Comparative Analysis of Convolutional Neural Networks and Vision Transformers in Medical Imaging

Mahmoud Alhihi
University of Minnesota - Twin Cities
Department of Science & Engineering
alhih001@umn.edu

November 2025

1 Introduction

Medical Imaging represents one of the most critical components of modern medicine that facilitate detection, diagnosis, and treatment. The era of imaging began in 1895 with the discovery of X-ray, the first way to view the internal organs without invasive surgery, and has evolved over the years with CT, MRI, Ultrasound, PET scans and more that have further advanced precise, non-invasive diagnostics.

Digital methods have taken over in recent years as a result of rapid expansion of digital health data and improvements in computational capacity to process such volume over time. Deep learning techniques have emerged over the last few years for medical image analysis that powers automated systems to learn complex parameters from digitized imaging solutions that boast sufficient volume to rival (or surpass) expert derived conclusions. For example, deep learning has been trained in applications that process image classification (tumor versus non-tumor), segmentation (define edges of an organ or tumor) and anomaly detection (diagnosing pneumonia in a chest X-ray).

As the medical datasets grow, therefore, it's essential to understand the comparative performance of varied deep learning approaches within the sub-domain. In computer vision alone, two predominant architectures emerge in the Convolutional Neural Network (CNN) and Vision Transformer(ViT) variety. When it comes to medical imaging, CNNs have been well-regarded as the gold standard

due to their inductive biases and relative ease to learn positional features from spatially aligned data. However, in recent discoveries, ViTs have emerged as promising challengers; especially on larger datasets, due to their self-attention architecture that empowers relative understanding of image context across vectors.

In this paper, I will compare Convolutional Neural Networks(CNNs) and Vision Transformers (ViTs) in the imaging domain through literature review and comparative experiments to assess variances of findings through different data set levels to expose the pros and cons of each model relative to real-world application feasibility.

2 Background

2.1 Convolutional Neural Networks (CNNs)

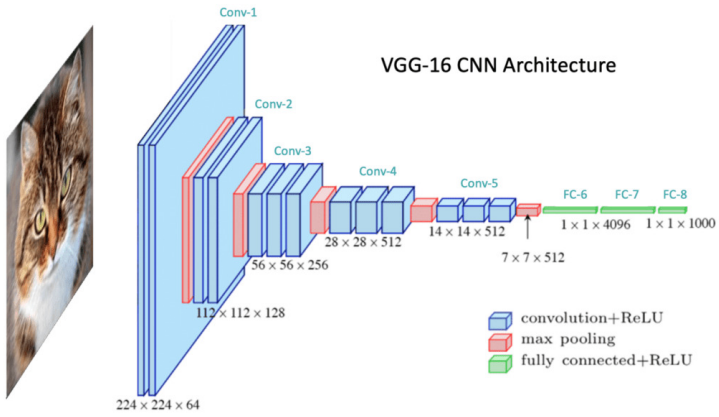


Figure 1: Example architecture of a CNN

Convolutional Neural Networks (CNNs) are one of the most widely used deep learning architectures for image analysis. As stated by Liu et al.[1], "CNN is a kind of feedforward neural network that is able to extract features from data with convolution structures." Simply, CNNs learn to automatically identify patterns, such as edges, shapes, and textures by applying filters over the input image. They are versatile: Beyond medical images, they can process text, speech signals, and time-series data such as sensor measurements.

Their architecture is partially inspired by biological visual processing. As de-

scribed in recent work, "A biological neuron corresponds to an artificial neuron; CNN kernels represent different receptors that can respond to various features; activation functions simulate the function that only neural signals exceeding a certain threshold can be transmitted to the next neuron." This biologically inspired pattern allows CNNs to learn from low-level features like edges to high-level structures such as organs or tumors

CNNs rely on a hierarchical architecture composed of convolutional layers, pooling layers, and fully connected layers. Convolutional layers extract spatial features using learnable kernels, pooling layers reduce spatial resolution to create more robust representations, and fully connected layers map these learned features to final predictions. This structured pipeline enables CNNs to efficiently capture local spatial patterns that are especially important in medical imaging, where subtle textural or structural differences can indicate disease.

Due to these strong points, CNNs have long been considered the gold standard architecture for analyzing medical images. However, recent works on transformer-based models have provided new alternatives that might potentially compete with CNN dominance, particularity on large-scale datasets

2.2 Vision Transformers (ViTs)

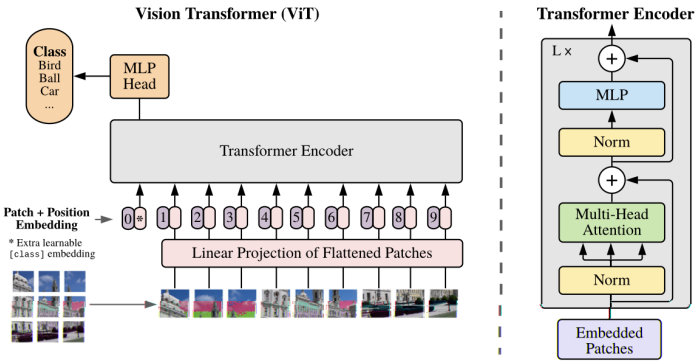


Figure 2: Example architecture of a Visual Transformer

Vision Transformers (ViTs) are a deep learning architecture that adapts the Transformer frameworks, it was originally developed for Natural Language Processing(NLP), and now it is used in images, "we split an image into patches and provide the sequence of linear embedding of these patches as an input to

a Transformer. Image patches are treated the same way as tokens(words) in an NLP application. We train the model on image classification in supervised fashion." (Dosovistikity et al., 2021)In simpler terms, the image is cut into small squares(patches), each patch is turned into a vector, and then these patches are processes by a Transformer Encoder just like a sequence of words.

ViTs use self-attention mechanisms to capture long-range spatial relationships across an image, enabling the model to understand global context rather than relying solely on local patterns. This also makes the ViTs particularly effective for large datasets where complex spatial dependencies play an important role. Thus, Vision Transformers have made their way to becoming strong competitors of CNNs in many computer vision tasks, including medical imaging.

2.3 Prior Comparative Studies

Recent comparative studies investigating CNN and Vision Transformer performance in medical imaging reveal that architectural effectiveness is highly task-dependent. Kawadkar [**?**] conducted a comprehensive evaluation across three modalities, chest X-ray pneumonia detection, brain tumor MRI classification, and dermoscopic skin cancer detection and reported that "ResNet-50 achieved 98.37% accuracy on chest X-ray classification", while "DeiT-Small excelled at brain tumor detection with 92.16% accuracy", and "EfficientNet-B0 led skin cancer classification at 81.84% accuracy"

Across studies, CNNs consistently show stronger performance on large, high-resolution datasets. For instance, the chest X-ray dataset; containing 4,172 training images, favored convolutional architectures, with Kwadakar noting that "CNNs demonstrated superior average performance (98.18%) compared to Vision Transformers (95.55%) on this modality. In contrast, Vision Transformers tend to outperform CNNs on smaller, specialized datasets. In the brain MRI classification task, ViTs achieved "89.22% average accuracy compared to CNNs at 72.55%, reflecting the strength of the self-attenition mechanisms in capturing subtle pathologic features under limited data conditions

For fine-grained texture analysis tasks, such as skin cancer detection, CNNs remain more effective. EfficientNet-B0 achieved the highest performance at 81.84%, while ViT-Base produced the lowest accuracy (77.82%), according to Kawadkar, "the self-attention mechanism may be less effective for fine-grained texture analysis required in dermatoscopic images"

Additionally, ViTs show greater cross-task stability, exhibiting "4.01% higher

average accuracy with signigicantly lower variance (7.82% vs 13.83%)" compared to CNNs, although CNNs generally train faster and require fewer computational resources.

Taken together, previous comparative work emphasizes that CNNs remain powerful for Large-scale or texture-heavy tasks, whereas Vision Transformers excel in scenarios needing global context understanding or where data are limited. The findings indeed motivate further investigation into architecture selection under differing medical imaging conditions.

3 Methods

3.1 Datasets

Two publicly available medical imaging datasets were used in this study to evaluate the performance of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) for disease classification tasks

3.1.1 Chest X-ray Pneumonia Dataset

The Chest X-ray Pneumonia dataset [**?**] contains 5,863 radiographic images categorized into two classes: Normal and Pneumonia. The dataset is already organized into predefined training, validation, and test splits, with a total of 4,237 Pneumonia images and 1,583 Normal images. This dataset represents a binary medical image classification task.

3.1.2 LC25000 Lung and Colon Cancer Histopathology Dataset

The LC25000 dataset [**?**] contains 25,000 histopathology patch images divided into five classes: Lung benign tissue, Lung adenocarcinoma, Lung squamous cell carcinoma, Colon benign tissue, and Colon adenocarcinoma.

For the purpose of this experiment, a balanced subset of 1,000 images was constructed by randomly sampling 200 images from each class. This subset was then partitioned into training, validation, and test splits using a 70% / 15% / 15% ratio, resulting in 700 training images, 150 validation images, and 150 test images.

3.2 Model Architectures

3.2.1 Convolutional Neural Network (CNN) - Custom Baseline

The first CNN architecture used in this study is a lightweight model implemented from scratch to serve a baseline. The model consists of three convolutional blocks followed by global average pooling and fully connected classification layer. Each convolutional block includes:

- A 2D convolution layer with 3x3 kernal and padding = 1
- Batch normalization to stabilize training
- ReLU non-linearity
- 2x2 max-pooling with stride 2 for spatial downsampling

These three blocks progressively increase the number of feature channels from 16 to 32 to 64, enabling the network to learn increasingly abstract representations of the input images. Following the last block, a global adaptive average pooling layer reduces each feature map to a single value such that the final 64-dimensional feature vector independent of the input spatial size. This is followed by a fully connected layer with output dimension equal to the number of classes, which produce the final class logits.

The architecture is kept very simple and compact, making it suitable as a baseline CNN to compare against deeper pre-trained models and the Vision Transformer.

3.2.2 Convolutional Neural Network (CNN) - ResNet18 with Transfer Learning

To evaluate a deeper and more powerful CNN architecture, ResNet18 was employed with ImageNet-1K pretrained weights. ResNet18 is a widely used residual network architecture that incorporates identity skip connections, enabling efficient training of deeper models without vanishing gradients.

For the purpose of this study, the original classification head of ResNet18 was replaced with a new fully connected layer. The input feature dimension of the original head was retained, and the final layer was replaced. This modification allows the model to output predictions corresponding to the number of classes in each dataset. The pretrained backbone provides strong initial feature representations, while the new classification layer enables fine-tuning on the medical imaging tasks.

3.2.3 Vision Transformer - ViT-B/16

A Vision Transformer pretrained on ImageNet-1K was used as the Transformer-based model for this study. ViT-B/16 processes the input image by splitting it into 16x16 patches, converting each patch into a vector, and feeding the sequence of patch embeddings into a transformer encoder.

Only the final classification head was modified to match the number of output classes. The original fully connceted layer was replaced. The pretrained backbone was kept intact, allowing the model the leverage strong initial representations while being fine-tuned on the medical imaging datasets.

3.3 Training Setup

All models were implemented in PyTorch and trained on the same training,validation, and test splits described in Section 3.1. Images were resized to 224 x 224 pixels and converted to three-channel tensors to match the input requirements of the CNNs and the Vision Transformer.

Training was formulated as a supervised classification problem using the cross-entropy loss. For optimization, the Adam optimizer was used with an L2 decay of 1 X 10^{-4}. The learning rates were set separately for each model:

- Simple CNN : learning rate = 1 X 10^{-4}
- ResNet18 : backbone frozen, only the final fully connected layer trainable, learning rate = 1 X 10^{-3}
- ViT-B/16: backbone frozen, only the classification head trainable, learning rate = 1 X 10^{-3}

Each model was trained for 10 epochs using mini-batch training. After each epoch, performance was evaluated on the validation set in terms of loss and accuracy. Once training was completed, the final models were evaluated on the test set using accuracy, precision, recall, F1-score, and confusion matrices to assess classification performance.

4 Results

4.1 Quantitative Evaluation

4.1.1 Lung/Colon Cancer Evaluation

Table 1–3 summarize the quantitative results of the Simple CNN, ResNet18, and ViT-B/16 models on the lung/colon cancer dataset. The reported metrics include precision, recall, F1-score, and support for each class.

Table 1: Simple CNN for Lung/Colon Cancer dataset

Class	**Precision**	**Recall**	**F1-score**	**Support**
COLON ADENOCARICINOMAS	0.89	0.53	0.67	30
COLON BENIGN	0.69	0.97	0.81	30
LUNG ADENOCARCINOMAS	0.69	0.67	0.68	30
LUNG BENIGN	1.00	0.83	0.91	30
LUNG SQUAMOUS	0.75	0.90	0.82	30
Accuracy			0.78	150
macro avg	0.80	0.78	0.78	150
weighted avg	0.80	0.78	0.78	150

Table 2: ResNet18 for Lung/Colon Cancer dataset

Class	**Precision**	**Recall**	**F1-score**	**Support**
COLON ADENOCARICINOMAS	0.91	0.97	0.94	30
COLON BENIGN	1.00	0.93	0.97	30
LUNG ADENOCARCINOMAS	0.75	0.90	0.82	30
LUNG BENIGN	1.00	0.87	0.93	30
LUNG SQUAMOUS	0.93	0.87	0.90	30
Accuracy			0.78	150
macro avg	0.80	0.78	0.78	150
weighted avg	0.80	0.78	0.78	150

Table 3: ViT-B/16 for Lung/Colon Cancer dataset

Class	**Precision**	**Recall**	**F1-score**	**Support**
COLON ADENOCARICINOMAS	1.00	0.97	0.98	30
COLON BENIGN	1.00	1.00	1.00	30
LUNG ADENOCARCINOMAS	0.90	0.90	0.90	30
LUNG BENIGN	1.00	1.00	1.00	30
LUNG SQUAMOUS	0.87	0.90	0.89	30
Accuracy			0.95	150
macro avg	0.95	0.95	0.95	150
weighted avg	0.95	0.95	0.95	150

4.1.2 Chest X-ray Evaluation

Table 4–6 show the quantitative results for the Simple CNN, ResNet18, and ViT-B/16 models on the Pneumonia Chest X-ray dataset.

Table 4: Simple CNN Performance on Chest X-ray Pneumonia Dataset

Class	**Precision**	**Recall**	**F1-score**	**Support**
NORMAL	0.79	0.53	0.64	234
PNEUMONIA	0.77	0.92	0.84	404
Accuracy			0.78	638
macro avg	0.78	0.73	0.74	638
weighted avg	0.78	0.78	0.76	638

Table 5: ResNet18 Performance on Chest X-ray Pneumonia Dataset

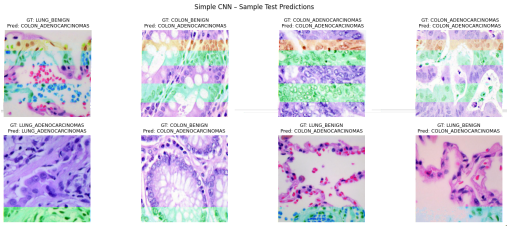
Class	**Precision**	**Recall**	**F1-score**	**Support**
NORMAL	0.97	0.60	0.74	234
PNEUMONIA	0.81	0.99	0.89	404
Accuracy			0.84	638
macro avg	0.89	0.79	0.81	638
weighted avg	0.87	0.84	0.83	638

Table 6: ViT-B/16 Perfomrance on Chest X-ray Pneumonia Dataset

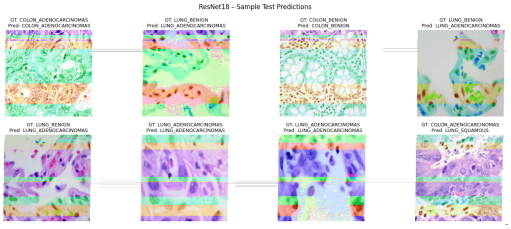
Class	Precision	Recall	F1-score	Support
NORMAL	0.95	0.77	0.85	234
PNEUMONIA	0.88	0.98	0.93	404
Accuracy			0.90	638
macro avg	0.92	0.88	0.89	638
weighted avg	0.91	0.90	0.90	638

4.2 Visualizations

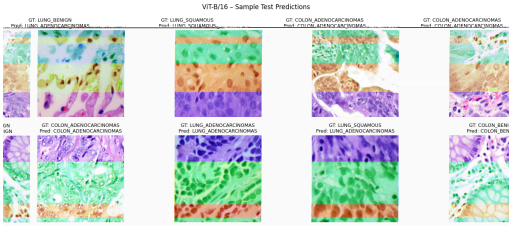
4.2.1 Lung/Colon Cancer Predictions



(a) Simple CNN



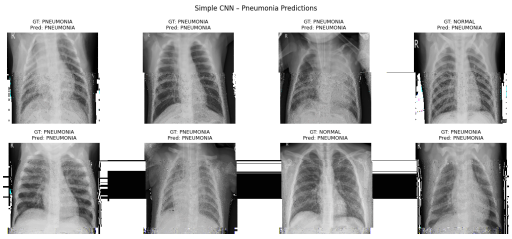
(b) ResNet18



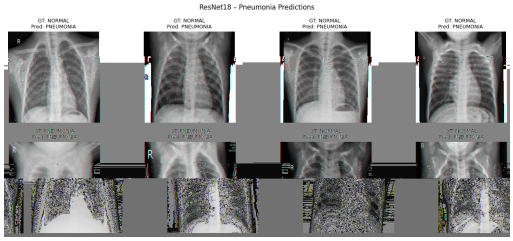
(c) ViT-B/16

Figure 3: Sample predictions for lung and colon cancer across the three models.

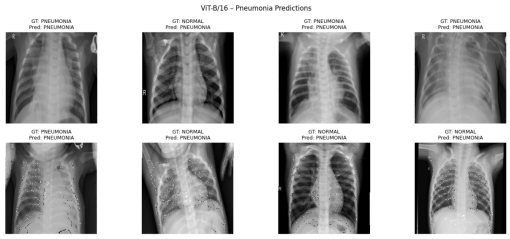
4.2.2 Chest X-ray Pneumonia Predictions



(a) Simple CNN



(b) ResNet18



(c) ViT-B/16

Figure 4: Sample chest X-ray pneumonia predictions across the three models.

4.3 Summary

An increasingly better performance with increasing complexity can be observed in quantitative results across both medical imaging datasets. For the Lung/Colon Cancer classification experiment, ViT-B/16 yielded the best results with macro precision, recall, and F1 of 95%. The simple CNN and ResNet18, however, both reached a macro F1 of only 78%, indicating the model dimensionality gap had a serious impact on results. Moreover, ViT-B/16's near perfect class-level precision and recall show that global self-attention had an increasingly good value for effectively understanding different image components structural and textural nuances with histopathology images that were more difficult for convolutions to assess

A similar performance pattern emerged in the Chest X-ray Pneumonia clas-

sification task. Although this task involves only two classes and was therefore easier across all models, ViT-B/16 again obtained the highest results, achieving a macro F1-score of 89% and demonstrating excellent sensitivity toward pneumonia cases (Recall = 98%, F1 = 93%). ResNet18 also performed strongly, achieving a macro F1-score of 81%, while the simple CNN baseline produced the weakest results with a macro F1-score of 74% and lower recall for normal cases. These findings highlight the importance of model depth and pretrained feature representations for reliable medical image classification.

Thus, the results across both datasets suggest ViTs generalize better; CNN based approaches struggle in both multiclass complex pathology assessments as well as simple radiographic classification approaches despite being less ideal. ResNet18 is still a competitive architecture, however, boasting strong relevance for chest X-ray analysis where convolutional inductive biases support better classifications. Simple CNN baseline performance metrics indicate that low dimensionality models do not produce refined results in medical imaging which is further compounded by these trends; ViTs seem to offer the best support for heterogeneous medical imaging modalities.

5 Discussion

Results of the present study indicate some interesting trends with respect to the comparative performance of CNNs and ViTs across two diverse medical imaging tasks. First, in both the Lung/Colon Cancer and Chest X-ray Pneumonia datasets, the Vision Transformer ViT-B/16 produces the best results. The capability of capturing long-range dependencies and global contextual information seems to be more beneficial for medical images, where subtle variations in texture or structure usually distinguish the classes of diagnosis. This was especially evident in the histopathology dataset, where ViT-B/16 achieved near-perfect macro precision, recall, and F1-scores of 95%, demonstrating a strong capacity to distinguish between visually similar cancer subtypes.

Second, deeper convolutional architectures such as ResNet18 proved to be robust across the datasets, outperforming the simple CNN baseline by a large margin. The respective macro F1-scores of 78% for the cancer dataset and 81% for the pneumonia dataset obtained by ResNet18 testify to the benefit of pretrained feature extractors and residual connections in capturing meaningful spatial features. This performance on the Chest X-ray dataset, in particular,

underlines that CNNs remain powerful tools for radiographic image analysis, where anatomical structures are globally consistent and well-suited to the inductive biases of convolutional kernels.

By contrast, the CNN baseline demonstrated limited generalization capability and performed significantly worse in terms of performance across all metrics. This was most severe in the case of the pneumonia classification task, as it was unable to identify normal scans due to the limited depth in its representation. This demonstrates the importance of model complexity and prior knowledge when implementing medical AI systems for various tasks that rely on identifying subtle abnormalities.

Overall, this study reveals that Vision Transformers provide the best performance for heterogeneous medical imaging applications for now, while deep CNN architectures represent competitive options for large-scale or organized imaging tasks.

6 Conclusion

This study compared three deep learning architectures; a baseline simple CNN, ResNet18, and the Vision Transformer ViT-B/16, across lung/colon cancer histopathology images and chest X-ray pneumonia detection. Results suggest that the Vision Transformer is the best performing for both datasets, achieving the highest precision, recall and f1 scores. Its construction as a means to model global relationships within the image allows for better discrimination of subtle patterns, which is important in specialized histopathological images where cancer subtypes may only differ by subtle textural differences.

In addition, ResNet18 performed well in comparison to the baseline CNN and although it wasn't as high performing as the Vision Transformer, it's likely that pretrained convolutional networks remain viable for medical image endeavors, especially where more structured anatomical components exist, such as chest x-rays. On the contrary, the CNN baseline performed decently enough but not enough to exceed ResNet18 - this suggests that a lighter architecture is not a good option in fields where so much diagnostic information could exist and feature layering is a possibility.

In summary, these findings suggest that the Vision Transformer is a good option for medical image classification across various imaging endeavors but deep CNNs are more practical and computationally easier. The next step may

be to investigate hybrid options, larger datasets or interpretability efforts to better facilitate trustable and clinically relevant models in the future.

7 References

[1] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, pp. 6999–7019, Dec. 2022.

[2] K. Kawadkar, *Comparative Analysis of Vision Transformers and Convolutional Neural Networks for Medical Image Classification*. 2025.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," June 2021. arXiv:2010.11929.

[4] L. B. T. C. P. W. L. A. D. S. M. M. Andrew A. Borkowski, Marilyn M. Bui, "Lc25000 lung and colon histopathological image dataset,"