Vithurshan Varenthirarajah

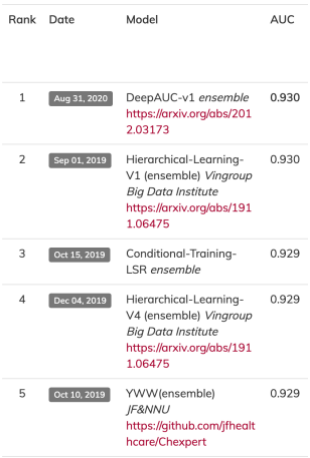
1. Abstract

Manual interpretation of Chest X-rays (CXRs) for detecting diseases like pneumonia, cardiomegaly, and edema is a critical process in determining lungs related disease. However, it is a time-consuming process and subject to inter-observer variability. Although, existing models based on Convolutional Neural Networks (CNNs) demonstrate strong performance in classification, they often struggle to capture global contextual understanding. Therefore, this project aims to explore the transformer-based architectures such as Vision Transformer (ViT), Swin Transformer, and Data efficient image transformers (DeiT) for CXR classification through the process known as fine-tuning pre-trained transformer-based models.

2. Introduction

Chest X-rays are crucial for the detection of various lung-related diseases, including pneumonia, cardiomegaly, and edeme. However, reading CXRs manually is time consuming, and the observation may vary from person to person (inter-observer variability). To automate this process with computational tools, people in the past developed a dataset containing X-ray images of chest with labels for 14 different diseases.

Current state-of-the-art models for CXR disease classification often leverages CNN based architectures and achieved good Area Under the Curve (AUC) of 0.93 as shown in the leaderboard in figure 1. However, these CNN based models lack the ability to effectively capture global contextual understanding within chest X-ray images due to their strong locality inductive bias and susceptibility to transition variance, necessitating model architectures with global contextual understanding.



Rank	Date	Model	AUC
1	Aug 31, 2020	DeepAUC-v1 *ensemble* https://arxiv.org/abs/2012.03173	0.930
2	Sep 01, 2019	Hierarchical-Learning-V1 (ensemble) *Vingroup Big Data Institute* https://arxiv.org/abs/1911.06475	0.930
3	Oct 15, 2019	Conditional-Training-LSR *ensemble*	0.929
4	Dec 04, 2019	Hierarchical-Learning-V4 (ensemble) *Vingroup Big Data Institute* https://arxiv.org/abs/1911.06475	0.929
5	Oct 10, 2019	YWW(ensemble) *JF&NNU* https://github.com/jfhealthcare/Chexpert	0.929

Figure 1 Top models achieving high AUC scores, including DeepAUC-v1 (0.930), Hierarchical-Learning (0.930/0.929), and others, ranked by performance and date.

Currently, we are in the era of Large Language Models (LLMs) such as ChatGPT, Gemini, and Grok. The main engine behind these models is the Transformer architecture and its core mechanism, self-attention, introduced in the paper *"Attention Is All You Need"* by the Google Research team in 2017. After the Transformer architecture achieved a major breakthrough in Natural Language Processing (NLP), researchers began exploring its capabilities in other domains as well, including biology, time-series analysis, audio, and computer vision. In biology, the Transformer architecture was used to predict the three-dimensional structure of proteins from their amino acid sequences through tools such as AlphaFold, which contributed to the Nobel Prize in Chemistry in 2024.

The core idea of this project is to investigate the efficacy of these transformer-based pre-trained models such as Vision Transformer (ViT), Swin Transformer, Data efficient image transformers (DeiT), and hybrid ResNet + ViT (figure 2) for our down-stream task of CXR disease classification.

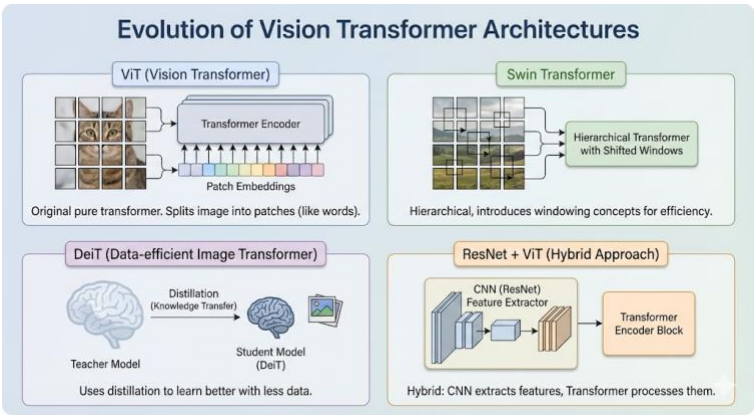


Figure 2 The evolution of Vision Transformer (ViT) architectures, detailing the original pure ViT, the data-efficient DeiT, the hierarchical Swin Transformer, and the hybrid ResNet + ViT approach

3. Methodology

3.1 Dataset

This project aims to use CheXpert dataset from Stanford ML Group (Stanford AIMI Shared Datasets) . This contains 224,316 chest radiographs taken from 65,240 patients with labels for 14 different thoracic diseases such as Pneumonia, Atelectasis, Cardiomegaly and Edema. This contains grey X-ray images of both frontal and lateral views with the resolution of 320 x 320 as shown in figure 3.

Figure 3 Chest X-ray showing a standard posteroanterior (PA) view (left) and a lateral view (right) of the lungs and thorax.

However, the original dataset contains uncertainties in the labels, denoted by **–1**. Moreover, some diseases are not clinically relevant. Therefore,

samples with uncertain labels and non-relevant diseases are excluded from this project. In this study, we focus only on the following clinically relevant conditions: Enlarged Cardiomediastinum, Cardiomegaly, Lung Opacity, Lung Lesion, Edema, Consolidation, Atelectasis, Pneumothorax, Pleural Effusion, and Support Devices**.** After all these operations, the training dataset contains 124,386 images and validation dataset contains 234 images.

As shown in figure 4, the CheXpert dataset shows severe class imbalance, with the majority of samples being negative (0) for most diseases. For an example, the Lung Lesion positive to negative ratio is only 0.04, indicating 25 times more negative examples. To mitigate this imbalance, a WeightedRandomSampler is used during training to assign a higher sampling weight to the minority (positive) class. Therefore, this ensures that the model is not biased towards the dominant negative class.

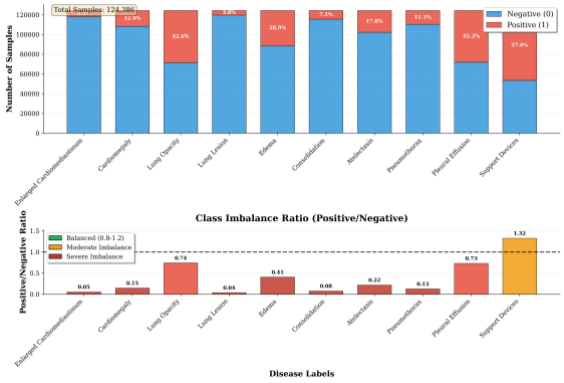


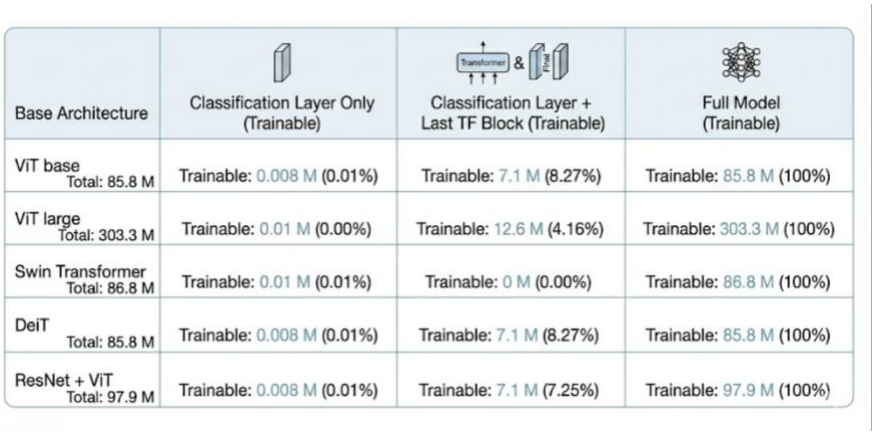
Figure 4 Bar charts illustrating the total sample distribution and class imbalance ratio (Positive/Negative) across various thoracic disease labels

Moreover, to improve the model's generalization and prevent overfitting, a data augmentation strategy was employed. As a chest X-ray is a medical image with

a fixed, anatomical orientation, aggressive transformations could alter the clinical meaning. Therefore, only minor, non-anatomically disruptive

augmentations were applied, specifically a small rotation (up to 5 degree) and a small zoom.

3.2 Model Architecture Overview



Base Architecture	Classification Layer Only (Trainable)	Classification Layer + Last TF Block (Trainable)	Full Model (Trainable)
ViT base Total: 85.8 M	Trainable: 0.008 M (0.01%)	Trainable: 7.1 M (8.27%)	Trainable: 85.8 M (100%)
ViT large Total: 303.3 M	Trainable: 0.01 M (0.00%)	Trainable: 12.6 M (4.16%)	Trainable: 303.3 M (100%)
Swin Transformer Total: 86.8 M	Trainable: 0.01 M (0.01%)	Trainable: 0 M (0.00%)	Trainable: 86.8 M (100%)
DeiT Total: 85.8 M	Trainable: 0.008 M (0.01%)	Trainable: 7.1 M (8.27%)	Trainable: 85.8 M (100%)
ResNet + ViT Total: 97.9 M	Trainable: 0.008 M (0.01%)	Trainable: 7.1 M (7.25%)	Trainable: 97.9 M (100%)

Figure 5 A set of images detailing the process of developing computer vision models for chest disease detection, including a standard chest X-ray view, bar charts showing data distribution and class imbalance across disease labels, the evolution of Vision Transformer (ViT) architectures, and an analysis of their parameter efficiency

The models compared in this study mainly differ in how big and deep they are, and in how much attention capacity they have. ViT-Base uses 768-dimensional embeddings, a 3072-dimensional hidden layer, 12 attention heads, and 12 Transformer blocks. Whereas ViT-Large scales this up to 1024-dimensional embeddings, a 4096-dimensional hidden layer, 16 attention heads, and 24 Transformer blocks, which lets it capture richer global information but also makes it more computationally expensive. When it comes to Swin Transformer, it takes a different approach with shifted window attention as it processes images in stages with gradually increasing embedding sizes (for example, 96 → 192 → 384 → 768) and focuses attention within local windows, which makes it more efficient for vision tasks. Additionally, DeiT is a data-efficient version of ViT-Base that keeps the same architecture as ViT-Base but adds a distillation token to boost performance on smaller datasets. Finally, the ResNet + ViT hybrid model combines convolutional layers from ResNet for local feature extraction with Transformer layers for global context modeling, leveraging

both the strong local biases of CNNs and the long-range reasoning ability of Transformers.

3.3 Fine-Tuning Striates

To study how efficiently the models use their parameters, we consider three fine-tuning strategies. In the first, all pre-trained Transformer layers are frozen and only the final classification layer is trained. In the second strategy, both the classifier and the last Transformer block are trained. In the third strategy, the entire model is fine-tuned, with all parameters updated.

3.4 Training Configuration and Experimental Setup

To train the models, we chose hyperparameters that keep learning stable and reasonably fast. We used a learning rate of $2 \times 10^{-4}$and batch sizes of 64, 128, or 256, depending on model size and GPU memory. All models were trained for 20 epochs with the AdamW optimizer and a weight decay of 0.01 to help prevent overfitting.

Because CheXpert is a multi-label task (a single X-ray can show several findings at once), we used BCEWithLogitsLoss and treated each disease as its own binary prediction. To deal with class imbalance, we used a WeightedRandomSampler so that rare conditions are seen more often during training and don't get overshadowed by common ones. All experiments ran on NVIDIA H100 GPUs; for example, training ViT-Large for 20 epochs took about 110 minutes.

#	Hyperparameter	Value
1	Learning Rate	2e-4
2	Batch Size	64, 128, 256
3	Epochs	20
4	Optimizer	AdamW (weight_decay=0.01)
5	Loss Function	BCEWithLogitsLoss

6	Sampler	WeightedRandomSampler
7	Hardware	NVIDIA H100

4. Results

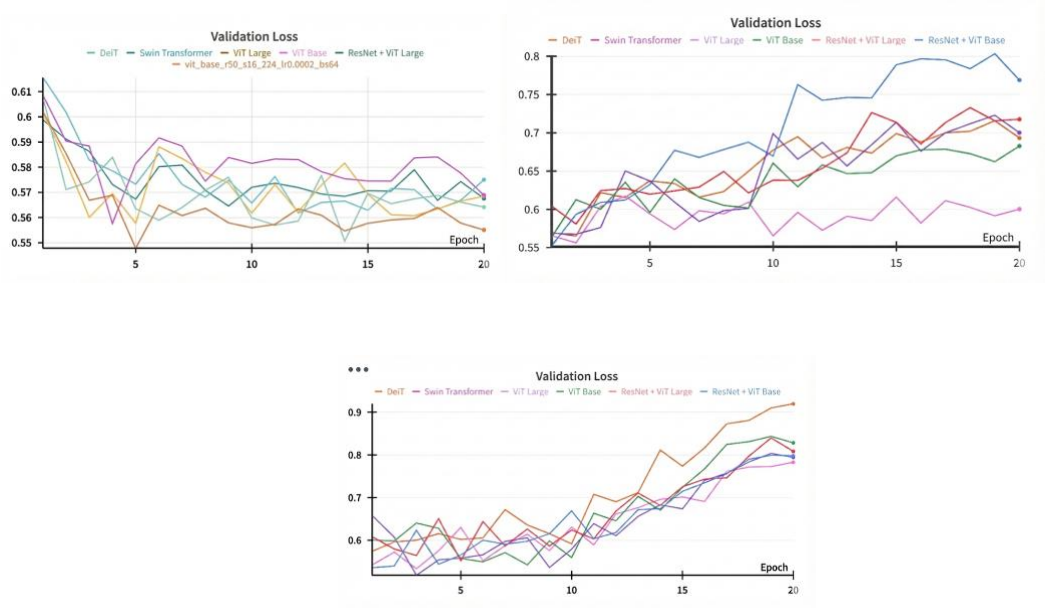


Figure 6 Validation Loss plot for various Vision Transformer (ViT) and CNN architectures on the CheXpert dataset (batch size 64) are shown across three fine-tuning strategies: (a) classification head only, (b) classification head and last transformer block, and (c) whole model trainable.

As shown in figure 6, in setup (a), where only the classification head is trained, the validation loss is the lowest and also relatively stable, staying roughly between 0.55 and 0.61, even though it still fluctuates. In setup (b), where we train the classification head and the last Transformer block, the loss is higher and, for most models, starts to climb noticeably after about 10 epochs. In setup (c), where the whole model is trainable, the validation loss rises sharply for almost all models, reaching the highest values (around 0.9). This strong upward trend is a clear sign of severe overfitting at this batch size.

From the Validation AUC curves as shown in figure 7, performance actually becomes less stable as we fine-tune more of the model, and there is no clear AUC gain over just training the classification head.

In scenario (a), where only the head is trained, all models quickly improve and then level off around 0.65–0.67 after about five epochs. The curves are noisy but don't drift downward, which suggests the pre-trained features are already very strong.

In scenario (b), where we also train the last Transformer block, the AUC curves become much more volatile, jumping around between about 0.60 and 0.70 from one epoch to the next.

In scenario (c), with the whole model trainable, the instability is strongest. The curves show sharp saw-tooth patterns; some models briefly reach higher AUCs (around 0.74), but the performance swings a lot and doesn't stay high. With a batch size of 64, this makes full fine-tuning unreliable and does not provide a consistent improvement over the simpler, more stable head-only setup.

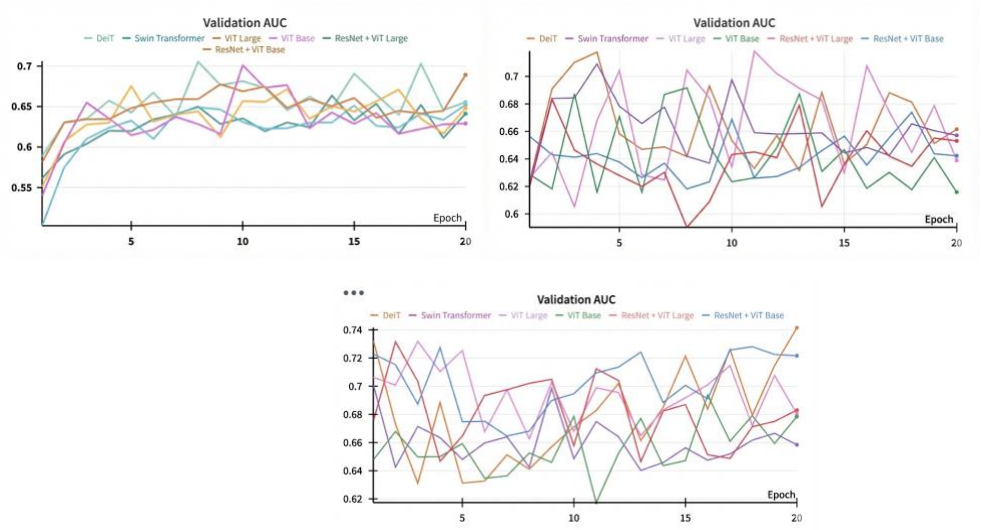


Figure 7 Validation AUC plot for various Vision Transformer (ViT) and CNN architectures on the CheXpert dataset (batch size 64) are shown across three fine-tuning strategies: (a) classification head only, (b) classification head and last transformer block, and (c) whole model trainable.

Similar to the above with a batch size of 64, other batch sizes such as 128 and 256 also tried as shown in figure 8 and figure 9.

When looking at the validation loss curves, both batch sizes, 128 and 256, show the same pattern: after about step 10, the loss starts climbing steadily and doesn't stop. Many of the experiments end with loss values above 0.8, which is a clear sign that the models are overfitting heavily, no matter which of the two batch sizes is used during full fine-tuning.

The validation AUC tells a similar story. For both batch sizes, the AUC jumps up and down dramatically, swinging between roughly 0.62 and 0.74. None of the models manage to keep a stable or consistently strong performance. Instead of improving smoothly over time, the curves remain noisy and highly variable, suggesting that using larger batch sizes does not fix the high variance or instability that appears during full fine-tuning.

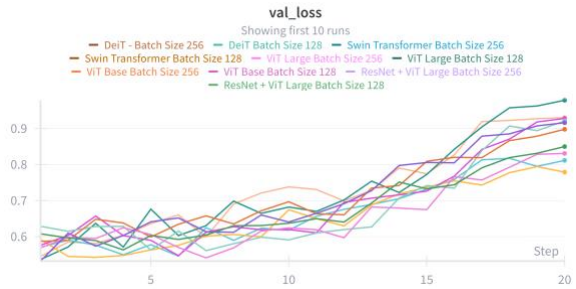


Figure 8 Validation Loss plot for various Vision Transformer (ViT) and CNN architectures on the CheXpert dataset (batch size 64) are shown for whole model trainable.

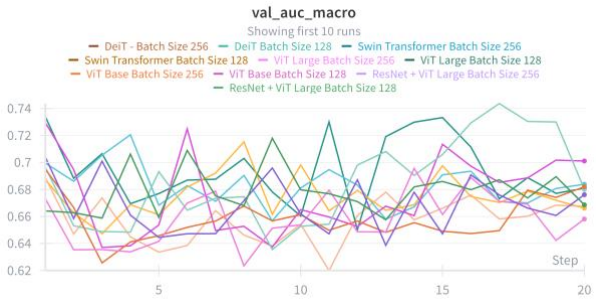


Figure 9 Validation AUC plot for various Vision Transformer (ViT) and CNN architectures on the CheXpert dataset (batch size 64) are shown for whole model trainable.

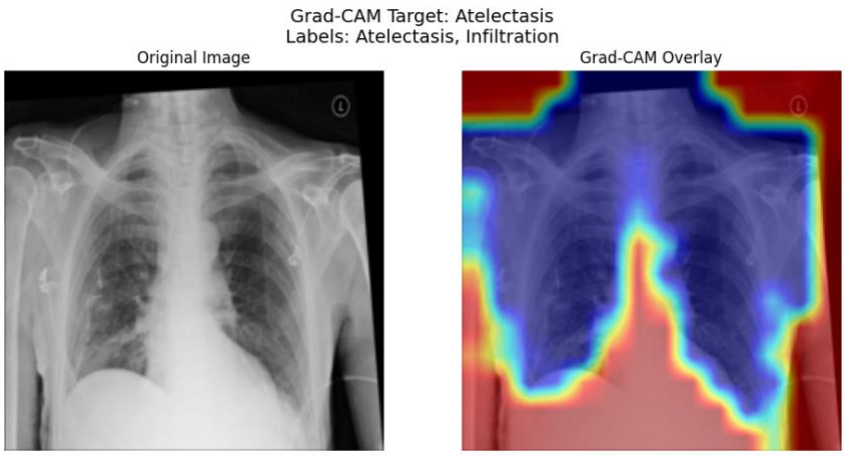


Figure 10 Grad-CAM Visualization for Interpretable Atelectasis Classification

Figure 10 shows a Grad-CAM explanation for an Atelectasis prediction on a chest X-ray. Next to the original image, there is a heatmap overlay that highlights the regions the model paid the most attention to. The bright red and yellow areas, mainly in the lower parts of the lungs, indicate where the model focused when deciding whether Atelectasis was present. This kind of visualization helps us check whether the model is looking at the right anatomical areas and making decisions based on clinically meaningful features, rather than irrelevant parts of the image.

5. Conclusion

Across all batch sizes, one pattern stands out very clearly: when the entire model is made trainable, the validation loss rises sharply and keeps climbing, often reaching values around 0.9 or higher. This steady increase shows that the models are severely overfitting, likely because full fine-tuning gives them more capacity than the dataset can realistically support.

The validation AUC curves also reveal another issue: instability. For all full fine-tuning runs—whether the batch size is 64, 128, or 256, the AUC fluctuates in a noisy, saw-tooth pattern, making the results inconsistent and hard to trust.

In contrast, the simplest approach—training only the classification head, produces the most stable and reliable behavior. With batch size 64, this shallow tuning

maintains validation AUC values around 0.65–0.67 and validation loss in the range of 0.55–0.61, making it the most dependable configuration among all experiments.

Overall, the transformer-based models are underperforming in this setup. Their best AUC values are still well below the CNN benchmark, and the high variability and instability suggest that, at least for this dataset and current setup, these transformer models are not yet reliable alternatives to the established CNN approach.

6. Reference

Dosovitskiy et al., 2021 – *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, ICLR, Vienna.

He et al., 2016 – *Deep Residual Learning for Image Recognition*, CVPR, Las Vegas.

Irvin et al., 2019 – *CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison*, AAAI, 33(01): 590–597.

Liu et al., 2021 – *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*, ICCV, 10012–10022.

Touvron et al., 2021 – *Training data-efficient image transformers & distillation through attention*, ICML, 10347–10357, PMLR.