopsy	MRI

TABLE I: Summary of Medical Imaging Datasets

SARS-CoV-2 Dataset (Soares et al. [13]): This dataset consists of 2,482 computed tomography (CT) scan images, including 1,252 scans from patients diagnosed with SARS-CoV-2 and 1,230 scans from non-infected individuals. These scans were collected from real patients in hospitals in São Paulo, Brazil, providing a valuable dataset for studying imaging patterns associated with SARS-CoV-2.

Break Cancer Dataset (Spanhol et al. [14]): The original dataset comprises 7,909 histopathology images of breast cancer, collected from 82 patients, with images categorized into two classes: benign and malignant. The dataset serves as a benchmark for evaluating machine learning models in breast cancer diagnosis. For this study, a subset of 1,693 images was

selected, focusing only on samples captured at 400× optical zoom. This selection ensures a consistent imaging scale while maintaining the dataset's relevance for automated classification tasks.

Brain Tumor Dataset (Nickparvar et al. [8]): This dataset comprises 7,023 magnetic resonance imaging (MRI) scans categorized into four classes. It integrates data from three sources, including MRI scans from 20 subjects diagnosed with glioblastoma. Each subject has two MRI exams: one within 90 days after completing concomitant chemo-radiation therapy (CRT) and another at tumor progression. The dataset includes various imaging modalities such as T1-weighted (pre- and post-contrast), FLAIR, T2-weighted, ADC, and perfusion images derived from dynamic susceptibility contrast (DSC) imaging.

The diversity in dataset characteristics—notably in terms of image modality, class distribution, and sample size—ensures a comprehensive evaluation of the transfer learning strategy. By employing these datasets, the study aims to benchmark the adaptability and generalization capabilities of modern vision architectures when confronted with varying medical imaging challenges.

2) Hyperparameters and Training Configuration

The training protocol is standardized in all experiments to ensure a fair evaluation of the transfer learning capabilities of the selected models. The following hyperparameter are settings:

Epochs: The model is trained for 10 epochs on the brain tumor dataset, and 20 epochs on the SARS-CoV-2 and BreaKHis datasets . This selection was made to balance training time with the potential for convergence, considering the varying complexities and sizes of the datasets.

Learning Rate: An initial learning rate $\alpha_{\text{init}} = 0.001$ is used for all CNN models. For the ViT architecture, we experiment with two learning rate $\alpha_{\text{init}} = 0.001$ like CNNs and $\alpha_{\text{init}} = 1 \times 10^{-5}$ to assess the influence of the starting learning rate on model convergence and performance.

Dual Monitoring Fine-tuning Parameters: We set the threshold $T = 0.9$ to monitor the progress of the training. The learning rate patience is set to $p = 2$, meaning that if there is no improvement for two consecutive epochs,a reduction in the learning rate is triggered. The stopping patience is set as $P = 2$, limiting the total number of allowed learning rate reductions before stop training. The learning rate is reduce by a multiplicative factor of $\phi = 0.5$.

Data Splitting: The dataset is partitioned into training, validation, and test sets using a split ratio of [0.8, 0.1, 0.1].

Batch Size: Batch size 32 is used to efficiently balance memory constraints and training speed.

Model Parameter Freezing: All models are fine-tuning across the entire network, meaning that all parameters were trainable during the fine-tuning process.

Device Configuration: All the experiments are conduct on a computational setup NVIDIA Tesla P100 GPU, with 3,584 CUDA cores and 16GB of GDDR6 VRAM. The CPU is a single-core, hyper-threaded Intel Xeon processor (1 core, 2 threads) clocked at 2.2 GHz, without Turbo Boost. The system memory is 16GB RAM, and the disk storage amounts is 155GB.

This configuration by experiment show that the model sufficient escape potential local minima while prevente overfitting during training. The hyperparameter choices are maintain consistently to accommodate the diverse sizes and complexities of the datasets, ensuring that the evaluation remains comparable across different experimental conditions.

B. Metrics

We report Accuracy (ACC), Precision, Recall, and F1-score as the standard evaluation metrics for evaluation models performance in classification task. Accuracy is a threshold-based metric that quantifies the overall proportion of correct predictions among all samples. The formula for Accuracy is given by:

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

Precision evaluate the ratio of true positives to all positive predictions, emphasizing the model's ability to avoid false alarms. Its formula is defined as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

The recall measures the ratio of true positives to the total number of actual positive instances,thereby assessing the model's capability to capture all relevant cases. The formula for Recall is:

$$\text{Recall} = \frac{TP}{TP + FN}$$

The F1-score, defined as the harmonic mean of Precision and Recall, provide a balance metric that accounts for both false positives and false negatives. It is expressed as:

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

C. Result

Training Processs

As illustrated in Figures 3a, 3c, and 3b, the training process for all models across the different datasets reveal several observations :

First, models leveraging transfer learning consistently demonstrate a robust learning capacity and converge within just a few epochs. Notably, these models do not exhibit sign of overfitting—even though we let all parameters are updated during training—indicating that the use of modern architectures in conjunction with transfer learning can lead to strong generalization capabilities across diverse datasets.

Second, when comparing CNN-based and Transformer-based models under the same transfer configuration (i.e., identical learning rates and minimal architectural modifications),

the CNN-based models tend to converge more effectively than the Vision Transformer (ViT). This difference likely comes from the way CNNs are designed, making them fine-tune better by focusing on local features. ViT, on the other hand, depends on global attention, which seems to need more data to work at full potential. Models like DenseNet, ResNet, and ConvNeXt show strong feature learning in transfer learning, leading to steadier training and better results.

Lastly, in further investigation , we observe that reduce the learning rate for ViT (e.g., to 1×10^{-5}, referred to as ViT*) can yield convergence behavior comparable to that of CNN-based models. This observation suggests that a more gradual optimization process allows ViT to effectively learn global relationships within the data without being adversely affected by large gradient updates. By employing a smaller learning rate, attention-based model can incrementally adjust its weights, thereby reducing the risk of overshooting optimal solutions in fine-tuning tasks .

Collectively, these observations demonstrate the effectiveness of transfer learning in facilitating rapid convergence and strong generalization for modern vision architectures ,especially CNN-based models. While ViT requires more careful hyperparameter tuning to achieve comparable performance, it remains a promising approach,where attention mechanism can benefit from train on sufficiently large datasets.

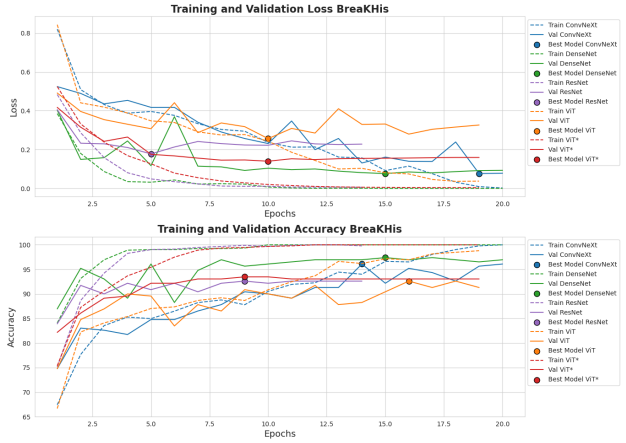
Evaluation Model

From the experiment, the results with difference metrics shown in Table II further confirm the trends observed during training. CNN-based architectures maintain high and stable accuracy across all three datasets, with DenseNet and ConvNeXt often outperforming ResNet, particularly on BreaKHis and BrainTumor . Validation and test accuracies remain closely aligned, supporting the earlier observation that these models do not suffer from excessive overfitting although fine-tuning all the model parameters. Performance metrics, including Precision, Recall, and F1-score, remain consistently high (95-99%), reinforcing the generalization ability of modern CNNs while transfer learning on small datasets.

Model ViT ,when trained with the standard learning rate of 0.001, show lower validation and test accuracy (85-92%) compare to CNNs (95-98%), which aligns with the earlier finding that ViT struggles to converge effectively on smaller datasets. For instance, on BreaKHis, the training accuracy reaches 93.46%, but validation accuracy drops to 90.00%, reflecting difficulties in optimization and possible underfitting. These results support the notion that ViT requires more data or finer hyperparameter tuning to reach its full potential.

When the learning rate reduce to 1e-5 (ViT*), we observe significant improvements, with validation and test accuracies increasing to 93-98%, making model have comparable to CNNs. Performance metrics such as Precision, Recall, and F1-score also improve (94-98%), which is consistent with the earlier observation that a lower learning rate allows ViT to refine its weight updates more effectively. Notably, on the larger BrainTumor dataset (7022 images), ViT* achieve performance on par or even slightly better than CNNs, further

(a) SARS-COV-2 dataset



(b) BreaKHis dataset

(c) Braintumor dataset

Fig. 3: Training and validation performance for different datasets

confirming that ViT benefits from larger datasets where its self-attention mechanisms can be fully leveraged .

Overall, these evaluation results reinforce observation from training analysis. CNNs remain a strong choice for smaller datasets due to their efficient feature extraction and stable convergence . ViT can initially struggles under standard learning rate but, when apply proper tuned learning rate (1e-5),

can achieve competitive performance. For datasets with fewer than 2000 images (SARS-COV-2, BreaKHis), CNNs retain a clear advantage. However, for larger datasets (>5000 images, as in BrainTumor), ViT* demonstrates potential to match or exceed CNN performance, suggesting that transformers can be effectively utilized in medical image analysis when optimization strategies are appropriately adjusted.

TABLE II: Model Performance and Classification Metrics on Medical Image Datasets

Dataset	Model	Train Acc	Val Acc	Test Acc	Precision	Recall	F1-score
SARS-COV-2	ResNet	100.0%	99.19%	96.79%	97%	97%	97%
	DenseNet	100.0%	98.39%	97.99%	98%	98%	98%
	ConvNeXt	99.65%	97.18%	97.19%	97%	97%	97%
	ViT	94.35%	86.69%	91.57%	93%	91%	91%
	ViT*	99.9%	93.55%	94.38%	94%	94%	94%
BreaKHis	ResNet	99.89%	90.87%	94.31%	93%	94%	94%
	DenseNet	100.0%	97.39%	94.50%	94%	93%	94%
	ConvNeXt	100.0%	95.65%	95.96%	95%	96%	95%
	ViT	93.46%	90.00%	84.59%	83%	81%	82%
	ViT*	99.89%	93.48%	92.29%	91%	91%	91%
Brain Tumor	ResNet	99.98%	97.32%	98.55%	98%	98%	98%
	DenseNet	100.0%	99.42%	99.47%	99%	99%	99%
	ConvNeXt	99.96%	99.53%	99.16%	99%	99%	99%
	ViT	98.13%	92.53%	88.56%	89%	88%	88%
	ViT*	100.0%	98.48%	98.4%	98%	98%	98%

V. Discussion

While our experiments have demonstrated the effectiveness of leveraging pre-trained modern vision architectures such as DenseNet, ResNet, ConNeXt, and Vision Transformer (ViT) for medical image analysis through transfer learning, several avenues for further exploration remain open. One promising direction involves experimenting with alternative transfer learning strategies beyond full fine-tuning. Techniques such as layer-wise freezing, feature extraction, and discriminative layer-wise learning rates could offer more control over training dynamics, particularly when working with small-scale datasets.

For instance, selectively freezing early convolutional layers—often responsible for capturing general low-level features—and fine-tuning only higher-level semantic layers may improve both training efficiency and generalization performance. In our current work, full fine-tuning was applied across all architectures; however, a more granular investigation into the optimal number of trainable layers may yield valuable insights. This could involve systematically varying the number of unfrozen layers from top to bottom and evaluating the resulting impact on both accuracy and training time.

Furthermore, while the current study focuses on well-established modern architectures, the rapid evolution of vision models presents an opportunity to explore emerging architectures such as ConvNeXtV2, Swin Transformers, or SAM (Segment Anything Model) in similar medical contexts. These models incorporate architectural innovations like dynamic attention mechanisms and improved hierarchical representations, which may further enhance performance, especially in complex diagnostic tasks.

Another potential extension includes hybrid transfer learning setups where knowledge from multiple source domains is aggregated or models are pre-trained on more relevant medical image datasets when available. This could improve domain adaptation and better capture the unique characteristics of biomedical data.

VI. Conclusion

In this study, we demonstrates that the advancements in medical image analysis, driven by deep learning and modern vision architectures, have yielded significant improvements in diagnostic accuracy and efficiency. By leveraging convolutional neural networks (CNNs) and state-of-the-art models like ResNet, DenseNet, Vision Transformer (ViT), and ConvNeXt, we effectively extracted hierarchical features and captured complex patterns within medical imaging data.

Our results reveal that transfer learning plays a critical role in addressing the unique challenges of medical data, such as limited dataset size and class imbalances. By fine-tuning models pretrained on large-scale datasets, we not only mitigated the risk of overfitting but also accelerated convergence, leading to more accurate and efficient diagnostic tools in classification tasks.

The empirical evidence from our experiments underscores the robustness and scalability of these modern architectures in diverse medical imaging tasks. Moving forward, continued research will be essential for further enhancing AI-driven solutions and overcoming remaining challenges in medical image analysis.

Availability

All implementation details, training configurations, and experimental results are publicly available at our GitHub repository: https://github.com/DangKhoaAI/TransferlearningMedicalImage

Acknowledgments

We would like to express our heartfelt gratitude to Mr. Do Duc Hao from FPT University for his invaluable support and insightful guidance throughout the course of this project. His expertise, encouragement, and mentorship played a important role in shaping the direction and quality of our work.

References

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv **abs/2010.11929** (2020), https://api.semanticscholar.org/CorpusID:225039882

[2] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015), https://ieeexplore.ieee.org/document/7780459

[3] Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks (2018), https://ieeexplore.ieee.org/document/8099726

[4] Kasban, H.: A comparative study of medical imaging techniques. International Journal of Information Science and Intelligent System, **4**, 37–58 (03 2015)

[5] Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s (2022), https://ieeexplore.ieee.org/document/9879745

[6] Miranda, E., Aryuni, M., Irwansyah, E.: A survey of medical image classification techniques. In: 2016 International Conference on Information Management and Technology (ICIMTech). pp. 56–61 (2016). https://doi.org/10.1109/ICIMTech.2016.7930302

[7] Musa Adamu Wakili, Harisu Abdullahi Shehu, M.H.S.M.H.U.S.A.U.H.K.I.F.I.S.U.: Classification of breast cancer histopathological images using densenet and transfer learning **2** (2022). https://doi.org/https://doi.org/10.1155/2022/8904768, https://doi.org/10.1155/2022/8904768

[8] Nickparvar, M., Schmainda, K.M., Prah, M.: Brain tumor mri dataset. The Cancer Imaging Archive (2018). https://doi.org/10.34740/KAGGLE/DSV/2645886, https://doi.org/10.7937/K9/TCIA.2018.15quzvnb, data set

[9] Omid Nejati Manzari, Hamid Ahmadabadi, H.K.S.B.S.A.A.: Medvit: A robust vision transformer for generalized medical image classification **11**(1), 19–38 (2023), https://www.sciencedirect.com/science/article/abs/pii/S0010482523002561

[10] Puttagunta, M., Ravi, S.: Medical image analysis based on deep learning approach. Multimedia Tools and Applications **80**(16), 24365–24398 (2021). https://doi.org/10.1007/s11042-021-10707-4, https://doi.org/10.1007/s11042-021-10707-4

[11] Rashed, B.M., Popescu, N.: Critical analysis of the current medical image-based processing techniques for automatic disease evaluation: Systematic literature review. Sensors **22**(18) (2022). https://doi.org/10.3390/s22187065, https://www.mdpi.com/1424-8220/22/18/7065

[12] Sadia Showkat, S.Q.: Efficacy of transfer learning-based resnet models chest x-ray image classification for detecting covid-19 pneumonia (2022). https://doi.org/https://doi.org/10.1016/j.chemolab.2022.104534, https://doi.org/10.1016/j.chemolab.2022.104534

[13] Soares, E., Angelov, P., Biaso, S., Froes, M.H., Abe, D.K.: Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification. medRxiv (2020). https://doi.org/10.1101/2020.04.24.20078584, https://www.medrxiv.org/content/early/2020/05/14/2020.04.24.20078584

[14] Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L.: A dataset for breast cancer histopathological image classification. IEEE Transactions on Biomedical Engineering **63**(7), 1455–1462 (2016). https://doi.org/10.1109/TBME.2015.2496264