CheX-Swin: Chest X-Ray Classification using Swin Transformer

Nick DiSanto
Department of Computer Science
Vanderbilt University
nicolas.c.disanto@vanderbilt.edu

Abstract

Accurate classification of chest X-rays is critical for early diagnosis of pulmonary diseases, especially in clinical settings where professional opinions are increasingly difficult to come by. Thus, there exists a need for automated procedures that can provide quick and accurate diagnoses. In this work, we evaluate whether the **Swin Transformer**, a hierarchical Vision Transformer that employs shifted window self-attention, can outperform traditional CNN-based models such as ResNet18 and VGG19 on medical classification tasks. We perform our evaluation on the ChestMNIST dataset, a large and highly imbalanced benchmark for robust classification. In order to verify whether Swin is a comparable alternative to CNN baselines through its ability to capture global context and long-range dependencies, we implemented all three models with identical training conditions so that they focus exclusively on subtle disease patterns. While our results show high classification accuracy (>93%) across all three of the models, Swin consistently performs the highest in both overall and classwise AUC, which was our metric of interest due to its sensitivity to class imbalance. Additionally, all three of our implemented and fine-tuned models surpassed the performance of the original MedMNISTv2 benchmark. These findings reinforce the promise of transformer-based architectures in advancing clinical AI applications for medical imaging tasks requiring both local detail and global contextual understanding. All source code is publicly available at GitHub.

1 Introduction

Even in an era where chest X-rays are one of the most commonly utilized imaging modalities due to their wide availability and low cost, pulmonary diseases remain a leading cause of mortality worldwide [16]. However, interpreting chest X-rays requires specialized expertise since slight misinterpretation can lead to dangerously incorrect diagnoses. New developments in artificial intelligence, and deep learning in particular, have demonstrated substantial promise in supporting X-ray classification and clinical decision-making. Nevertheless, many traditional deep learning models have trouble with broad generalizability, especially in environments where subtle patterns may occur over large separated spatial regions.

An additional challenge in automated chest X-ray classification lies in the data's labeling and organization. The ChestMNIST dataset [21], for example, which is a subset of the gold-standard MedMNIST dataset [22], is a multilabel dataset where a sample from a single patient may contain a single disease, multiple diseases, or no disease at all. Furthermore, this dataset is severely imbalanced, with certain conditions (e.g., hernia, fibrosis) appearing much less often than others (e.g., effusion, infiltration). This imbalance can easily complicate model training and lead to biased classifiers that fail to detect rare but clinically critical pathologies because they learn to predict "false" on every sample to achieve high accuracy. Consequently, for these problems, sensitivity-based metrics such as

Preprint. Under review.

AUC and F1 score are increasingly valuable indicators of high performance, rather than just accuracy or loss alone. Addressing class imbalance and improving model sensitivity remains a critical research gap in clinical AI for chest radiography.

The prevailing paradigm for image classification in both the medical domain and in natural images has been on using convolutional neural networks (CNNs), due to their strength in capturing local low-level patterns. However, CNNs struggle to model long-range dependencies across images as a whole, which becomes particularly notable in medical imaging domains like epidemelogy, where different parts of an image can combine to paint a broader picture. To overcome these obstacles, Vision Transformers (ViTs) have been proposed as alternatives to CNNs since they offer self-attention mechanisms that can capture global relationships. Since standard ViTs are computationally expensive, the Swin Transformer was proposed by Liu et al. [10] to focus on this problem. Swin applies attention within local shifted windows while building hierarchical feature maps, which aims to maintain the efficiency of CNNs on a local level and balance larger-scale global context modeling. While Swin's unique architecture has made it a strong performer on many baseline datasets like ImageNet, its clinical potential in empirical domains remains underexplored.

Therefore, in this study, we investigate whether the Swin Transformer can perform well for multilabel classification on ChestMNIST by benchmarking its performance against ResNet18 and VGG19, two commonly-used convolutional models. Our experimental results show that the Swin Transformer rivals both models in loss and accuracy, while achieving an AUC of 0.7799, outperforming ResNet18 (0.7552), VGG19 (0.7341), and the original published MedMNISTv2 baseline (0.75). These results suggest that hierarchical transformer architectures like Swin may be worth further exploration in complex clinical classification tasks, especially with imbalanced datasets. Our study seeks to highlight the potential for integrating ViTs into the medical domain and motivate future work in continuing further development.

2 Related Work

2.1 Vision Transformers

While transformers were initially introduced in natural language processing [19], the Vision Transformer (ViT), introduced by Dosovitskiy et al. [3] in 2021, adapted self-attention mechanisms towards image data. This method of splitting images into patches and applying standard transformer encoders was powerful, but nevertheless early ViTs required extremely large training datasets because of their lack of inductive biases.

The Swin Transformer [10] was introduced with a unique architecture to address this shortcoming, as it applies attention locally within non-overlapping windows, as well as shifting windows between layers to capture cross-region dependencies. More specifically, Swin operates by building hierarchical feature representations through patch merging. This approach effectively mimicked CNN multiscaling and immediately achieved state-of-the-art results on standard vision tasks like image restoration [7]. Additionally, Swin has been applied to domain-specific applications, showing promising outcomes in both the medical domain [8] and in other unique modalities [5]. These outcomes not only demonstrate Swin's immediate impact in empirical applications, but also suggests its potential for continual development in the medical domain, where datasets are becoming increasingly powerful and complex.

2.2 Deep Learning for Chest X-Rays

Within the domain of chest X-ray classification specifically, deep learning has recently become a promising approach for automated and scalable analysis. For example, CheXNet [14] demonstrated that a DenseNet-121 CNN could achieve radiologist-level performance on some pneumonia detection tasks, and was later expanded into multilabel classification [13].

Other approaches have also explored this domain beyond convolutional architectures. For example, Park et al. [11] demonstrated that pretraining Vision Transformers using self-supervised learning on large unlabeled chest X-ray datasets can enhance downstream diagnostic performance on rare pathologies. Similarly, Tiu et al. [17] revealed that self-supervised ViT models can achieve radiologist-comparable accuracy. These modern approaches continue to show the possibilities of transformer-based methods in medical imaging, further motivating our approach.

2.3 CNNs vs. Transformers for Classification Tasks

There are a few architectural differences between CNNs and Transformers that have been explored in recent research and set the models apart in their empirical use. To start, a CNN processes an image through a hierarchy of layers, with convolutional filters that operate over a small receptive field and iteratively build complex representations. While this hierarchical design captures fine-grained structures well, it struggles to model long-range dependencies without stacking many layers. Transformers, in contrast, use self-attention mechanisms to allow every patch to directly attend to every other one, regardless of spatial distance, allowing for early layers to begin to capture global context.

While ViTs originally had a reputation of requiring massive pretraining datasets to outperform CNNs, newer strategies like DeiT [18] have demonstrated that ViTs can be trained effectively on mid-sized datasets using strong augmentation and knowledge distillation. Additionally, in the medical domain, Kim et al. [6] found that Vision Transformers like Swin can outperform CNNs like ResNet in a variety of medical image classification tasks with small datasets. There are also some unique hybrid architectures like CoAtNet [1] that have shown that combining both convolutional operations and attention layers can begin to leverage both strengths.

3 Methods

3.1 Dataset

We utilized the ChestMNIST dataset, a multilabel subset of MedMNISTv2 [22], as our classification baseline. MedMNIST is considered a gold-standard in the medical community: a curated benchmark composed of lightweight medical imaging datasets. ChestMNIST itself contains grayscale chest radiographs annotated for 14 different thoracic diseases, with each image labeled with binary indicators (presence or absence) for each condition.

Figure 1: ChestMNIST Samples with Example Classifications [X]

The ChestMNIST dataset comprises 112,120 frontal chest X-ray images, preprocessed to a uniform resolution of 28×28 pixels. Given the small image size and binarized labels, ChestMNIST serves as a computationally efficient proxy for larger, high-resolution clinical datasets. However, the dataset is a challenging task for large-scale classification because it is inherently imbalanced; most individual samples contain a single condition, so negatives dominate the labeling.

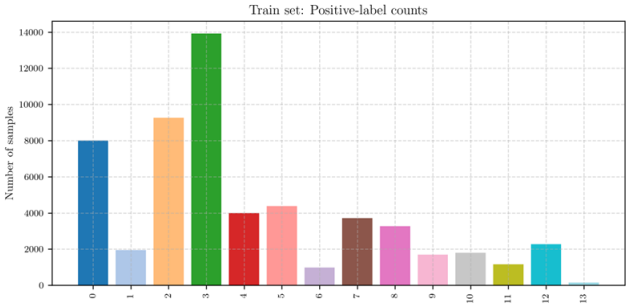


Figure 2: ChestMNIST Training Set Class Imbalance

3.2 Baseline Models

To establish an apples-to-apples baseline for comparison, we selected two widely-used and highly-considered convolutional architectures: ResNet18 and VGG19. Both of these models have demonstrated success in image classification tasks for years, and have even recently been validated within chest X-ray analysis [2, 12, 20].

ResNet18 [4] rapidly gained recognition upon its introduction in 2015. Its introduction of residual learning, which allows gradients to propagate more easily, allowed for deeper training without vanishing gradients. **VGG19** [15], on the other hand, is a more traditional deep CNN, with standard sequential 3×3 convolutional layers with periodic max-pooling. Although deeper than ResNet18, VGG19 lacks residual connections and thus often lacks deeper optimization potential.

Both ResNet18 and VGG19 encode strong spatial locality priors through convolutional operations. However, they contain key limitations in modeling long-range dependencies, which are critical for connecting spatially separated findings. This motivates our exploration of the Swin Transformer, a more advanced and modern architecture.

3.3 Swin Transformer

3.3.1 Architecture Overview

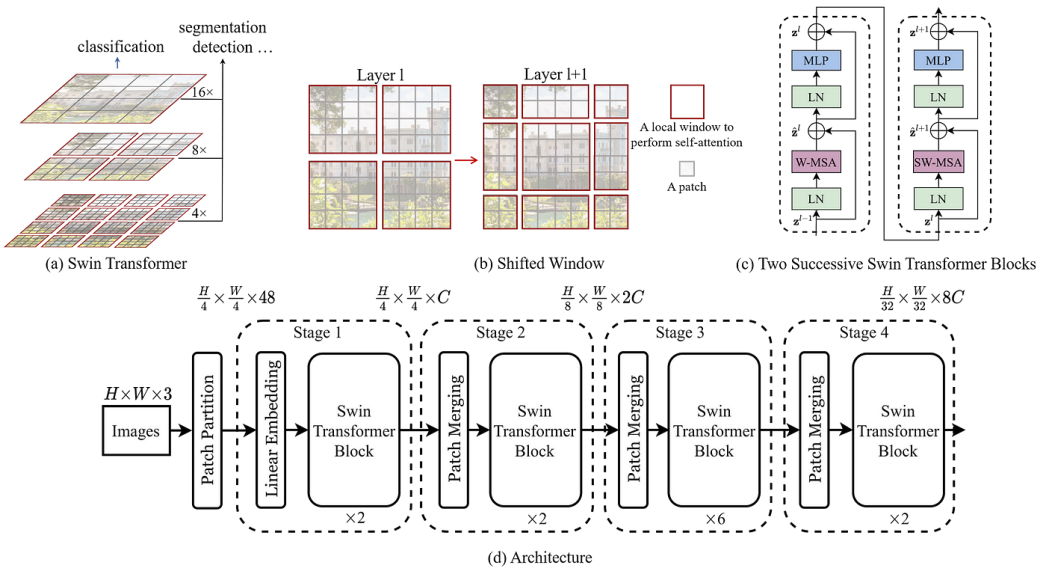


Figure 3: Swin Transformer Architecture

The Swin Transformer [10], as seen in Figure 3 marks a substantial advancement over traditional CNNs and vanilla ViTs for image processing tasks. Unlike standard ViTs, which apply global self-attention across all patches, the Swin Transformer applies self-attention within non-overlapping local windows and introduces a hierarchical feature representation through patch merging. These innovations combine the benefits of CNNs' local receptive fields and hierarchical features with the flexibility of transformer-based global modeling.

Swin processes an input image through four sequential stages. Each stage consists of multiple Swin Transformer blocks, where each block performs multi-head self-attention restricted to local windows followed by multilayer perceptrons (MLPs) applied to each patch. After every stage, patch merging reduces spatial resolution while expanding the feature dimension, creating a hierarchical representation.

The basic Swin building block includes:

- Window-based Multi-head Self-Attention (W-MSA), which is restricted to patches within a window.
- Shifted Window Multi-head Self-Attention (SW-MSA), which extends across shifted windows to enable cross-window connections.

Along with these attention features, each building block also includes residual connections and layer normalization. This emulates CNNs via progressively merging patches and allows for multiscale feature learning on complex visual tasks.

3.3.2 Shifted Window Attention

Since window-based attention can introduce a limited receptive field, Swin employs a novel shifted window mechanism. In every alternate layer, the window partitioning is shifted by half the window size, so that the boundaries of the previous windows are now covered by the new windows.

The advantage of this shifted-window approach is two-fold: it allows for increased information flow, and still maintains a similar cost. It also ensures that all tokens within an image can attend to each other after a few layers while keeping the quadratic attention complexity confined within small windows. This is essential for chest X-ray samples, where multiple conditions may be occurring at vastly different parts of a single image.

3.3.3 Hierarchical Feature Representation

The Swin Transformer constructs a feature hierarchy by performing patch merging between stages, with adjacent patches being concatenated and passed through a linear layer. This reduces spatial dimensions and increasing feature depth, a hierarchical representation that is particularly important for chest X-rays, where findings can vary dramatically in size and models must capture both fine-grained and coarse-level features.

4 Experiments

4.1 Implementation

We implemented all models using PyTorch, leveraging open-source libraries such as timm for model building and torchmetrics for evaluation. Additionally, we followed standard coding practices that recommend configurable and generalizable code, which allowed us to swap easily between models and hyperparameter settings as our experimentation progressed.

Each of our models was pretrained via ImageNet to accelerate convergence and allow computation to focus solely on our specific domain. Our choice of optimizer was AdamW (weighted Adam), which was configured with a learning rate of 1×10^{-3}, a momentum of 0.9, and a cosine annealing learning rate scheduling. We set the batch size to 128 and trained each model for 10 epochs to maintain impartiality when comparing between them.

Training was conducted on an NVIDIA CUDA GPU, operating over fixed seeds to maintain reproducibility. Additionally, as is standard with deep learning experimentation, early stopping was implemented based on validation loss, with the best-performing model being reloaded for downstream inference and evaluation. To maintain thoroughness, we conducted our experimentation over three independent trials for each model and reported averaged performance metrics to account for variance due to random initialization and batch sampling.

4.2 Data Preparation

We utilized the default official ChestMNIST splits, which consist of:

- 78,000 training images
- 11,200 validation images
- 22,920 test images

Before any training, each image was first resized to 224×224 pixels to match the input requirements of the pretrained networks. We also performed random data augmentation transformations on the training data, including cropping, flipping, and rotating, with a goal of simulating natural variability in data acquisition to improve generalization. Validation and test sets, on the other hand, underwent only resizing and center cropping, without any augmentation.

Additionally, given the severe class imbalance, we attempted a re-weighted sampling strategy for the training set. This included assigning a weight inversely proportional to an image's respective frequency, to attempt to scale minority classes proportionally. Further class imbalance mitigations were also considered in the loss function selection and implementation.

4.3 Evaluation Metrics and Loss Functions

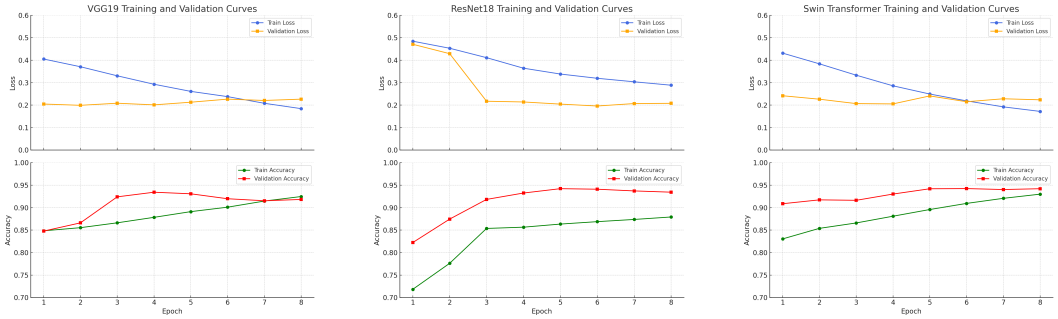
Given the severe class imbalance in ChestMNIST, we focused on evaluation metrics sensitive to rare classes rather than simply depending on overall accuracy. Thus, the primary metric that we focused on was the Area Under the ROC Curve (AUC), calculated both per-class and macro-averaged across all classes. AUC evaluates ranking quality and is an ideal metric for imbalanced multilabel tasks. While we still computed accuracy and loss over each model's epochs, we didn't assign much weight to their outcomes given that we expected all three models to converge to strong values quickly.

For our loss function, we implemented several different approaches but settled on Focal Loss [9]. Focal Loss is especially beneficial in highly imbalanced datasets because it dynamically scales to focus on more difficult examples (in our example, positives). More specifically, we implemented focal loss with $\alpha = 1.0$ and $\gamma = 2.0$, following common conventions to balance the contributions of positive and negative samples.

5 Results

5.1 Quantitative Results

As expected before our experimentation, all three of our models quickly converged to low training and validation loss, verifying our hypothesis that accuracy would not be a discriminatory metric between them. VGG19, ResNet18, and Swin Transformer all reached over 93% classification accuracy in fewer than 10 epochs each:



(a) VGG19 (b) ResNet18 (c) Swin Transformer

Figure 4: Comparison of loss (top) and accuracy (bottom) for training and validation across models.

While we found that Swin's CNN counterparts slightly outperformed it in accuracy, Swin consistently performed the best in AUC, which we hypothesized would be a far more indicative measure of performance in an imbalanced dataset. We even compared Swin's sensitivity-based output metrics with MedMNISTv2's original baseline, and found that Swin outmatched all three of its competitors in both overall and class-wise AUC performance. Table 2 shows even more specific results, with Swin outperforming its CNN counterparts in 12 of the 14 classes (including on all of the rarest labels):

Table 1: Performance Comparison Across Each Model

Model / Baseline	Accuracy	AUC
VGG19	0.9386	0.7341
ResNet18	0.9388	0.7552
Swin Transformer	0.9335	**0.7799**
MedMNISTv2 Baseline	0.9218	0.7500

Table 2: Class-wise AUC Comparison Across Each Model

Class	VGG19	ResNet18	Swin Transformer
Atelectasis	0.7089	0.7522	**0.7601**
Cardiomegaly	0.7622	0.8687	**0.8862**
Effusion	0.7888	0.8418	**0.8514**
Infiltration	0.6258	0.6236	**0.6570**
Mass	0.6306	0.6910	**0.7703**
Nodule	0.5953	0.6375	**0.6530**
Pneumonia	0.5868	0.6444	**0.6781**
Pneumothorax	0.7364	0.8183	**0.8540**
Consolidation	0.7285	**0.7470**	0.6981
Edema	0.8273	0.8597	**0.8744**
Emphysema	0.7166	0.8472	**0.8888**
Fibrosis	0.6458	0.7143	**0.7192**
Pleural Thickening	0.6315	**0.7018**	0.6973
Hernia	0.8494	0.8289	**0.8658**

5.2 Comparison to Existing Literature

As seen in Table 1, results exceed the published MedMNISTv2 baselines, which reported an average of 0.75 AUC and 0.9218 accuracy for ChestMNIST classification [22]. Compared to the MedMNISTv2 baselines, all three of our models surpassed the reported accuracy, but only the Swin Transformer significantly outperformed the baseline AUC.

Previous deep learning models for chest X-ray classification predominantly employed deep CNN architectures which, while highly successful, did not explicitly address the need for long-range contextual reasoning within images. Our results corroborate findings that show that self-attention architectures can rival CNN baselines with sufficient training and demonstrate that the Swin Transformer's shifted window attention and hierarchical representation confer measurable advantages on ChestMNIST.

Overall, our findings reinforce the emerging consensus that transformer-based models can rival, or even outperform, convolutional architectures in medical image classification, especially when sensitivity to rare pathologies is critical.

6 Discussion

6.1 Interpretation and Implications

Our results confirm the growing trend that vision transformers can match or exceed CNN performance in medical imaging tasks when adequately fine-tuned towards a specific domain. The Swin Transformer's skill of juggling local fine-grained feature extraction with global context modeling is particularly well-suited for chest X-ray interpretation, where abnormalities may be spatially diffuse, subtle, or involve multiple anatomical regions.

Additionally, Swin's high AUC score demonstrates better discriminatory power across both common and rare diseases and adds further emphasis to the significance of sensitivity-focused metrics in imbalanced datasets, showing that simply focusing on high accuracy may be naïve.

These results represent important indications and takeaways:

- Hierarchical vision transformers, such as Swin, may be beneficial clinical AI tools in tasks with rare pathology detection.
- It is possible to adapt ViTs to real-world clinical settings without overly large datasets.
- Model architecture selection in medical imaging should increasingly consider attention-based mechanisms when there is a domain need for capturing global relationships.

6.2 Limitations and Future Work

In light of our encouraging results, several limitations in our approach and methodology must be acknowledged. First, our ChestMNIST dataset consists of low-resolution 28×28 images, resized to 224×224 for compatibility with ImageNet-pretrained models. While this was sufficient for experimental validation, it likely fails to perfectly replicate the complexity of high-resolution clinical chest radiographs. Future work should evaluate Swin Transformers directly on full-resolution clinical datasets to ensure that high performance is maintained.

Additionally, although Swin outperformed CNNs on most of the class-wise AUC outcomes, there were still two disease classes (consolidation and pleural thickening) that posed notable challenges. Further optimization of class-balanced loss functions and uncertainty modeling may be needed to improve sensitivity on these rare, but extremely important classes.

Third, due to computational constraints, we conducted training on pretrained weights at a slightly higher learning rate for only 10 epochs. While all three of our models converged quickly and consistently, extended training would allow for particularly robustness. This is particularly true for Swin, given the data-hungry nature of ViTs, so we expect its high performance would likely continue to separate itself from its competitors if it is given further training and refinement.

7 Conclusion

To conclude, this work explored the hypothesis that the Swin Transformer would be a high-performing architecture on multilabeled chest X-ray classification. Our results validate this hypothesis and demonstrate a substantial outperformance compared to traditional convolutional neural networks. Swin achieved notably higher AUC scores while maintaining competitive accuracy, additionally outperforming MedMNISTv2's original benchmark metrics, indicating superior sensitivity and generalization to rare disease classes.

Our findings add to the growing body of evidence that vision transformers, when properly adapted, can function as effective tools for medical imaging applications. The Swin Transformer offers a promising way forward for building more accurate, sensitive, and clinically reliable AI models for diagnostic radiology tasks. This analysis emphasizes the critical role of architectural choices in model development and suggests that hierarchical attention-based methods will play an increasingly pivotal role in the next generation of medical AI systems.

References

[1] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3965–3977. Curran Associates, Inc., 2021.

[2] Nilanjan Dey, Yu-Dong Zhang, V. Rajinikanth, R. Pugalenthi, and N. Sri Madhava Raja. Customized vgg19 architecture for pneumonia detection in chest x-rays. *Pattern Recognition Letters*, 143:67–74, 2021.

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

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[5] Xin He, Yong Zhou, Jiaqi Zhao, Di Zhang, Rui Yao, and Yong Xue. Swin transformer embedding unet for remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022.

[6] Bum Jun Kim, Hyeyeon Choi, Hyeonah Jang, Dong Gu Lee, Wonseok Jeong, and Sang Woo Kim. Improved robustness of vision transformers via prelayernorm in patch embedding. *Pattern Recognition*, 141:109659, 2023.

[7] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1833–1844, October 2021.

[8] Ailiang Lin, Bingzhi Chen, Jiayu Xu, Zheng Zhang, Guangming Lu, and David Zhang. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement*, 71:1–15, 2022.

[9] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[10] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October 2021.

[11] Sangjoon Park, Gwanghyun Kim, Yujin Oh, Joon Beom Seo, Sang Min Lee, Jin Hwan Kim, Sungjun Moon, Jae-Kwang Lim, Chang Min Park, and Jong Chul Ye. Self-evolving vision transformer for chest x-ray diagnosis through knowledge distillation. *Nat. Commun.*, 13(1):3848, July 2022.

[12] Sheetal Rajpal, Navin Lakhyani, Ayush Kumar Singh, Rishav Kohli, and Naveen Kumar. Using handpicked features in conjunction with resnet-50 for improved detection of covid-19 from chest x-ray images. *Chaos, Solitons Fractals*, 145:110749, 2021.

[13] Pranav Rajpurkar, Jeremy Irvin, Robyn L. Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P. Langlotz, Bhavik N. Patel, Kristen W. Yeom, Katie Shpanskaya, Francis G. Blankenberg, Jayne Seekins, Timothy J. Amrhein, David A. Mong, Safwan S. Halabi, Evan J. Zucker, Andrew Y. Ng, and Matthew P. Lungren. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLOS Medicine*, 15(11):1–17, 11 2018.

[14] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, 2017.

[15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[16] Don D Sin and S F Paul Man. Chronic obstructive pulmonary disease as a risk factor for cardiovascular morbidity and mortality. *Proc. Am. Thorac. Soc.*, 2(1):8–11, 2005.

[17] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nat. Biomed. Eng.*, 6(12):1399–1406, December 2022.

[18] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers amp; distillation through attention. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 18–24 Jul 2021.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[20] A. Victor Ikechukwu, S. Murali, R. Deepu, and R.C. Shivamurthy. Resnet-50 vs vgg-19 vs training from scratch: A comparative analysis of the segmentation and classification of pneumonia from chest x-ray images. *Global Transitions Proceedings*, 2(2):375–381, 2021. International Conference on Computing System and its Applications (ICCSA- 2021).

[21] J Yang, R Shi, and B Ni. MedMNIST classification decathlon: A lightweight AutoML benchmark for medical image analysis. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*.

[22] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Nature Scientific Data*, 2022.