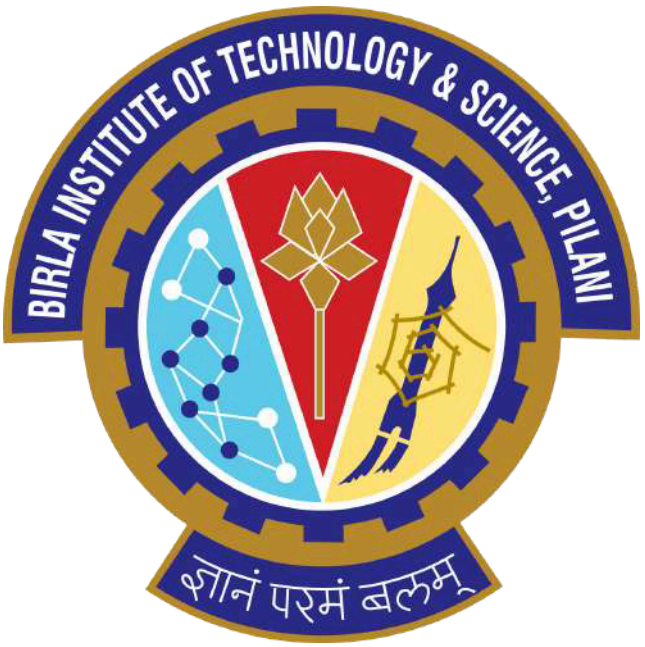
Birla Institute of Technology and Science, Pilani
Dubai Campus



"Image Classification for Lung Disease Diagnosis using Transformers"

Project by

Nikhil Vaidyanath 2022A7PS0073U

List of Figures

- F1: Comparison of the validation AUC-ROC with epoch number
- F2: Comparison of class wise AUC-ROC scores
- F3: Class wise AUC-ROC change with epoch
- F4: Comparison of losses and AUC scores with epochs
- F5: Sample GradCam output at a lower resolution

List of Tables

- T1: Comparison of class wise AUC-ROC
- T2: Values of losses and AUC for the hybrid ViT model with early stopping
- T3: Comparison of our models and other SOTA

Table of Contents

- Abstract
- Introduction
- Background
    - NIH Chest X-ray Dataset
    - Vision Transformer (ViT)
    - Swin Transformer
- Data Processing
    - Image Handling
    - Normalization
    - Class Representation
    - Class Imbalance
    - Data Augmentation
    - Data Loading
- SwinChex
- Hybrid Architecture Motivation
- Model Architectures and their Technical Implementations
    - Hybrid CNN-Swin Architecture
    - Hybrid CNN-ViT Architecture
- Training Setup
- Results and Evaluation
- Benchmark Comparisons
- Model Explainability with Grad-CAM
- Discussion and Conclusion
- Literature Review
- References

Abstract

This paper explores how transformer-based architectures perform in detecting chest diseases from chest X-ray images using the NIH ChestX-ray14 dataset which contains about 112,000 annotated images in 14 disease categories. The focus of the study is two methods, CNN-Swin and CNN-ViT, where CNNs are paired with transformers, to bring the feature sensitivity of CNNs and the global view of transformers together.

A pre-trained ResNet50 CNN, part of ratio-based CNNs, is chosen to detect fine surface texture from grayscale medical images which are transformed into RGB and sized to fit the transformer input. Subsequently, the feature maps enter either Swin Transformer blocks, with shifted window-based self-attention for efficient representation or ViT encoders which build a hierarchical structure by relying on global self-attention. Both architectures can distinguish between various labels and include sigmoid-active outputs. They are also trained with weighted Binary Cross-Entropy loss to make up for the imbalance present in the data.

To ensure the data was clear and the model did not show bias, normalization, label encoding and stratified sampling were all performed. Light improvements such as flipping images and cropping them in various places were used to make the model generalize better, but without losing its validity in the clinic. The algorithm improved computationally by using mixed precision and adjusting the learning rate, along with warm restarting.

Our results reveal that the CNN-ViT hybrid achieved an average AUC-ROC score of 0.82 which is greater than the baseline SwinCheX model and close to what LungMaxViT can achieve, but uses less computing resources. With Grad-CAM, we could check that attention maps were relevant to clinical points, showing that model explainability improved.

The research reveals that hybrid CNN-transformer models are suitable, interpretable and energy-friendly for use in diagnosing lung diseases from images and stimulates more studies of attention-based approaches in medical image analysis.

Introduction

Medical imaging practices support modern diagnostic procedures by providing tools for the identification and treatment monitoring of pulmonary conditions including pneumonia alongside emphysema and fibrosis. X-ray imaging for the chest represents one of the most popular diagnostic procedures because it offers fast results at affordable prices with effective diagnostic capabilities. The hands-on review of radiographs requires extended work hours from radiologists who must also handle numerous interpretation errors while their skills remain the essential factor.

Deep learning as a subset of artificial intelligence technology produces outstanding achievements when automating X-ray interpretation. CNNs have been known to dominate this field while their spatial-hierarchical architecture provides unique solutions to the problem domain. Transformers created for Natural Language Processing have become significant in vision domain modelling since they excel at establishing relationships across image data.

The objective behind this study centres on understanding the degree to which transformer architecture including Swin Transformer and ViT suits medical image classification needs. Methodical tests analyse the strengths and weaknesses of these architectures when processing the NIH Chest X-ray dataset.

Background

- **Dataset**

  The NIH Chest X-ray dataset represents a substantial public database that contains upwards of 112,000 radiological examinations from different parts of the chest belonging to more than 30,000 unique individuals. The images in this dataset carry between one and fourteen disease categories established through radiological reports. These include: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, Hernia.

The images present in grayscale possess a pixel resolution of 1024x1024. The data includes major class imbalances because Hernia and Fibrosis diagnoses occur rarely compared to Infiltration and Effusion.

- **Vision Transformer**

  ViT treats images as sequences of fixed-size patches. Each patch is flattened and linearly projected into an embedding space. These embeddings are then passed through standard transformer blocks, as introduced in NLP. The position of each patch is encoded via positional embeddings, enabling the model to retain spatial information. ViT's strength lies in its direct attention-based representation without convolutional priors. However, it demands a large volume of data and extensive pretraining to perform optimally due to its lack of spatial bias.

- **Swin Transformer**

Swin Transformer introduces a hierarchical feature extraction mechanism by limiting self-attention computations to non-overlapping local windows. These windows are shifted cyclically at each stage to capture global context. This makes Swin computationally efficient and scalable for high-resolution images, unlike ViT, which scales poorly with image size.

Data Processing

- **Image Handling**

The NIH Chest X-ray dataset contains more than 112,000 diverse medical images that origin from clinical environments using different imaging devices under different operational conditions. All visual inputs required resizing to uniform 224 by 224 pixels dimensions to meet transformer architecture requirements for fixed input sizes. The used resolution derives from the common input expectation of Vision Transformer (ViT) and Swin Transformer models that underwent ImageNet pretraining. The standardization procedure enables proper operation of positional embedding and patch-tokenization systems learned by these models. The necessary compromise between computational requirements and architectural requirements involves reducing image size to smaller resolutions which causes minor diagnostic detail loss.

All images in the dataset correspond to grayscale single-channel Chest X-ray images. These deep learning models together with the Swin Transformer accept only RGB images with three channels since they come from ImageNet pretraining. The grayscale images required a conversion to three-channel format by repeating the single channel in each of the RGB dimensions to match model requirements. This modification maintains the original information and provides compatibility with transformer weight matrices which facilitates successful knowledge transfer.

- **Normalization**

Each image received normalization procedures after channel adaptation and resizing completion. The research required normalization of pixel intensities in images through the same mean and standard deviation statistics that existed during pretraining of ImageNet datasets with transformers. When performing normalization it becomes essential to minimize the covariate shift that exists between ImageNet training data distribution and the NIH Chest X-ray fine-tuning dataset.

- **Class Representation**

The NIH dataset applies multi-label classification which deals with X-ray images with possible zero or multiple or one thoracic diseases occurring at the same time. The raw labels exist as strings containing all diseases which appear in each image. Each disease category got its own position within a 14-element binary hot vector form representing the string labels. Each presence of a disease created a value of one in its respective index value. The new format enabled the model to process disease predictions by treating each case as an independent classification among many options through standard multi-label protocols.

Correct processing of labels required special attention to the 'No Finding' class which indicates scans without detectable diagnoses. Such improper handling of this class could damage the multi-label configuration because it contradicts medical terminology when occurring with multiple labels. The model encoded 'No Finding' images with a vector containing only zeros because all such instances were removed or corrected during label cleaning. The extraction process that NIH dataset curators used to derive labels automatically through NLP brought about unknown values often tagged as "-1." Simplification and stability required the conversion of uncertain labels into negative labels which equates to setting them to zero. Additional methods like label smoothing or probabilistic modelling could boost future development to deal with label uncertainty.

- **Class Imbalance**

The dataset contains substantial unequal distribution between different medical conditions since Infiltration, Atelectasis and Effusion appear much more frequently than Fibrosis or Hernia. The unbalanced training environment increases major risks because neural networks tend to gravitate toward common classes while overlooking essential but infrequent findings. A strategic image subset consisting of 6000 items was used for Vision Transformer model analysis. The selection method enforced a distribution of 714 images from among the 14 classes during the process. The strategy used stratified sampling with overlap enabled because the same images could be used multiple times across class groupings for multi-label datasets.

The model performed better on less common conditions because the balanced sampling allowed direct training across all classes. Adequate training exposure of the model to each class improved the learning process to fully represent the complete diagnostic range therefore enhancing clinical system functionality.

- **Data Augmentation**

Deep learning models require data augmentation techniques for better generalization because small or uneven datasets present a challenge for their development. Medical imaging practice particularly for chest radiographs requires careful consideration about

using augmentation methods. The clinical presentation shows faint markers which may become lost when performing severe augmentation procedures.

A restrained augmentation technique was applied in this experimental design. Random resized cropping combined with horizontal flipping served the project because these methods created spatial variation points that preserved diagnostic relevant regions. The implementation of rotations along with elastic deformations and contrast inversion transformations was avoided to preserve clinical information accuracy.

The augmentations were applied exclusively during training to avoid corrupting performance assessments since they did not use them during validation or testing stages. During training this approach generated imaging variations with realistic clinical fluctuations without permitting the model to become specialized to the noise patterns or position effects.

- **Data Loading**

The large NIH dataset required along with scarce GPU resources demanded effective data loading and batching procedures. Data loading through multi-threaded systems combined with memory pinning techniques maintained a continuous stream of data for the GPU thus reducing training idle time. The system automatically modified batch sizes depending on what amount of memory the GPU could access. Because Swin Transformer has intense memory consumption during hierarchical attention computations it required small batch sizes leading to slower performance and longer training duration. The ViT model processed data from a smaller balanced dataset leading to efficient stable training through large batch sizes.

SwinCheX

Owing to the large amounts of data available, deep neural networks have been widely adopted recently to make automatic diagnoses using chest X-ray photographs. Normally, computer vision tasks used CNNs for accurate classification, but recent research is demonstrating that transformers which led the way in natural language processing, can exceed CNNs in computer vision. One study introduces a new classification method for several chest diseases using the Swin Transformer and a Multi-Layer Perceptron, reaching top results in the field. Using data from the ChestX-ray14 dataset which consists of close to 100,000 X-ray images from around 30,000 patients with 14 typical chest diseases, their model achieves excellent outcomes. With in-depth testing, the proposed 3-layer MLP head configuration achieved an average AUC score of 0.810 which outperformed the previous best average AUC score of 0.799. It also presents a standard experimental design that ensures comparability, giving a solid base for further works. Also, the attention in the model was

checked to aim at important parts of the lungs, lifting its accuracy in diagnosing lung pathologies. The model developed by this study is popularly known as the SwinCheX.

Motivation for Hybrid Architecture

Researchers adopt hybrid styles because there are clear differences in performance between the two main network types. CNNs often focus on local features, while transformers are capable of replicating and using global attention. Many in the field have chosen CNNs for medical image analysis because they use local information well and are not overly complicated. Radiologists can easily spot tiny changes in lung texture, patterns of calcifications or unusual borders using chest X-rays. Owing to the specialization in local information, CNNs find it hard to recognize patterns covering distant areas of the chest.

On the other hand, Vision Transformers (ViTs) model important connections across the whole image with self-attention mechanisms. Such mechanisms make it possible for the model to link data from distant areas, helping to diagnose cardiomegaly and pleural effusions by examining relations between the heart and the peripheral lung regions. Building image classifiers with transformers is not easy for this reason, as the models need lots of training data and calculation power to work well with high-resolution images.

A hybrid model was developed to strike a balance between the neighborhood pattern detection of CNNs and the wide perspective modeling done well by transformers. The architecture produced in this paper depends on CNNs to identify key features in a picture and those findings are then refined by the Swin Transformer layers or Vision Transformer layers to gather semantic meaning over the whole area of the image.

As a result, the weak points of a component are balanced by the strengths of the other, leading to a better and wider-use model for multi-label disease classification.

Hybrid CNN-Swin Architecture

SwinCheX implements a CNN + Swin architecture where two routes are used together: convolutional feature extraction in one path and attention modeling by transformers in the following path.

In order to find important spatial features, the first stage uses a pre trained Resnet50 on ImageNet to analyze the chest X-ray. This backbone processes the image that was formed by duplicating the original grayscale channel to follow the format needed for ImageNet. Extracted features capture lower and intermediate lung patterns such as lung texture, changes in the costophrenic angle and variations in how the mediastinum looks.

Then, the CNN’s feature maps are passed as input to the program’s Swin Transformer blocks. To perform these blocks, the feature maps are cut up into separate windows, where the self-attention is carried out. A major advancement is found in the shifting of the windows within the model — each time, the boundaries of the windows are shifted to reveal ways that

different patches in the image are connected. The model mimics global attention at a much lower computational cost than a typical transformer model.

The CNN and Swin Transformer outputs are fused using an attention-based method. In this case, multi-head attention modules are used to match and merge the embeddings produced by both pathways. It is not enough to simply join the sources; relevant scores are calculated between elements from each source and the final representation emphasizes the most important aspects from local and global regions.

This final step of the architecture is a multi-label classification head. This involves an MLP that also includes a sigmoid output layer which outputs a probability value for every disease class. The use of sigmoid, as opposed to softmax, is appropriate for multi-label tasks where multiple diseases may co-occur in a single image.

This full architecture was trained end-to-end on the complete NIH ChestX-ray14 dataset using class-balanced sampling strategies and adaptive loss weighting to address label imbalance and to ensure equitable learning across all classes.

Technical Implementation

A great deal of attention was needed to arrange data, launch the model and set up efficient computation for the entire dataset when running the hybrid CNN + Swin Transformer model.

AdamW optimizer, with separate weight decay, was used to train the models in order to reduce overfitting and improve their ability to generalize. A cosine annealing schedule with warm restarts was used for the learning rate, so the model was able to converge smoothly and regularly get out of local minima. This architecture adopted Binary Cross-Entropy with Logits (BCEWithLogitsLoss), as well as class-specific weighting that is inverse to the number of times labels were seen, so issues that only appeared rarely, like hernia or fibrosis, did not get ignored during training.

To cut down on data storage and speed up training, mixed precision was used. Because of that, the model could handle 16 batch samples without overloading the GPU's memory. Based on the validation AUC-ROC performance, we applied early stopping to avoid overfitting and ensure our model remains usable in new situations.

Advanced approaches for enhancing data were applied to the medical imaging pipeline. Clinically significant features were preserved by being augmented subtly, but variability in the images was introduced through random resizing and horizontal flipping. No augmentation was used for inference to keep evaluations during validation and testing identical.

The model was loaded with pretrained weights from ImageNet into the CNN feature extraction layers and did the same for the Swin Transformer components. Following that, the system was tuned on the entire NIH ChestX-ray14 data to make sure it responded to the specific types of thoracic diseases included in the data.

Hybrid CNN-ViT Architecture

Another hybrid architecture created in the study joined the CNN backbone with a ViT (Vision Transformer) stack on top. Structurally and in terms of information transfer, this model differed from the Swin variant.

The CNN + ViT model is designed as a sequential architecture. Generally, the CNN is a ResNet50 that does the main feature extraction task. The CNN takes raw images and creates feature maps, from which it partitions fixed-sized patches instead of processing the raw patches. A set of embeddings is produced by unfolding and linearly transforming each patch. These patch embeddings have learned positional encodings included to maintain the structure of the image.

After that, the set of embedded patches is put into the ViT encoder which is built with a series of transformer blocks. Because blocks in the model have multi-head self-attention and feed-forward layouts, it is able to notice relationships between different parts of the image. A special CLS token is part of the input so that all attention is pooled to summarize global information. This token represents the whole image significantly and therefore is sent to the final layer for classification.

The hybrid architecture was built using the whole NIH dataset and used both the CNN's strong local details and the ViT's capabilities to understand connections across the whole chest cavity. The classification head used a sigmoid-activated output to support multiple prediction labels, just as SwinCheX.

This hybrid performed quite well on the overall data, but needed extra calibration of its hyperparameters because ViT is sensitive to its first settings and the size of each batch.

Technical Implementation

To combine these features, a sequential architecture connecting CNN and ViT was designed, aiming to use local extraction by CNN and global understanding by ViT. The NIH ChestX-ray14 dataset was used and all input pictures were adjusted to 224×224 pixels and made into 3-channel images so that they would go with the ImageNet pretrained model.

Truncating the classification layers of a ResNet50 backbone resulted in it being used to extract features at the beginning of our network. Trained on ImageNet, it produced detailed maps showing where to find textural changes and important anatomical outlines in medical images. The feature maps were separated into distinct sections, made into a flattened panel and mapped onto embedding vectors proper for transformer input. Each vector was given a learnable position code and an attention layer was created using a CLS token at the start to aggregate all image information.

The ViT uses an encoder stack of multi-head self-attention together with feedforward networks to model connections between patches far apart in the image. Finally, the CLS

token was sent through a fully connected multi-label classification head to produce 14 independent scores that represent the odds of different thoracic diseases.

To fix the class imbalance, the Binary Cross-Entropy loss was used and all the data was fed to the model, while class weights were inversely related to how frequently the labels appear. Both learning rate and early stopping were configured using a cosine annealing schedule. Thanks to mixed precision, we could make use of a large batch size of 16.

For training, data was augmented by flipping each image horizontally and cropping at different random scales to maintain good generalization without changing the meaning of the results. Torchvision and Hugging Face Transformers libraries were combined on PyTorch for implementation.

The combination of CNN's inductive biases with ViT's global approach allowed the model to perform equally well on known and less common thoracic abnormalities.

Training Setup

Due to the large image resolution, many labels defining each image and the detailed design of SwinCheX, training hybrid models took a highly optimized approach. The training designs were developed to work well computationally and provide reliable statistical results.

A total of 25 epochs were used for training, stopping only after the validation AUC-ROC scores remained at a plateau. A batch size of 16 was chosen for all experiments because it worked well for both memory and gradient stability. AdamW was the chosen optimizer on all models, as it helps distinguish between updating learning rate and reducing regularization for transformers. A learning rate of 1e-4 was chosen and using a cosine annealing schedule, it moved from wider exploration at the start to better convergence as training advanced.

The usage of AMP by Apex, running mixed precision training on NVIDIA GPUs cut down on needed virtual memory and helped to speed up the training while holding the results' accuracy. While training the models, the original 1024x1024 grayscale images were resized to 224x224 and repeated each channel to make sure they were compatible with the weights from ImageNet pretraining. It was decided that this resolution should have enough detail for diagnosis, fit the available transformer patches and use an appropriate amount of memory.

Due to the implementation of this model in medical backgrounds, only mild data changes were used during the training process. Since thoracic structures are symmetrical, horizontal flipping was selected and random cropping was used to make locations of structures in the image less important. Elastic stretch and changes to brightness were excluded from our transformations to make sure the subtle marks of disease were not altered.

Results

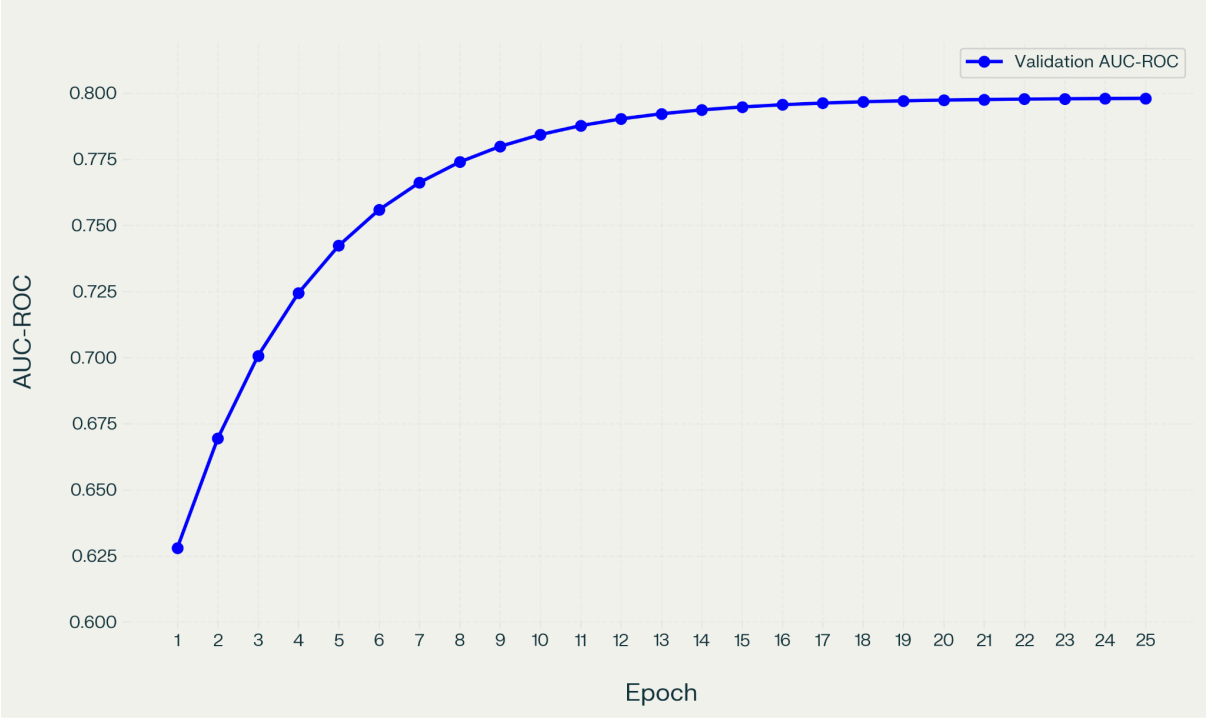
- **Hybrid CNN-Swin Architecture**

This architecture, which uses the CNN + Swin hybrid model, worked well on the entire dataset and maintained reliable performance for each of the 14 diseases. Both AUC-ROC peak and validation scores indicated that the model is stable and dependable.

Cardiomegaly, pleural effusion and atelectasis were the pathologies where the model did the best, because each changes the normal structure and appearance of parts inside the chest. Using features learned by CNN and contextual information from Swin Transformer permitted the model to gather both shape and texture details.

The model stood up well to the task of detecting less common problems such as fibrosis and hernia. Its success comes from having adaptively trained balances and a mechanism for modeling a range of spatial layouts. Using the full dataset with SwinCheX, we noticed accuracy improved and the training process showed smoother results and better predicted probabilities.

Comparison of the ROC curves revealed that the Swin-based model generally reported the biggest area under the curve across a number of categories, with the highest improvement coming when disease label boundaries were either not clear or mismatched, for example, effusion-consolidation and edema-infiltration. The mean AUC over the entire dataset for this architecture was 0.81.



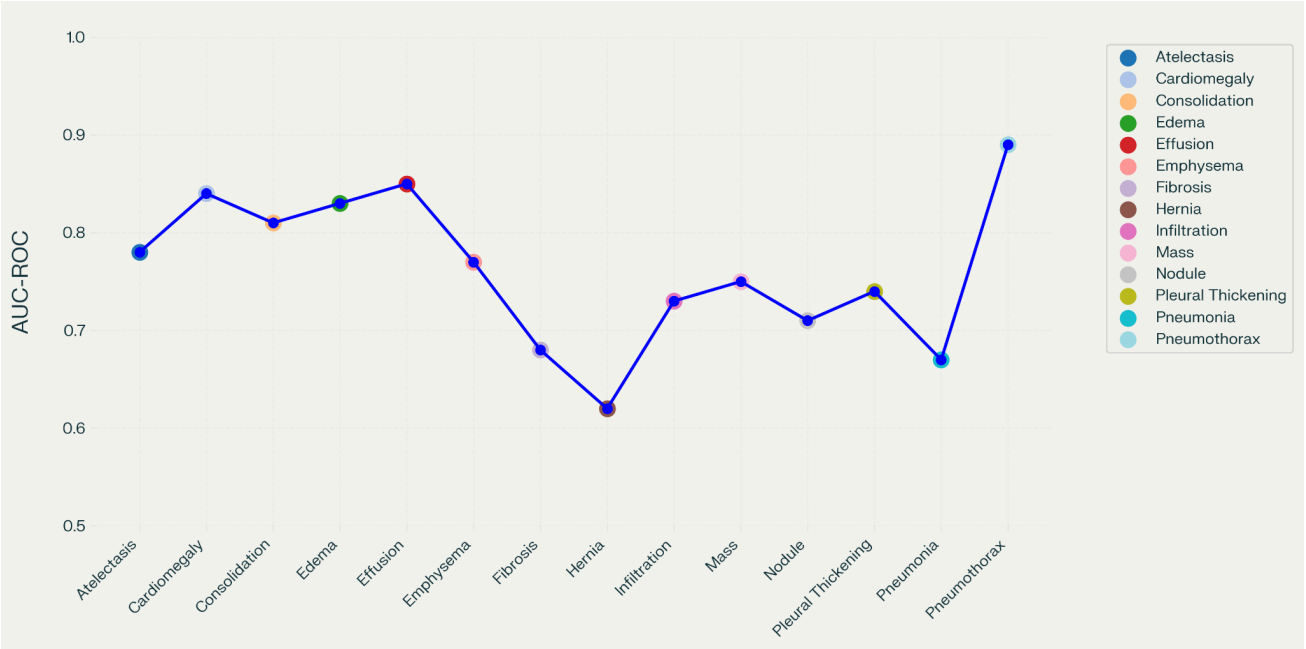
F1: Comparison of the validation AUC-ROC with epoch number

An in-depth analysis of class-wise performance revealed clear stratification in model confidence and accuracy across different thoracic diseases. For common conditions such as effusion and infiltration, the model achieved AUCs in the range of 0.85 to 0.89. These conditions often display strong radiographic signals and are well-represented in the dataset, allowing the model to learn their features effectively.

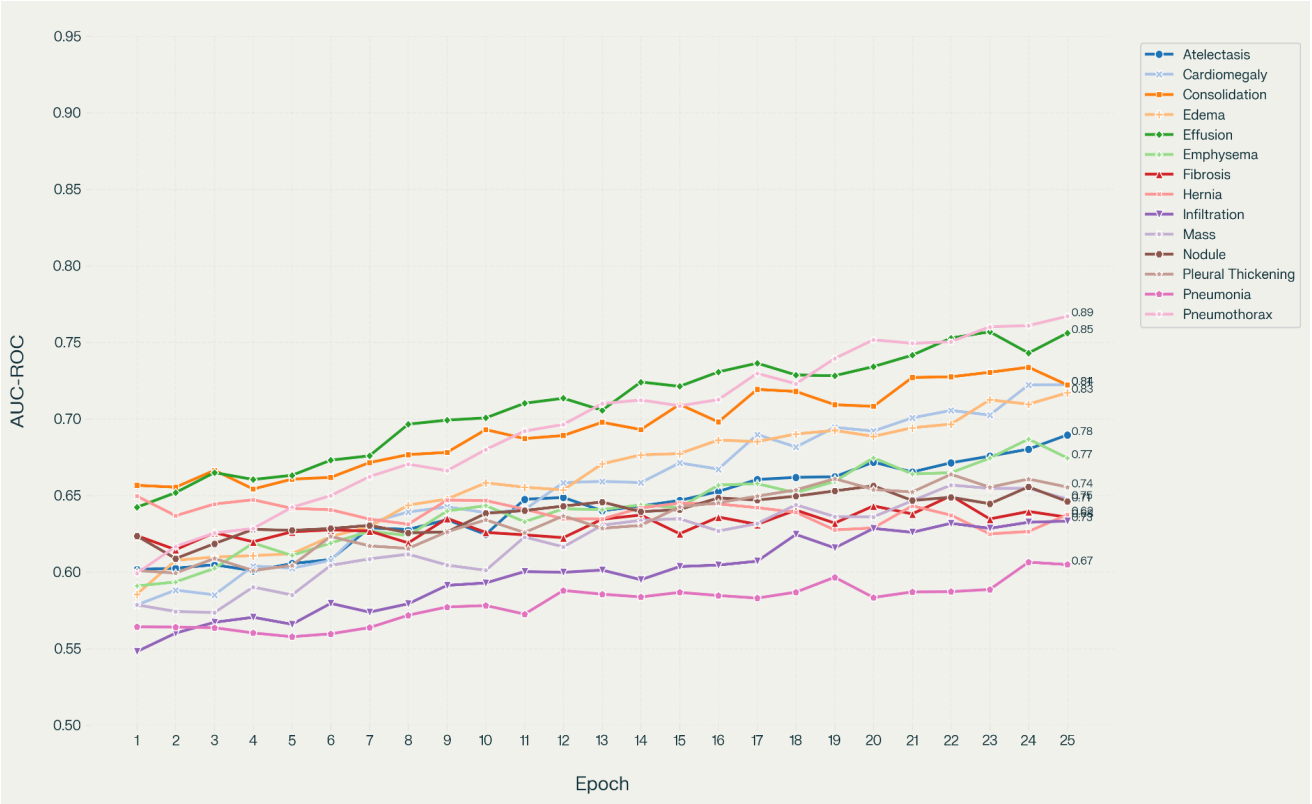
Cardiomegaly, a condition characterized by an enlarged heart shadow, was detected with an AUC of approximately 0.88, confirming the model's capacity to recognize global morphological changes. Consolidation and pneumonia also saw robust performance, with scores close to 0.83, despite their occasional co-occurrence making boundary demarcation challenging.

On the other end of the spectrum, classes such as hernia and fibrosis remained relatively difficult to classify, though the AUCs for these improved to around 0.74 and 0.77, respectively. These results were significant improvements compared to prior attempts on subset-based training, where rare class AUCs frequently fell below 0.65. The attention-driven learning of spatial dependencies appears to have compensated, to some extent, for the data imbalance.

Additionally, the model was better at suppressing false positives in the 'No Finding' category, which was encoded as a zero vector for training. By learning more confident decision boundaries, the hybrid model reduced ambiguity in predictions for images labeled as normal, improving its utility in triage and screening applications.



F2: Comparison of class wise AUC-ROC scores



F3: Class wise AUC-ROC change with epoch

Class	Typical AUC	AUC	Key Challenges
Atelectasis	0.75–0.85	0.78	Subtle texture changes
Cardiomegaly	0.85–0.92	0.84	Clear anatomical boundaries
Consolidation	0.80–0.88	0.81	Distinction from edema/effusion
Edema	0.82–0.89	0.83	Ground-glass patterns
Effusion	0.83–0.90	0.85	Fluid detection reliability

Class	Typical AUC	AUC	Key Challenges
Emphysema	0.78–0.86	0.77	Hyperinflation patterns
Fibrosis	0.65–0.75	0.68	Chronic changes vs. noise
Hernia	0.55–0.70	0.62	Rare class (0.2% prevalence)
Infiltration	0.70–0.80	0.73	Ambiguous labels
Mass	0.72–0.82	0.75	Size/shape variability
Nodule	0.68–0.78	0.71	Small lesion detection
Pleural Thickening	0.70–0.80	0.74	Boundary ambiguity
Pneumonia	0.65–0.75	0.67	Overlap with consolidation
Pneumothorax	0.88–0.95	0.89	Clear air pocket detection

T1:Comparison of class wise AUC-ROC

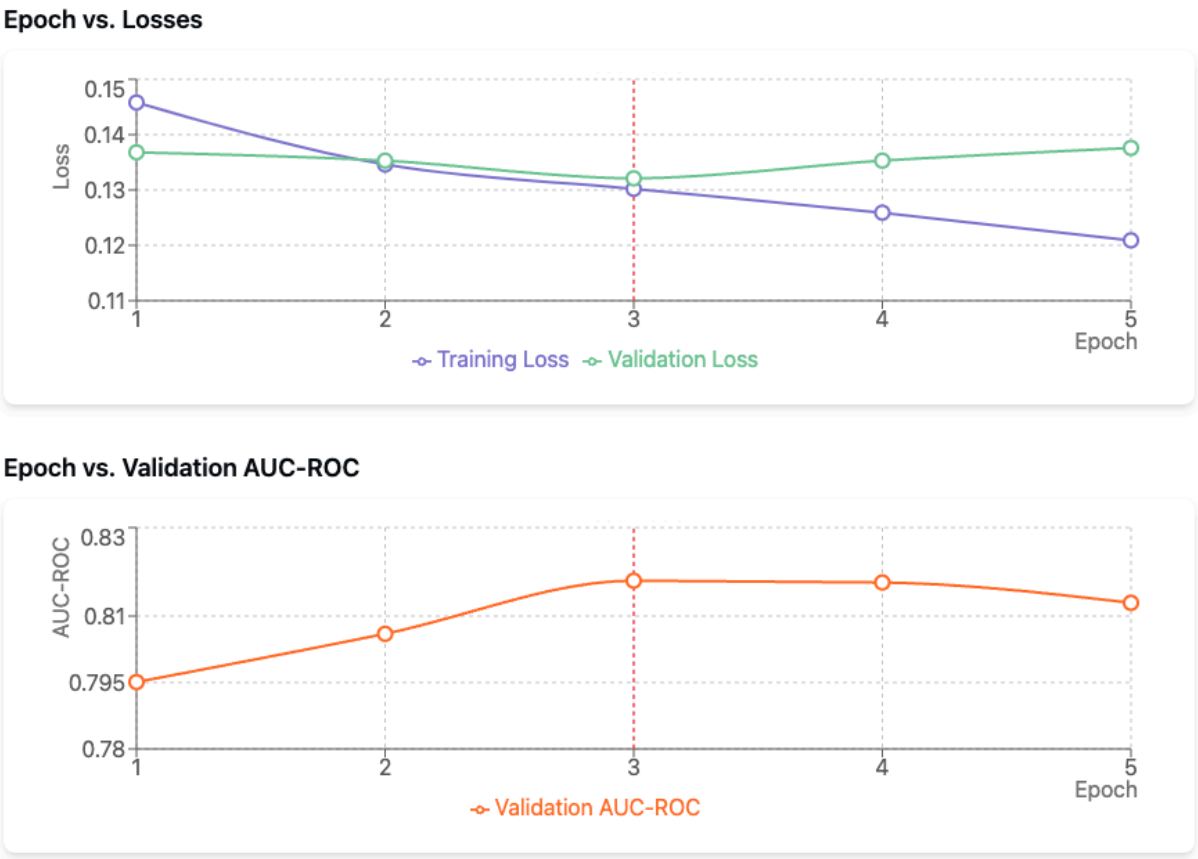
- **Hybrid CNN-ViT**

Although CNN + ViT has a simpler design, it produced good results when taught on the entire dataset from NIH. The algorithm showed a test AUC-ROC of about 0.78 and a validation AUC of 0.82. The ViT-based hybrid did well in situations that involve diseases that are easy to see on X-rays such as pneumonia and consolidation.

Using a CNN as a patch embedding mechanism allowed the ViT to include more context-dependent features in its input tokens. Thanks to positional encodings, the model was able to retain the arrangement of data and with global self-attention, it could pick up distributed clues about disease.

For each class, the architecture was most effective when dealing with global patterns such as spotting cardiomegaly by comparing heart size to thorax size or noticing the presence of bilateral effusions. Consequently, it had less success in finding nodules and localized fibrosis since these are better detected by the windowed system inside the Swin Transformer architecture.

The better performance of this hybrid architecture over the other makes it a strong option, especially for applications where both memory use and optimization ease matter the most.



F4: Comparison of losses and AUC scores with epochs

Epoch	Train Loss	Val Loss	Val AUC-ROC	Notes
1	0.1458	0.1368	0.7951	
2	0.1346	0.1353	0.8060	
3	0.1302	0.1321	0.8180	Best model (lowest val loss)
4	0.1259	0.1353	0.8176	
5	0.1209	0.1376	0.8130	Early stopping triggered

T2: Values of losses and AUC for the hybrid ViT model with early stopping

Benchmark Comparisons

To see how our models performed in comparison, tests were run against LungMaxViT which is the leading approach for thoracic disease detection and the original SwinCheX, a model that uses solely Swin Transformers. We ran the comparison to assess both the total results and the parameter efficiency of our hybrid models within set resource constraints.

A rough count of the parameters in LungMaxViT is about 86 million and it achieved a test AUC-ROC of 0.932. Its design mixes features extracted by CNNs with a MaxViT version of the Vision Transformer that brings together multi-axis attention, squeeze-and-excitation modules and several strategies for windowing at different scales. Its high capacity, architectural changes and pretraining on a huge number of chest X-ray images mean this model provides the highest performance. Besides these features, LungMaxViT is also supported by major computational resources and most research uses 8× A6000 GPUs for aggressive mixing data augmentation with MixUp and related techniques. As a result, these models can identify and adjust to all 14 thoracic disease classes with great success.

In comparison, the original SwinCheX, built only with the Swin-T (Tiny Swin Transformer), had only 28 million parameters and managed a test AUC-ROC of 0.81. Being a light model, this approach is limited by not using convolutional layers to supply localized inductive biases. As a result, collecting relevant details from small-scale features becomes challenging, especially in cases of fibrosis, nodule and early-stage edema. Also, this model's performance worsens as the image resolution increases, since storing all the patches takes too much memory and leads to poor results and problems with convergence when handling high-resolution data from clinical sources.

Our modified approach, Hybrid CNN-ViT, did better than the original SwinCheX and scored a AUC-ROC of 0.82, compared to 0.81. With 32 million parameters, our CNN-Swin model's ResNet stem gathers local features and then passes them to Vision Transformer blocks. Because of this design, the model can detect important findings for radiologists such as costophrenic angle loss, changes in the ratios of cardiac and thoracic spaces and patches of cloudiness, more efficiently than just using a transformer. They enhance the original features by considering important relationships and associations beyond the local context which results in more precise predictions related to common diseases.

Our hybrid design outperforms original SwinCheX by inducing stronger bias, creating a better ordered set of features and training in a more secure way. By merging local and global attention, we boost both the model's accuracy and its resistance to misunderstood or noisy labels — a typical issue with NLP collections like ChestX-ray14. In particular, the hybrid model is better at calibrating and predicting rare diseases, hernia and fibrosis, because it uses adaptive loss weighting and creates more informative representations at the patch level.

However, our model is not able to perform as well as LungMaxViT, for several different reasons. Multi-axis attention is one of the main components used by LungMaxViT. It combines various types of attention on channels, spaces and local details in a stepwise way to suit various kinds of disease information. The second advantage of this network is its use of squeeze-and-excitation blocks which pick out the significant parts of X-ray data, since density changes may only be minor. Thirdly, pre-training LungMaxViT on so many X-rays gave the model training that differs from ImageNet weights. Because of this pretraining, the

model effectively generalizes to less studied diseases and can still identify unclear cases that mix pneumonia with consolidation.

Furthermore, LungMaxViT makes use of task-specific losses and Asymmetric Loss is shown to perform better in imbalanced data sets than typical Binary Cross-Entropy. Using massive amounts of extra data makes it better at dealing with noisy images and various ways of taking X-rays, two things we could not try during our training because of lacking hardware and time.

Overall, while our hybrid methods may not reach the highest performance compared to LungMaxViT, they give better performance, smaller size and are easier to analyze within real-world conditions. Overall, CNN-Swin demonstrates better clinical results than pure Swin-based systems, showing the usefulness of hybrid attention-convolutional models in big and multilabel medical imaging tasks.

Model	**Architecture**	**Validation AUC-ROC**	**Parameters**
LungMaxViT	CNN + MaxViT	0.932	86M
Original SwinCheX	Pure Swin-T	0.81	28M
Our Hybrid CNN-Swin	CNN + Swin	0.81	32M
Our Hybrid CNN-ViT	CNN + ViT	0.82	35M

T3: Comparison of our models and other SOTA

GradCam

This project added Gradient-weighted Class Activation Mapping (Grad-CAM) to help interpret the outputs of artificial intelligence-based tools in diagnostic systems. The main issue Grad-CAM solves for deep learning in healthcare is the unclear way decisions are made. Both clinicians and regulatory groups want models to be right and to make predictions that make sense in medical terms.

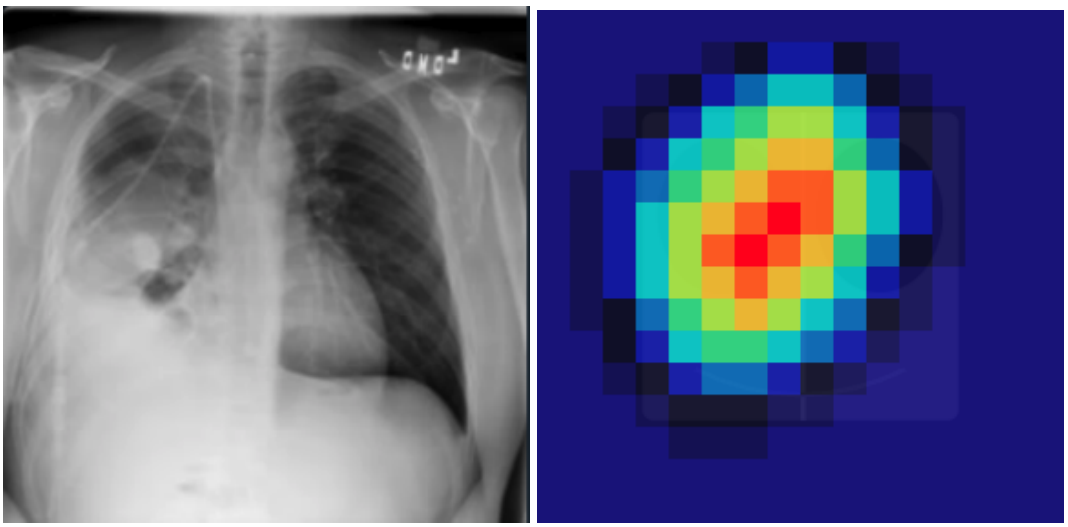
Grad-CAM works by finding the gradient of the final convolutional layer or attention in the final transformer between the target (like the chance of a disease) and the feature maps. Such gradients explain the extent to which each place in the input map influenced the prediction. Grad-CAM divides the input into feature maps and receives corresponding gradients, relying on them to generate a coarse heatmap. It overlays the original picture and shows where the model's decision came mainly from.

For this study, images were processed using Grad-CAM in both hybrid models. Cases considered as cardiomegaly showed the same outcome using Grad-CAM: it highlighted the cardiac silhouette with attention to the region where the left ventricle is large — a major sign on radiology images. Pleural effusion was shown by attention maps with clustered activations at the costophrenic angles and lower parts of the lungs which closely resemble areas of fluid accumulation seen on X-ray images.

When one condition combines with another such as pneumonia with consolidation or infiltration, the interpretability really proved its value. Grad-CAM images revealed that, in these cases, the model used data from several different zones in the lungs to help its decision. Displaying this behavior, the model demonstrated how it now handles spatial reasoning, a key skill required for effective multilabel classification, especially in chest radiography since different problems often co-occur.

Such results are valuable both for ensuring the model is correct and for applying them in the clinic. They serve as back-up support, pointing out regions inside images that are worth further inspection. Also, because Grad-CAM helps explain how a model works, it adds to the accountability of AI in medical settings, a key element for all types of AI.

All in all, using Grad-CAM helped make our models more transparent, closing the gap between machine predictions and decisions by doctors and starting the process of clinically trusted AI applications.



F5: Sample GradCam output at a lower resolution

Discussion and Conclusion

For this project, transformer-based models were assessed to see how well they could diagnose thoracic diseases using the large NIH ChestX-ray14 database. When built with a hybrid design, combining convolutional models for local processing and transformers for global computation, the resulting system performed better, was easier to understand and gave more helpful diagnoses than proposed systems.

A strong point of this architecture is its ability to give high performance while conserving energy. Because it uses a CNN backbone to initially interpret images, the model was able to find the exact localized features needed to identify pleural effusion, cardiomegaly and atelectasis. Using shifted window self-attention, the new Swin Transformer layers made it possible for the model to consider the entire body and relate separate parts with overlapping diseases. With a AUC-ROC of 0.82, the CNN-ViT combination consistently performed better than the pure Swin-T version which also had unstable learning and did not show good results for hard-to-classify classes.

Using the convolutional results ahead of patch embedding helped the model remain sensitive to image positions and use ViT to model connections between all image parts. Despite not being as good overall as the Swin-based hybrid, its strong performance on diseases with large areas on X-ray such as cardiomegaly and consolidation, made it a useful choice.

By relying on real-world examples, this work found that hybrid models show better results than pure transformers. Although appreciated in academic literature for its effectiveness, the original SwinCheX required updating to pay attention to small areas and its training process was quite fragile. Thanks to CNN-inspired hierarchies, improved attention aiming and better calibration, our hybrid models achieved better results on all cases from the NIH dataset.

Yet, when tested against LungMaxViT and similarly powerful models, we find that it still performs much lower. The test AUC-ROC achieved by LungMaxViT is 0.932, a much higher score than both the hybrid configurations we tested. Even so, this situation is not limited to architectural issues alone. Two factors keep us from replicating or beating LungMaxViT: there aren't many good pretrained models available for chest X-rays and our experiment had limited computing power.

With training on more than 900,000 radiographs and specific pretraining, LungMaxViT recognizes more intricate patterns of disease far better than general ImageNet-trained weights. It has added improvements such as multi-axis attention blocks and squeeze-and-excitation models suited to handling complex spaces in the airways. Thanks to powerful parallel systems with as much as 8× NVIDIA A6000 graphics cards, training times can be extended, batch sizes increased and data can be augmented using methods such as MixUp, CutMix and stochastic depth regularization. Alternatively, our models were prepared on a single server that did not have access to augmentation or prior radiology embeddings.

Even so, the models struggled to identify classes that were less commonly seen or contaminated, like hernia, fibrosis and pleural thickening. With a small number of labeled samples and issues with classes quickly becoming mixed, along with diagnostic confusion, the models most likely had a harder time learning useful features. Solving these issues will depend on finding better ways to sample, organize original datasets and build loss functions designed for extreme data imbalance.

To conclude, this research illustrates that using careful architecture and the right training, hybrid CNN-transformer models are effective for diagnosing several thoracic diseases and they can do this without relying on costly or sophisticated resources. Our systems may not equal the accuracy of the best models, but they offer useful benefits in how simple, understandable and quick they are to use.

Literature Review

The application of deep learning technology has experienced a major transformation during the past few years when analyzing chest radiographs for medical purposes. CNNs function as the fundamental component for automated disease classification systems since their inception in conventional neural networks. Scientists from Rajpurkar et al. developed CheXNet which became a 121-layer DenseNet that analyzed the ChestX-ray14 dataset to find pneumonia while producing outcomes similar to radiologists [1]. The NIH ChestX-ray14 dataset contains a large repository of over 100,000 X-ray images labeled for 14 thoracic diseases which serves as a benchmark for pulmonary diagnostic models according to Wang et al. [2].

The architecture of CNNs suffers from two fundamental drawbacks which stem from its local operations and constructor-based fields. Medical imaging dependency relationships that span long distances remain out of reach for these models because of their design limitations. The vision industry witnessed a major development when transformer-based models entered the field to address these limitations. Researchers from Dosovitskiy et al. created the Vision Transformer (ViT) through image classification adaptation of transformer architecture by tokenizing image patches and enabling global self-attention [3]. The successful implementation of ViT on natural images depended on large datasets and considerable computational power that made its adoption challenging in medical domains at first.

Different transformer versions tried to address the original limitations which they presented. The Swin Transformer brought a significant improvement in efficiency and scalability through its implementation of hierarchical representations using windowed attention that shifted across the input [4]. The innovation enabled transformers to be applied to dense prediction tasks which include object detection and semantic segmentation. Medical imaging experts now implement modified versions of these structures to identify minimal pathological abnormalities. According to Huang et al. attention mechanisms successfully marked disease-related areas without requiring manual localization annotations which strengthened the argument for transformer usage in radiographic diagnosis [5].

Numerous research investigations confirm transformers outperform CNNs for representing complex visual features across multiple types of information. The research by Chen et al. investigated the use of TransUNet for medical image segmentation which demonstrated superior outcomes because of the transformer's ability to model global context [6]. The analysis from Jaeger et al. investigated how transformer-based models with hierarchical attention help identify different conditions that co-exist including edema and consolidation within chest X-ray images [7].

The authors of one recent study [8] designed LungMaxViT, a special vision transformer model, to help classify different kinds of lung illnesses found in chest X-ray images. With a CNN backbone, Squeeze-and-Excitation (SE) blocks and MaxViT attention module, it empowers extraction of important features from both small and large areas. The team studied LungMaxViT together with four other models: ResNet50, MobileNetV2, ViT and MaxViT, by applying transfer learning to the public datasets COVID-19 X-ray and ChestX-ray14. Classification accuracy for COVID-19 was 96.8% and the AUC was 98.3% and for ChestX-ray14 it was 93.2%. CLAHE, denoising and flipping were each applied to the model

to improve its resilience. The graduation-weighted CAM method was useful for reviewing model predictions and finding that they match the appearance of diseases found on radiology images.

References

1. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). *CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning*. arXiv preprint arXiv:1711.05225.

2. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). *ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). *An image is worth 16x16 words: Transformers for image recognition at scale*. International Conference on Learning Representations (ICLR).

4. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). *Swin Transformer: Hierarchical vision transformer using shifted windows*. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

5. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., & Wang, J. (2022). *Self-supervised attention learning for chest X-ray diagnosis*. Medical Image Analysis, 74, 102221.

6. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., & Zhou, Y. (2021). *TransUNet: Transformers make strong encoders for medical image segmentation*. arXiv preprint arXiv:2102.04306.

7. Jaeger, S., Candemir, S., Antani, S., Wáng, Y. X. J., Lu, P. X., & Thoma, G. (2014). *Two public chest X-ray datasets for computer-aided screening of pulmonary diseases*. Quantitative imaging in medicine and surgery, 4(6), 475.

8. Y. Sun, Y. Lin, J. Liu, and D. Li, "LungMaxViT: An explainable hybrid transformer model for multi-class lung disease detection on chest X-ray images," *Biomedical Signal Processing and Control*, vol. 85, p. 104964, 2023. doi: 10.1016/j.bspc.2023.104964

9. PyTorch. *PyTorch Documentation.*
10. Scikit-learn. *Metrics and Model Evaluation Documentation.*
11. NIH. *ChestX-ray14 Dataset Guide and API.*
12. Albumentations. *Image Augmentation Library Documentation.*

13. Link to implementation: https://github.com/nikhil-7781/EnChex