Leveraging Modern Vision Model Architectures and Transfer Learning for Medical Image Analysis

Do Tran Dang Khoa* and Thai Nguyen Tri**

FPT University, Ho Chi Minh City , VietNam

Abstract. Medical imaging datasets, particularly in specialized domains like radiology and pathology, are often limited in size due to data acquisition challenges, privacy concerns, and the need for expert annotations. This limitation makes transfer learning a crucial strategy, as it enables leveraging knowledge from large-scale natural image datasets to enhance model performance on medical imaging tasks. Deep learning, particularly Convolutional Neural Networks (CNNs) and Transformer-based models, has revolutionized medical image analysis by enabling automatic feature extraction and pattern recognition from complex visual data. Recent advances in modern vision architectures, such as DenseNet , ResNet, ConvNeXt , and Vision Transformers (ViTs), have demonstrated state-of-the-art performance across various medical imaging tasks. These models, when pretrained on large-scale datasets, learn rich hierarchical representations that can be effectively transferred to domain-specific applications through transfer learning.

In this study, we extend the paradigm of transfer learning by systematically evaluating and adapting multiple modern vision architectures pretrained on large data set for medical image classification tasks. Our research is conducted on a diverse set of medical imaging datasets, covering a range of imaging modalities such as CT scans, histopathological images , and MRI. By fine-tuning these pretrained models on relatively small-scale medical datasets , we investigate their ability to extract meaningful features and generalize to domain-specific challenges. Additionally, we explore strategies to mitigate overfitting, and improve training efficiency. The findings from this study contribute to the growing body of research on deep learning for medical imaging, offering insights into the optimal selection and adaptation of modern vision architectures for real-world healthcare applications.

1 Introduction

1.1 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 modalities such as CT scans, MRI, and X-rays (see Figure 1) has

* First author. Main contributor.

** Co-author.

led to a growing demand for efficient and reliable image interpretation 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 down the diagnostic process but also increase the likelihood of inconsistencies in patient care. Moreover, medical image classification presents additional challenges, 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].

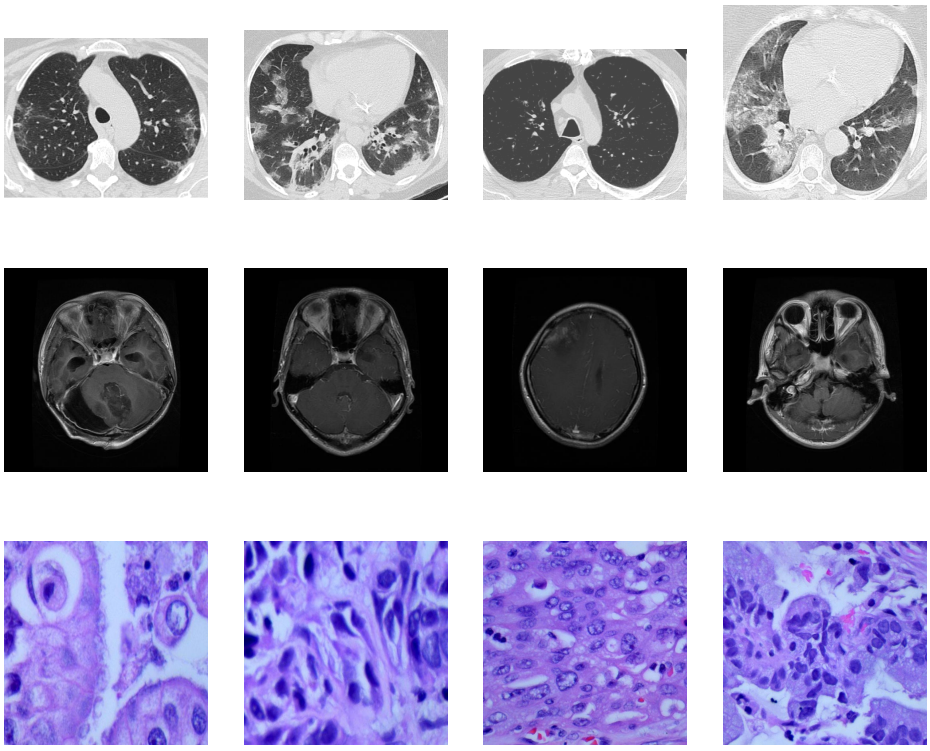


Fig. 1: Medical Images: (Top) SARS-COV-2 CT-scan images, (Middle) Brain tumor MRI images, (Bottom) Lung cancer histopathological images.

To address these issues, artificial intelligence could emerge 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 them 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, machine learning may not only enhance diagnostic precision but also accelerate the diagnostic process considerably. Deep learning-based automation can assist radiologists by streamlining workflows, reducing human errors, and facilitating decision-making for less experienced practitioners.

1.2 Modern Vision Architectures

In parallel, modern architectures such as ResNet [2], DenseNet [3], Vision Transformer [1] (ViT), and ConvNeXt [5] are evolving with an increasing number of parameters, enhanced data extraction capabilities, and rising accuracy. These state-of-the-art models, characterized by their deep architectures and exceptional performance on large-scale data, 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. ResNet [2], with its deep residual learning framework, effectively addresses 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. Meanwhile, Vision Transformers [1](ViTs) leverage self-attention mechanisms to capture long-range dependencies and spatial relationships in medical images, making them highly effective for large-scale medical imaging datasets where global context is critical. 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, despite these promising advancements, the application of such sophisticated models in medical imaging is not straightforward 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 increases the risk of overfitting and restricts the model's ability to generalize, a problem compounded 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 inherently computationally intensive, requiring high-performance hardware to efficiently process data. To achieve optimal performance, training these models necessitates dedicated infrastructure, including high-end GPUs or TPUs, often 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.

1.3 Addressing the Challenges with Modern Vision Architectures and Transfer Learning

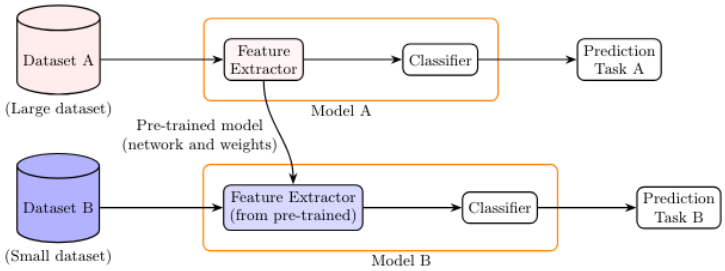


Fig. 2: Transfer Learning

To leverage this potential, transfer learning has emerged as a promising approach whereby models pretrained on large-scale datasets, are fine-tuned on smaller, domain-specific datasets. This strategy not only significantly enhances generalization and accelerates convergence but also facilitates the integration of AI-driven solutions into clinical workflows . In our study, we systematically investigate the performance of these advanced architectures on a range of medical imaging tasks with varying complexities and dataset sizes. The empirical evidence gathered from these experiments offers practical guidelines for selecting and adapting pretrained models, ultimately contributing to the development of more robust and scalable AI-driven diagnostic tools for medical applications.

In our study, by leveraging these modern vision architectures, we aim to provide a comprehensive evaluation of their effectiveness through systematic experiments on diverse medical datasets. Transfer learning in image analysis is

widely adopted for its ability to improve performance on tasks with limited labeled data while reducing computational costs .By leveraging pretrained modern vision architectures, our study aims to provide a comprehensive evaluation of their effectiveness in various medical imaging tasks . Through systematic transfer learning experiments on diverse medical datasets, we seek to improve diagnostic accuracy, optimize model selection, and facilitate the integration of AI-driven solutions into clinical workflows.

1.4 Paper Structure

Section 2: Related Work provides a comprehensive review of prior research on medical image classification, focusing on the contributions of deep learning models, particularly Convolutional Neural Networks (CNNs) and Transformer-based architectures. It discusses the impact of transfer learning in improving model performance across various medical imaging applications.

Section 3: Methodology involves evaluating modern vision architectures, including ResNet, DenseNet, Vision Transformers (ViTs), and ConvNeXt. It details the fine-tuning strategy, adaptive learning rate scheduling, and key mathematical formulations governing the training process.

Section 4: Experiments describes the experimental setup, including dataset selection, hyperparameter configurations, and training protocols. It also outlines the performance evaluation metrics and comparative analysis of different architectures.

Section 5: Discussion The results are analyzed to compare the effectiveness of CNNs and Transformer-based models in medical imaging. Key challenges, such as data scarcity, computational overhead, and overfitting, are discussed, along with potential strategies for mitigating these issues.

Section 6: Conclusion The study highlights the effectiveness of transfer learning in medical image classification and suggests future research directions. These include exploring hybrid architectures that integrate CNNs and Transformers to further improve performance in medical imaging tasks.

2 Releted Work

As outlined in the introduction, the application of deep learning in medical image classification has gained 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, numerous studies have explored 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 proposed a novel model, **DenTnet**, which builds

upon 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 substantial improvement over the baseline DenseNet model, which achieved an accuracy of only 80%. The study highlights the efficacy of integrating transfer learning with CNNs to optimize classification performance in medical imaging.

Another relevant study was conducted by Sadia Showkat et al., who 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 achieved 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].

Additionally, Omid Nejati Manzari et al. proposed a transfer learning-based approach utilizing a Vision Transformer (ViT) as the backbone model. Their study demonstrated remarkable accuracy results of 84%, 85. 1%, and 84. 2% in 12 datasets. A key contribution of their work was they introduced a novel patch moment changer augmentation, which enriched 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].

In summary, prior research has demonstrated that CNN-based models, particularly when combined with transfer learning, can significantly improve the accuracy and reliability of medical image classification. However, challenges such as dataset availability, computational cost, and overfitting remain critical areas for further investigation. Building upon these findings, this study aims to further explore and refine deep learning methodologies to enhance the accuracy and efficiency of medical image classification models.

3 Methodology

3.1 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:

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

where $\mathrm{LN}(\cdot)$ 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**
  $L^n_{\text{train}} \leftarrow$ compute_training_loss(n)
  $A^n_{\text{train}} \leftarrow$ compute_training_accuracy(n)
  $L^n_{\text{val}} \leftarrow$ compute_validation_loss(n)
  $A^n_{\text{val}} \leftarrow$ compute_validation_accuracy(n)
  if $A^n_{train} < T$ **then**
    `// Monitoring Training Accuracy`
    $A_{\text{best\_train}} \leftarrow \max(A_{\text{best\_train}}, A^n_{\text{train}})$
    if $A^n_{train} > A_{best\_train}$ **then**
      $A_{\text{best\_train}} \leftarrow A^n_{\text{train}}$
      $c \leftarrow 0$
      if $L^n_{val} < L_{best\_train\_val}$ **then**
        $L_{\text{best\_train\_val}} \leftarrow L^n_{\text{val}}$
        $\theta^* \leftarrow \theta_n$
      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`
    $L_{\text{best\_val}} \leftarrow \min(L_{\text{best\_val}}, L^n_{\text{val}})$
    if $L^n_{val} < L_{best\_val}$ **then**
      $L_{\text{best\_val}} \leftarrow L^n_{\text{val}}$
      $c \leftarrow 0$
      $\theta^* \leftarrow \theta_n$
    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

3.2 Fine-tuning Procedure

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

The fine-tuning process involves initializing the model with pretrained 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.

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 .

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:

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

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

4 Experiments

4.1 Experiment Setup

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 1).

Datasets	Soares et al. [13]	Spanhol et al. [14]	Nickparvar et al. [8]
Disease	SARS-CoV-2	Breast Tumor	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 1: 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. The images are grayscale with dimensions ranging from 275 to 400 pixels. 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. Image dimensions range from 200 to 1,000 pixels .

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.

Hyperparameters and Training Configuration The training protocol is standardized across all experiments to ensure a fair evaluation of the transfer learning capabilities of the selected models. We adopt the following hyperparameter settings:

Epochs: The model is trained for 10 epochs on the lung cancer and brain tumor datasets, and 30 epochs on the COVID-19 dataset. 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, experiments are conducted with both $\alpha_{\text{init}} = 0.001$ and $\alpha_{\text{init}} = 1 \times 10^{-5}$ to assess the influence of the starting learning rate on model convergence and performance.

Scheduling 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 significant improvement for two consecutive epochs, a reduction in the learning rate is triggered. The stopping patience is also defined as $P = 2$, limiting the total number of allowed learning rate reductions. Upon meeting these conditions, the learning rate is reduced 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: A batch size of 32 is used to efficiently balance memory constraints and training speed.

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

Device Configuration: All experiments were conducted on a computational setup featuring an NVIDIA Tesla P100 GPU, equipped with 3,584 CUDA cores and 16GB of GDDR6 VRAM. The system is powered by a single-core, hyper-threaded Intel Xeon processor (1 core, 2 threads) clocked at 2.2 GHz, without Turbo Boost, and featuring a 56MB cache. The available system memory is 16GB RAM, while the available disk storage amounts to around 155GB.

This configuration is chosen to allow the model sufficient opportunity to escape potential local minima while preventing prolonged stagnation during training. The parameter choices are maintained consistently to accommodate the diverse sizes and complexities of the datasets, ensuring that the evaluation remains comparable across different experimental conditions.

4.2 Metrics

We report Accuracy (ACC), Precision, Recall, and F1-score as the standard evaluation metrics. 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 evaluates the ratio of true positives to all positive predictions, emphasizing the model's ability to avoid false alarms. Its formula is defined as:

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

In contrast, 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, provides a balanced metric that accounts for both false positives and false negatives. It is expressed as:

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

4.3 Result

Training Processs

As illustrated in Figures 3a, 3c, and 3b the training process for all models across the different datasets reveals several noteworthy 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 signs of overfitting—even though 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.

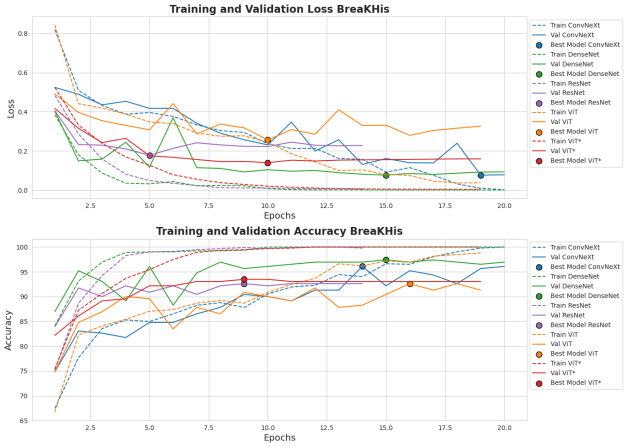
Second, when comparing CNN-based and Transformer-based models under the same transfer configuration (i.e., identical learning rates and minimal architectural modifications), the CNN-based models tend to converge more effectively than the Vision Transformer (ViT). This disparity is likely attributable to the inductive biases inherent to CNNs, which facilitate more efficient fine-tuning by leveraging local feature extraction. In contrast, ViT, which relies on global attention mechanisms, appears to require larger datasets to fully harness its potential. Architectures such as DenseNet, ResNet, and ConvNeXt demonstrate superior representational power in transfer learning scenarios, leading to more stable convergence and better performance metrics over the training epochs.

Lastly, further investigation shows that reducing 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

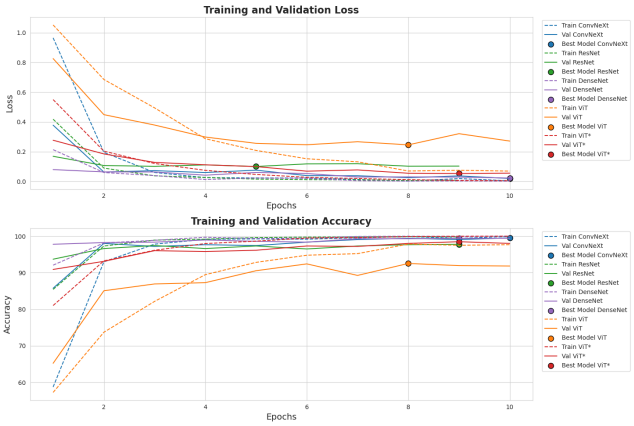
the data without being adversely affected by large gradient updates. By employing a smaller learning rate, the model can incrementally adjust its weights, thereby reducing the risk of overshooting optimal solutions during training.

Collectively, these observations emphasize the importance of appropriate hyperparameter tuning when applying transfer learning to various model architectures. They also underscore that while CNN-based models may exhibit more efficient convergence due to their inherent design advantages, Transformer-based models can achieve comparable performance with carefully adjusted learning rates, thereby highlighting their potential applicability in medical image analysis tasks.

(a) SARS-COV-2 dataset



(b) BreaKHis dataset



(c) Braintumor dataset

Fig. 3: Training and validation performance for different datasets

Evaluation Model

The accuracy of models in training, validation, and test sets, as shown in Table 2 demonstrates the effectiveness of transfer learning with modern architectures, including both CNN-based models (ConvNeXt, DenseNet, ResNet) and attention-based models (ViT, ViT*).

Overall Performance

Across all datasets, models exhibit high performance across all metrics—accuracy, precision, recall, and F1-score—on the training, validation, and test sets. The average performance is indicative of the robustness of transfer learning when applied to modern architectures. For instance, DenseNet consistently achieves superior results, with an average test accuracy of approximately 98.82%, while ConvNeXt and ResNet follow closely with average test accuracies of 98.45% and 97.84%, respectively. Additionally, ViT*, which benefits from a lower learning rate, shows notable improvements over the original ViT, underlining the importance of fine-tuning.

In the SARS-COV-2 and Brain Tumor MRI datasets, CNN-based models such as DenseNet and ConvNeXt dominate with near-perfect performance, while ResNet maintains competitive accuracy. The original ViT model, which relies on global attention mechanisms, initially struggles with feature extraction—likely due to its dependency on larger datasets. However, when fine-tuned (as in ViT*), the performance significantly improves, aligning its metrics closer to those of the CNN-based models. These results underscore the capacity of transfer learning to leverage pre-trained models for effective feature representation in medical imaging tasks.

In summary, the comprehensive evaluation across multiple metrics confirms that transfer learning with modern architectures is highly effective for medical image analysis. CNN-based models exhibit efficient convergence and robust performance, while Transformer-based models can achieve comparable results with appropriate hyperparameter tuning. Nevertheless, the exceptional results observed in the Lung Cancer dataset necessitate further investigation to address potential data leakage and ensure the validity of the experimental findings.

Table 2: 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%

5 Discussion

6 Conclusion

In summary, our study 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 task.

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.

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 (2021), https://arxiv.org/abs/2010.11929
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015), https://arxiv.org/abs/1512.03385

3. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks (2018), https://arxiv.org/abs/1608.06993
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://arxiv.org/abs/2201.03545
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://doi.org/https://doi.org/10.1016/j.compbiomed.2023.106791, https://doi.org/10.48550/arXiv.2302.09462
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