TRANSFER LEARNING IN CHEST X-RAY PNEUMONIA DIAGNOSE

JIARUI BI [BIJIARUI@SEAS.UPENN.EDU], SHOUWEN GU [WILLIEGU@SEAS.UPENN.EDU],

ABSTRACT. This study examines the application of transfer learning to chest X-ray image classification for detecting pneumonia. Through fine-tuning and validation of state-of-the-art pre-trained models on a balanced dataset, we enhance prediction accuracy and model robustness. Our results indicate that transfer learning offers a promising avenue for medical image analysis, with the best-performing model achieving superior accuracy and an optimal F1 score. The research highlights the potential of transfer learning in diagnostic imaging, while also noting the challenge of model overfitting.

1. INTRODUCTION

Pneumonia, a severe acute respiratory disease, is prevalent globally and is particularly dangerous for individuals at the extremes of the age spectrum. This condition arises from the invasion of infectious agents, which may be viral, bacterial, or fungal, triggering an inflammatory response in the lungs. This inflammation leads to the filling of bronchioles and alveoli with fluid, causing significant respiratory difficulty.

In children under five, pneumonia is especially lethal, accounting for over 15% of global deaths in this age group. In 2015, there were approximately $920,000$ child fatalities worldwide due to pneumonia [9]. The disease also has a substantial impact in the United States, with over $500,000$ emergency hospital admissions and $50,000$ deaths in the same year, making it one of the top 10 causes of death in the country [1].

The challenges in diagnosing pneumonia are significant. Chest X-rays, the primary diagnostic tool, often yield unclear images that can be easily misclassified, leading to incorrect treatment. This problem is exacerbated by a lack of trained radiologists, particularly in low-resource countries.

1.1. Contributions.

Our study utilized different pre-trained Convolutional Neural Network or Transformer models, including Vgg16, ResNet34, EfficientNetV2, and ViT (Vision Transformer). Employing transfer learning and fine-tuning techniques, the resulting test accuracies were 84.29%, 85.10%, 80.45%, and 87.18% correspondingly. These findings underscore the potential of computer-aided diagnosis (CAD) systems in supporting clinicians, especially where there is a shortage of trained professionals, by providing immediate and accurate analysis of chest X-ray images for pneumonia detection.

2. BACKGROUND

2.1. Transfer Learning.

According to Weiss, T et al [8], transfer learning is a machine learning framework where a model's knowledge from a source domain D_S and learning task T_S is applied to improve the learning of a predictive function $f_T(\cdot)$ in a target domain D_T with a task T_T. This process is predicated on the notion that while the domains or tasks are related, they are not identical, i.e., $D_S \neq D_T$ or $T_S \neq T_T$.

A domain D consists of a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where $X = \{x_1, \ldots, x_n\} \in \mathcal{X}$. Correspondingly, a task T includes a label space $\mathcal{Y}$ and a predictive function $f(\cdot)$, learned from the training data consisting of feature and label pairs $\{x_i, y_i\}$.

In the context of transfer learning, the source domain D_S and source task T_S are characterized by a feature space $\mathcal{X}_S$, a marginal probability distribution $P(X_S)$, and a label space $\mathcal{Y}_S$ with the source predictive function $f_S(\cdot)$. The target domain D_T and task T_T, on the other hand, are similarly defined but with their respective feature spaces, probability distributions, and label spaces denoted as $\mathcal{X}_T$, $P(X_T)$, and $\mathcal{Y}_T$, with the target predictive function $f_T(\cdot)$.

The objective of transfer learning is to enhance the predictive function of the target domain $f_T(\cdot)$ by utilizing the knowledge contained in $f_S(\cdot)$, under the condition where $\mathcal{X}_S \neq \mathcal{X}_T$ or $P(X_S) \neq P(X_T)$, or both. This is known as heterogeneous transfer learning when $\mathcal{X}_S \neq \mathcal{X}_T$, and homogeneous transfer learning when $\mathcal{X}_S = \mathcal{X}_T$ but $P(X_S) \neq P(X_T)$.

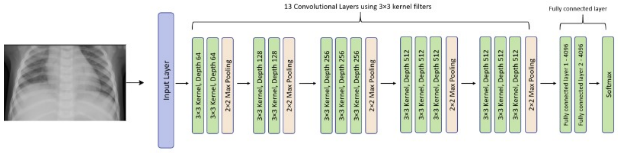


FIGURE 1. VGG-16 Architecture

2.2. **VGG-16.** The **VGG-16 model architecture**, introduced by *Karen Simonyan* and *Andrew Zisserman* in 2014 [5], is a seminal architecture in the field of deep learning for image recognition. Their work, titled *"Very Deep Convolutional Network for Large Scale Image Recognition,"* presented a deep network with 16 layers that contributed significantly to advancements in image processing. The architecture of VGG-16 is characterized by its simplicity and depth. It comprises:

- **13 Convolutional Layers:** These layers use a kernel size of 3×3, demonstrating the effectiveness of deep, small-kernel convolutional networks.
- **2 Fully Connected Layers:** Following the convolutional layers, these layers serve to synthesize the features extracted by the convolutions.
- **1 SoftMax Classifier:** The final layer in the network, responsible for the classification task.

The VGG-16 model is distinguished by its use of numerous but relatively small convolutional filters, specifically 3×3 filters, stacked on top of each other. This design choice emphasized the depth of the network as a key factor in improving performance for large-scale image recognition tasks.

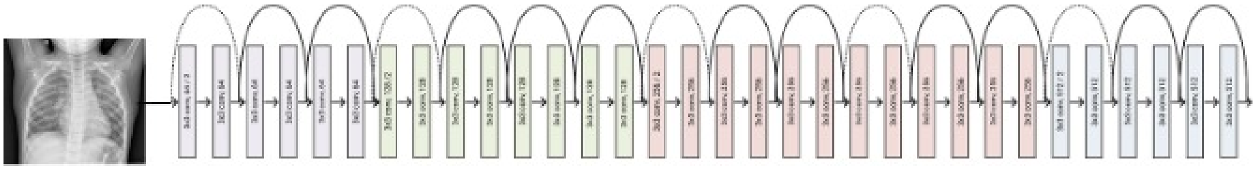


FIGURE 2. ResNet34 Architecture

2.3. **ResNet34.** The ResNet-34 architecture is a deep convolutional neural network with 34 layers, known for its use of residual blocks that facilitate the training of much deeper networks than was previously possible. It was introduced by Kaiming He et al [3]. in their seminal paper on deep residual learning. The core idea behind ResNet is to introduce direct connections between the layers in the form of *skip connections*. These connections allow the network to preserve the input information by bypassing non-linear transformations and combining outputs from multiple layers. A residual block in ResNet can be represented by the following formula:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x} \tag{1}$$

The ResNet-34 architecture consists of an initial convolutional layer followed by 16 residual blocks, with convolutional layers within each block. The network uses batch normalization after each convolution and employs a global average pooling layer after the last residual block, followed by a fully connected layer for classification.

2.4. **EfficientNet.** EfficientNet, which is introduced by Mingxing Tan and Quoc V. Le [7], is a scalable neural network architecture that achieves high accuracy with significantly reduced computational cost. It introduces a compound scaling method that uniformly scales network width, depth, and resolution with a set of fixed scaling coefficients. The compound scaling is defined by the equation:

$$\begin{aligned} d &= \alpha^{\phi}, \\ w &= \beta^{\phi}, \\ r &= \gamma^{\phi}, \end{aligned} \tag{2}$$

where: d is the network depth (number of layers), w is the network width (number of channels), r is the input image resolution, ϕ is the compound coefficient that scales the network, α, β, γ are constants that define how each of the dimensions is scaled. The scaling coefficients are subject to the constraint $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$ to balance the increase in computational cost. EfficientNetV2 is a state-of-the-art convolutional neural network architecture that represents an advancement in efficient and scalable deep learning models. It is an iteration on the original EfficientNet architecture, known for achieving excellent accuracy and efficiency in terms of computational resources.

2.5. **Vision Transformer.** The Vision Transformer (ViT), which is introduced by Alexey Dosovitskiy et al [2], marks a significant shift in the field of computer vision, moving away from conventional convolutional neural networks (CNNs) towards the application of transformers, a concept borrowed from natural language processing. Transformers process data in a manner that allows for attention-driven, contextually aware learning. In the context of image processing, this translates to a model's enhanced ability to focus on relevant parts of the input data. One notable variant of ViT is the '$vit\_base\_patch16\_224$' model. This model architecture involves dividing an image into fixed-size patches (in this case, 16x16 pixels) and linearly embedding each of them, akin to tokens in NLP. The model name, '$vit\_base\_patch16\_224$' , signifies a base-size transformer applied to 16x16 patches from images resized to 224x224 pixels. This approach allows the Vision Transformer to capture both local and global context effectively, leading to impressive performance in various image classification tasks. The '$vit\_base\_patch16\_224$' model, pre-trained on large datasets like ImageNet, has demonstrated its efficacy and robustness, making it a popular choice in advanced image recognition tasks.

3. Related Work

Srikanth Tammina [6] showcases transfer learning with a pre-trained VGG-16 model, enhanced for image classification through a Deep Convolutional Neural Network. By applying image augmentation techniques, they fine-tuned this model, achieving an accuracy of 79.20%. In a similar vein, Rahman T et al. [app10093233] focus on the automatic detection of bacterial and viral pneumonia using digital x-ray images. Their approach, grounded in transfer learning, employs four renowned Convolutional Neural Networks (CNNs) — AlexNet, ResNet18, DenseNet201, and SqueezeNet. This methodology was applied to process 5247 chest X-ray images, resulting in classification accuracies of 98%, 95%, and 93.3% for normal versus pneumonia, bacterial versus viral pneumonia, and a combined classification scheme, respectively. These accuracies exceed those previously reported in the literature. The implications of this study are significant, offering the potential to aid radiologists in rapidly diagnosing pneumonia and enhancing the efficiency of airport screenings.

4. Approach

Our approach centers on evaluating the performance of four pre-trained models—Vgg16, ResNet34, EfficientNetV2, and Vision Transformer—on chest X-ray image classification. We address class imbalance through data augmentation and fine-tune each model to our specific dataset. The models are trained with a keen eye on preventing overfitting by using a dedicated validation set, with performance measured by accuracy and F1 scores to ensure both precision and recall are considered in our assessment.

5. Experimental Results

5.1. **Dataset.** In this project, we utilize the ”Chest X-Ray Images (Pneumonia)” dataset, available on Kaggle [4]. This dataset includes chest X-ray images from pediatric patients aged one to five years at Guangzhou Women and Children's Medical Center. All imaging was part of routine clinical care. Quality control was stringent, with low-quality or unreadable scans being removed. Image diagnoses were validated by two expert physicians, and a third expert reviewed the evaluation set to ensure accuracy, thus enhancing the dataset's reliability. The dataset is structured into three folders: `train`, `test`, and `val`, each containing `Pneumonia` and `Normal` subcategories. It encompasses $5,863$ JPEG images, divided into Pneumonia and Normal classes. For model validation, we divided the dataset into training and validation sets. This division is crucial for assessing the model's performance and identifying any overfitting or underfitting, thus providing a more accurate performance estimate on new data. Due to the original validation set's limited size, we will form a new validation subset from the training set. To prevent data leakage, we grouped images by patient ID, as multiple pneumonia images exist per patient. 'Normal' images, lacking patient IDs, were randomly distributed between the training and validation sets.

5.2. **Data Augmentation.** To enhance our training dataset, we utilize data augmentation, applying rotations, flips, and translations to create varied examples. This approach helps in preventing overfitting and improves the model's generalization on new data. Additionally, the current dataset exhibits a class imbalance, with a higher number of pneumonia images compared to normal images. This imbalance could lead to a biased model with suboptimal performance on the under-represented class.

To mitigate this issue, we will employ techniques such as oversampling, undersampling, or a combination of both, aiming to equalize the number of samples in each class. These measures are expected to enhance the model's accuracy and reliability, particularly for the under-represented class.

Implementation Highlights:

(1) Distinct transformations are applied to the training data to avoid augmentation on the test set.
(2) Normalization is continued for consistency, despite minimal performance impact.
(3) The enriched dataset allows for more training epochs, thereby diminishing overfitting risks.
(4) Data augmentation is excluded from the validation set to ensure accurate model evaluation.
(5) Class imbalance is addressed through oversampling and/or undersampling techniques to balance the dataset.

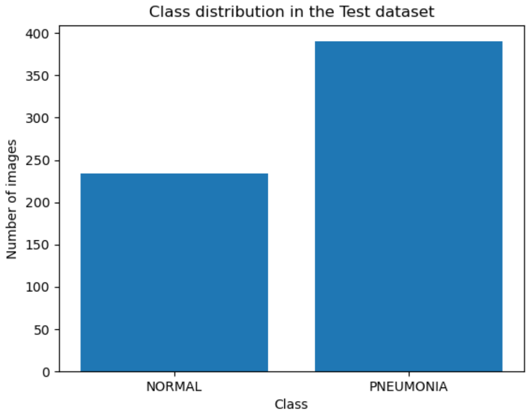


FIGURE 3. Class distribution in the Test dataset

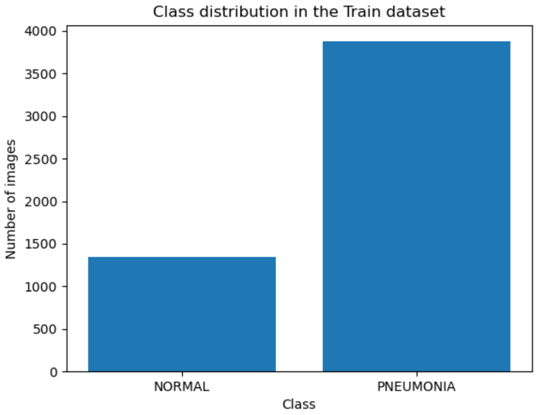


FIGURE 4. Class distribution in the Train dataset

5.3. **Model Evaluation.** We assess the efficacy of transfer learning with pre-trained models through the lens of the F1 score. This metric harmoniously integrates precision and recall, providing a comprehensive measure of a model's performance, especially in scenarios where both false positives and false negatives carry significant weight. The F1 score is particularly relevant for evaluating models in imbalanced datasets or in classification tasks where both aspects of the error are crucial. The formula for the F1 score is:

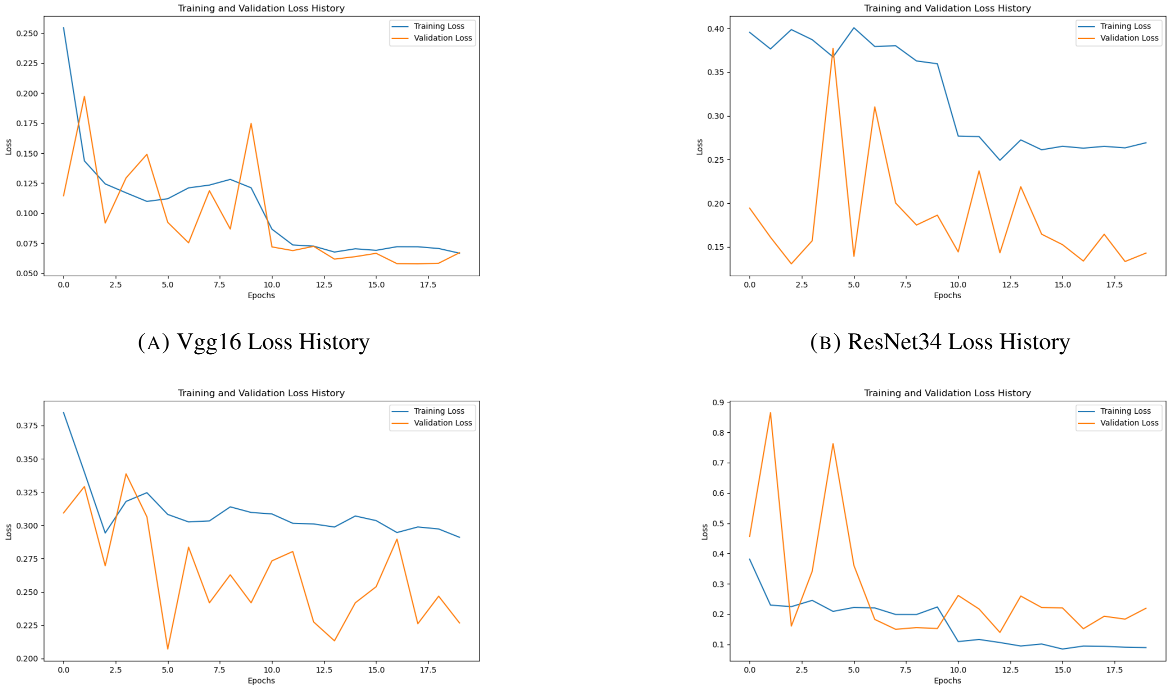
$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

In this context, precision and recall are derived from the model's classification results, and the F1 score thus serves as a robust indicator of the model's overall effectiveness in categorizing data accurately.

5.4. **Result.** Figure 5 presents the training and validation loss history for four distinct pre-trained models: Vgg16, ResNet34, EfficientNetV2, and Vision Transformer (ViT). Each subfigure illustrates the loss across epochs during the models' training phase.

(1) Subfigure (a) shows that the Vgg16 model's training loss decreases steadily, indicating consistent learning, while the validation loss shows fluctuations suggesting some variability in validation performance.
(2) Subfigure (b) demonstrates the ResNet34 model's training and validation loss trends with a decrease over epochs but with notable variability in validation loss.
(3) Subfigure (c) reveals that the EfficientNetV2 model has a relatively stable training loss with less variation in validation loss compared to the previous models.
(4) Subfigure (d) displays the ViT model's loss history, which exhibits sharp peaks in both training and validation loss, suggesting potential overfitting or instability during training.

The loss trends provide insights into each model's learning dynamics, with particular attention to the disparities between training and validation loss, which could signal overfitting or model instability.



(A) Vgg16 Loss History

(B) ResNet34 Loss History

(C) EfficientNetV2 Loss History

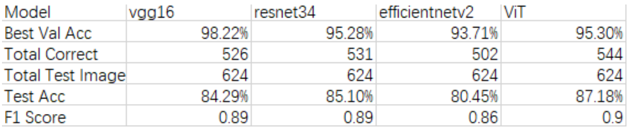
(D) ViT History

FIGURE 5. Loss history diagrams of the four different pre-trained models.

The comparative performance of four pre-trained models is presented in Table 6. The models evaluated include Vgg16, ResNet34, EfficientNetV2, and Vision Transformer (ViT), with their effectiveness gauged by best validation accuracy, total correct predictions, test set size, test accuracy, and F1 Score.

(1) Vgg16 leads in validation accuracy at 98.22%, albeit with lower test accuracy and F1 Score than ViT.
(2) ResNet34 displays consistent performance with a validation accuracy of 95.28% and matches Vgg16's F1 Score.
(3) EfficientNetV2 has the lowest test accuracy and F1 Score, suggesting it may be less effective for this dataset.
(4) ViT, despite a slightly lower best validation accuracy than Vgg16, excels in test accuracy at 87.18% and achieves the highest F1 Score of 0.9.

In summary, ViT emerges as the most balanced model, exhibiting superior test performance and the highest F1 Score, making it potentially the best model for generalizing to new data within this dataset context.



Model	vgg16	resnet34	efficientnetv2	ViT
Best Val Acc	98.22%	95.28%	93.71%	95.30%
Total Correct	526	531	502	544
Total Test Image	624	624	624	624
Test Acc	84.29%	85.10%	80.45%	87.18%
F1 Score	0.89	0.89	0.86	0.9

FIGURE 6. Summary Table

6. DISCUSSION

The outcomes of our investigation reveal that the Vision Transformer model outperforms traditional convolutional neural networks in classifying chest X-ray images, suggesting its superior capability in capturing complex patterns. This finding corroborates the growing evidence that attention-based models can achieve remarkable results in medical image analysis. However, the slight overfitting observed indicates a need for further optimization of regularization techniques. Future work could explore the integration of domain-specific augmentations and the potential of ensemble methods to enhance predictive performance and reliability in clinical settings.

REFERENCES

[1] Enes Ayan and Halil Murat Ünver. Diagnosis of pneumonia from chest x-ray images using deep learning. In *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, pages 1–5, 2019.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[4] Daniel Kermany, Kang Zhang, and Michael Goldbaum. Labeled optical coherence tomography (oct) and chest x-ray images for classification. Mendeley Data, v2, 2018.

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[6] Srikanth Tammina. Transfer learning using vgg-16 with deep convolutional neural network for classifying images. *International Journal of Scientific and Research Publications (IJSRP)*, 9(10):143–150, 2019.

[7] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

[8] Karl Weiss, Taghi M Khoshgoftaar, and Ding Wang. A survey of transfer learning. *Journal of Big Data*, 3(9), 2016.

[9] WHO. Standardization of interpretation of chest radiographs for the diagnosis of pneumonia in children, 2001.