Medical Image Classification with Deep Neural Networks

Alon Shapiro
Department of Decision Sciences
HEC Montreal University
Montreal, Canada

Abstract

The medical image analysis field, traditionally attributed to clinical radiologists, has also seen enormous advancements as a result of the recent Artificial Intelligence-fuelled image analysis boom, with the introduction of advanced deep learning models such as deep convolutional neural networks and vision transformers. Here, we dive into some of the associated challenges of adopting such a new technology in a clinical setting, including model interpretability and training of such models. Various deep neural network models are proposed and trained on chestMNIST, a publicly available dataset of chest X-Rays. Models are compared to previously published benchmarks and results and implications are discussed. Code is available in: `https://colab.research.google.com/drive/1IuAH_wrK0HGx7tHjEmbbVJI4NgCApBT7?usp=sharing`

1 Introduction

Deep learning (DL) models for image data analysis have been at the forefront of recent advancements, with recent introduction of state-of-the-art architectures such as ResNet (He, 2016) and DenseNet (Huang, 2017), which are both based on the traditional convolutional neural network (CNN) architecture, and the more modern Vision Transformer (ViT, Dosovitskiy, 2021). However, the adoption of such models has not been equally quick to spread across different domains, particularly in highly regulated industries such as healthcare and clinical settings, which pose particular barriers for adoption such as high reliability, safety, and regulatory compliance requirements. As stated by Chan et al. in 2019, there have been numerous speculations about the future of radiologists as medical professionals given the impressive advancements in machine learning and Artificial Intelligence. However, developing DL models capable of performing as good or better than trained clinical radiologists is no easy task.

2 Background

2.1 Summary of Existing Literature

CNN's and more recently ViT's have been shown to perform very well on medical image analysis across various types of tasks - image segmentation, detection, and classification, just to name a few (Shamshad, 2023). ViT's have also proven to better learn highly complex representations as well as long-range interactions, compared to CNN's. However, some challenges still exist with these highly complex models.

For one, the already high computational complexity of DL architectures typically scales quadratically as a function of the input image size, which is exacerbated by the fact that medical images are usually

Foundation Models for Science Workshop,38th Conference on Neural Information Processing Systems (NeurIPS 2024).

very dense in pixel resolution (Azad, 2024). This leads to often requiring downsizing of images prior to analysis and therefore loss of information.

Interpretability is another challenge that is amplified when DL is applied in a high-stakes clinical environment (Reyes, 2020). DL models are often viewed as 'black-box' models due to the numerous levels of calculation, which make it difficult for humans to understand the impact of input features and predictions.

With both these challenges (computational complexity and interpretability), the determining factor is often the model complexity - more complex models such as ViT's incur higher computational costs and are harder to interpret when compared to simpler models such as a simple CNN. However, the benefit more complex models provide are often far better performance and generalization abilities.

2.2 Positioning of Current Work

In this project, we set out to explore the challenges and feasibility related to training DL models to classify frontal thoracic X-ray images for a number of pathologies. Aspects of model architecture complexity, classification performance, and training computational costs and associated challenges are summarized, and the feasibility of deployment in a real clinical setting is discussed. The trade-off between model complexity and performance and computational cost are discussed. Model performance is compared to model and human performance benchmarks found in literature.

3 Methodology

3.1 Dataset Description

ChestMNIST dataset, introduced by Yang et al. 2023, compiles 112,120 chest X-Ray images with corresponding 14 disease binary labels (78,468:11,219:22,433 training:validation:testing split). Each image is of size 1 x 224 x 224, which contains one grayscale channel, and target vector contains 14 binary variables, indicating the presence of one of 14 pathologies which can be diagnosed from a chest X-Ray scan. Along with the dataset introduction, Yang et al. provide numerous Area Under Curve (AUC) and accuracy benchmarks for multi-label classification with different models which are used to compare models trained in our work.

As the pre-trained models that were chosen (see section 3.2) all take 3-channel (RGB) images as input, images were preprocessed to contain 3 identical (duplicated) grayscale channels.

Given the nature of the data, the dataset is slightly imbalanced, with the most balanced class only having 17.7% positive labels, and the least balanced class having only 0.2% positive labels (see Fig. 1).

3.2 Model Architectures

Eight different computer-vision model architectures were considered in this work (see table 1). A Vanilla Convolutional Neural Network (Vanilla CNN) trained from scratch with 3 convolution layers followed by a linear layer was used as the simplest model for baseline performance benchmarking. More advanced variants of a CNN, namely, pre-trained ResNet-50 (~26M parameters), ResNet-152 (~60M parameters), DenseNet-121 (~8M parameters), DenseNet-201 (~20M parameters), and DenseNet-161 (~29M parameters) models (He, 2015, and Gao, 2018), available from the pyTorch Python library (Paszke, 2019), were fully fine-tuned on the full training data set. Lastly, a pre-trained vision transformer (ViT_10, ~86M parameters) model (Dosovitskiy, 2021), available from the HuggingFace Python library (Wolf, 2020), was both fully fine-tuned, as well as efficiently fine-tuned using Low-Rank (ViT_LoRA, ~12M trainable parameters) adaptation (Hu, 2021).

All these models were selected in order to allow comparison across different model complexities and depths in regard to performance and training times.

3.3 Training Process

Except for the ViT model, all other models were exclusively fully fine-tuned starting from an acquired pre-trained model due to the lower number of model parameters. The ViT model was both fully

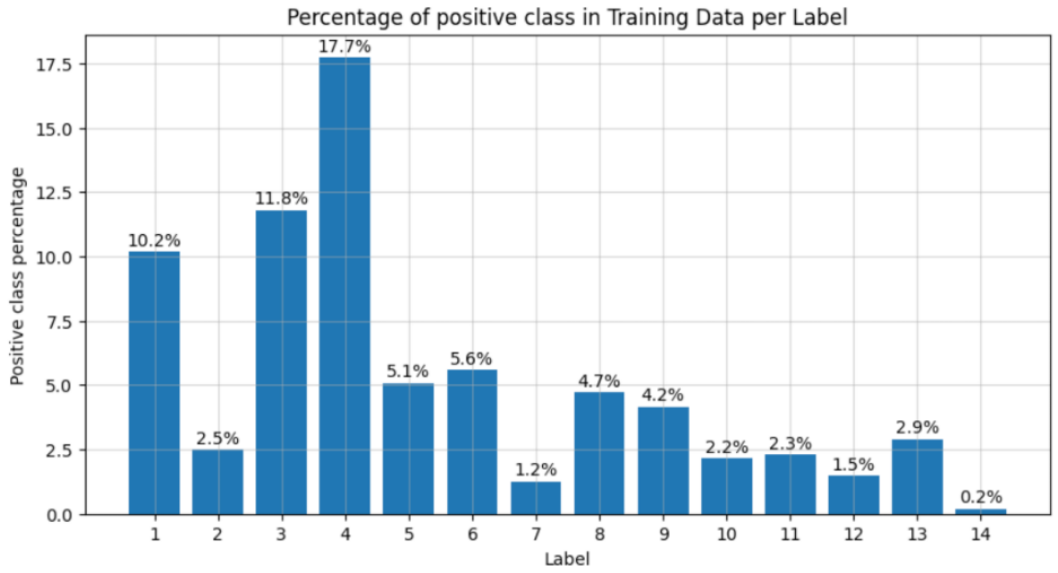

Figure 1: Proportion of positive class for each label in the training data set.

fine-tuned and LoRA fine-tuned. All models were fine-tuned for 10 epochs due to computational cost and time constraints.

AdamW (Loshchilov, 2019) as implemented in pyTorch is used as the optimizer, which explicitly applies parameter regularization (weight decay) and usually leads to improved generalization (compared to traditional Adam optimizer).

As the prediction task is a binary multi-label classification, binary cross-entropy variant was chosen as the objective (loss) function, to properly penalize each label in the output independently. Positive label weight in the objective function was weighted 10 times (also 50 times for ViT, see ViT_50 in table 1) as important compared to negative class in order to induce the models to prioritize positive prediction and reduce effects of slight data imbalance found in the dataset.

Hyperparameters related to the architecture of the models (such as depth and activation functions), these were predetermined and fixed according to the pre-trained models that were chosen, and therefore no tuning of these hyperparameters was performed.

Lastly, a hyperparameter search of batch size and learning rate was performed, and 32 and 0.0001, respectively, were chosen. The search was brief due to limitations on computational resources, and these values were fixed across all trained models due to the high computational cost and the amount of models trained.

3.4 Evaluation Metrics

Accuracy, sensitivity, and AUC are used to evaluate and compare models. Given the nature of the data at hand, it is often more harmful to misclassify a medical diagnosis as false (predict false when pathology is present) than as true (predict true when pathology is not present), as leading a pathology untreated has more potential of causing harm compared to running more tests to confirm a positive initial diagnosis. Therefore, emphasis was placed on evaluating sensitivity (true positive rate). Also, the choice of sensitivity and specificity over precision and recall was made in order to be more aligned with traditional terminology used in the biomedical field (rather than precision and recall which are more often used in computer science).

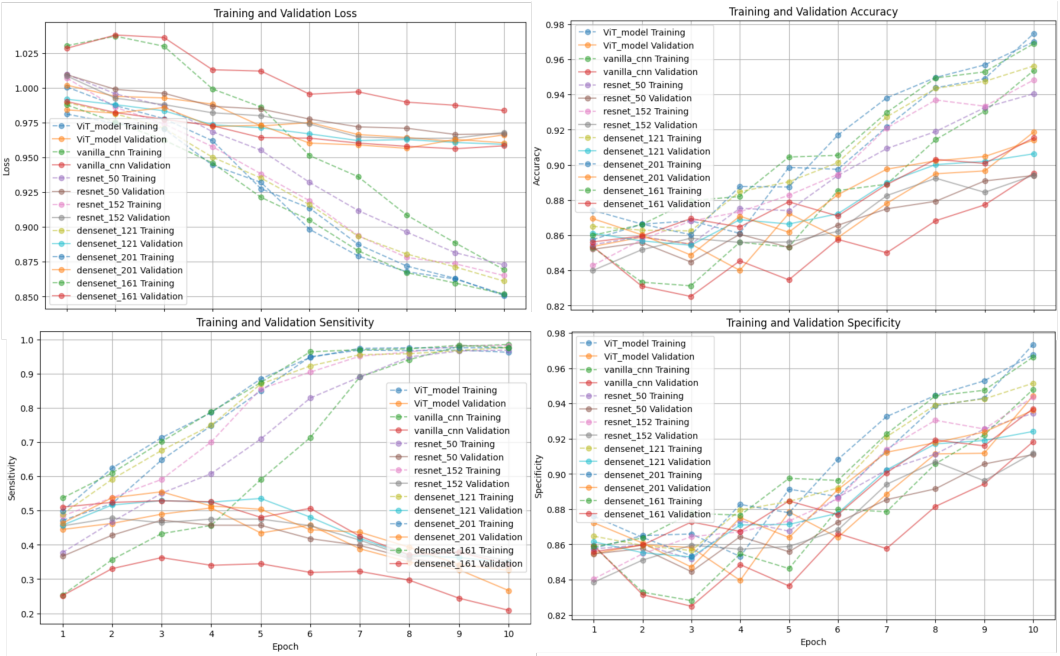

Figure 2: **Training and Validation metrics plots.** Metrics are plotted for models trained with 10-to-1 positive label to negative label modified weight in objective function. Top left: Loss. Top right: Accuracy. Bottom left: Sensitivity. Bottom right: Specificity

4 Results and Analysis

4.1 Data imbalance and modified objective weights

Initial training of all models yielded very high accuracy across all models (above 0.98), however, an examination of the sensitivities and specificities revealed that all specificities were very close to 1.0 and all sensitivities were very close to 0.0, indicating that all models took advantage of the data imbalance and always predicted the negative class for all labels regardless of input image. This is an indication that the models were not really learning during training. To counteract this, a positive weight was added to all positive labels, to penalize the objective more significantly when positive labels are missed, and reward it more significantly when they are predicted correctly.

4.2 Accuracy and Sensitivity vs. Specificity

The modified objective allowed all models to better predict the positive classes as the sensitivities initially increased in early epochs (see Fig. 2). However, as the training progressed sensitivities on validation set for all models degraded while specificities increased (see Fig. 2). However, this was not observed on the training set - sensitivities kept increasing until last epoch, which could mean that all models were overfitting the positive labels in the training set but failing to generalize this to new unseen data.

Lastly, looking at table 1, the trade-off between increased sensitivity at the expense of reduced accuracy can be observed across all models. For example, ViT_50, which was trained with a modified objective in which the positive class had a weight 50 times bigger than negative class, reached the highest sensitivity across all models (0.54 on validation set), but also had the lowest recorded accuracy (0.81 on validation set).

4.3 Model Performance

Overall, all DenseNet variants achieved the lowest validation loss score of 0.96, and the highest validation AUC scores of 0.79 and 0.80 (see table 1). ViT_10 and DenseNet-161 achieved the best

Table 1: Training and validation performance of chosen models

Model	Params #	Training / Validation			Val. AUC	Total train time 10 Epochs [s] (minutes)
		Loss	Accuracy	Sensivitiy		
ViT_10	86M	0.85 / 0.97	0.97 / **0.92**	0.96 / 0.27	0.76	16,706 (278)
ViT_50	86M	1.59 / 1.85	0.84 / 0.81	0.99 / **0.54**	0.78	16,782 (279)
ViT_LoRA	12M	1.06 / 1.08	0.88 / 0.87	0.98 / **0.44**	0.76	18,197 (303)
Vanilla CNN	51M	0.87 / 0.98	0.95 / 0.90	0.99 / 0.21	0.69	3,017 (50)
ResNet-50	26M	0.87 / 0.97	0.94 / 0.89	0.97 / 0.33	0.76	6,136 (102)
ResNet-152	60M	0.87 / 0.97	0.95 / 0.89	0.98 / 0.35	0.77	11,064 (184)
DenseNet-121	8M	0.86 / **0.96**	0.96 / 0.91	0.98 / 0.35	**0.79**	6,873 (115)
DenseNet-201	20M	0.85 / **0.96**	0.97 / 0.91	0.98 / 0.33	**0.79**	9,943 (166)
DenseNet-161	29M	0.85 / **0.96**	0.97 / **0.92**	0.98 / 0.34	**0.80**	12,770 (213)

validation accuracy of 0.92. However, DenseNet-161 achieved a higher validation sensitivity of 0.34 compared to 0.27 of ViT_10, which means that it is more likely to classify true positive cases correctly, and hence has a better performance.

Therefore, the DenseNet architecture is found to be the best performing architecture when considering all the evaluation metrics. Furthermore, the largest DenseNet model contains less than half the parameters in ViT_10 (86M parameters for ViT_10 compared to 29M parameters in DenseNet-161) and ran in roughly 23% less time (total training time of 278 minutes for ViT_10 compared to 213 minutes for DenseNet-161), making it computationally preferred as well, both in terms of training runtime and memory usage.

Lastly, in their dataset benchmarking, Yang et al. 2023 reported highest AUC of 0.778 using Google AutoML Vision model, and the highest accuracy score of 0.948 achieved with both ResNet-50 and Google AutoML Vision, both trained for 100 epochs from scratch. Compared to DenseNet-161 results in our work (AUC of 0.80 and accuracy of 0.92), the performance is comparable despite only fine tuning the pre-trained model for 10 epochs. As Yang et al. postulate, it appears that factors such as hyperparameter tuning and data pre-processing play a more important role in determining model performance for the chosen use case than the model architecture or learning scheme, which is confirmed by our findings as well.

4.4 Model Architecture Complexity

Considering only the models that were trained with the positive classes weighed 10 times heavier than the negative classes (see Fig. 2), it is apparent that the vanilla CNN model performed worse across the loss, accuracy, sensitivity, and specificity on the validation set over almost all epochs. This could indicate that the simpler CNN architecture is not able to learn representations as complex as the more elaborate model architectures, and therefore performs worse on both training data and unseen data. This can also be seen in Fig. 3 for the validation sensitivity and AUC scores per label, where the vanilla CNN model performs significantly worse compared to all other models on all labels.

4.5 Pseudo Multi-Task Learning and Per-Label Performance

Multi-task learning is a regularization method for deep neural network models in which the same model is trained on multiple tasks. The problem formulation in this work, i.e. a 14 class binary multi-label prediction, could be seen as 14 separate binary prediction tasks which our models were trained on all at once. This could potentially help the model learn to identify more general pathology markers (abnormalities) which are common to all pathologies, and therefore provide better generalization performance.

Looking at the per-label validation accuracy, sensitivity, and specificity in Fig. 3, we note that data imbalance did play a role when training with the modified objective function. More balanced classes, such as classes 1, 3, 4 which had the smallest label imbalance rates (10.2%, 11.8%, and 17.7% positive labels, respectively) achieved the highest sensitivity across all models, and in contrast, the

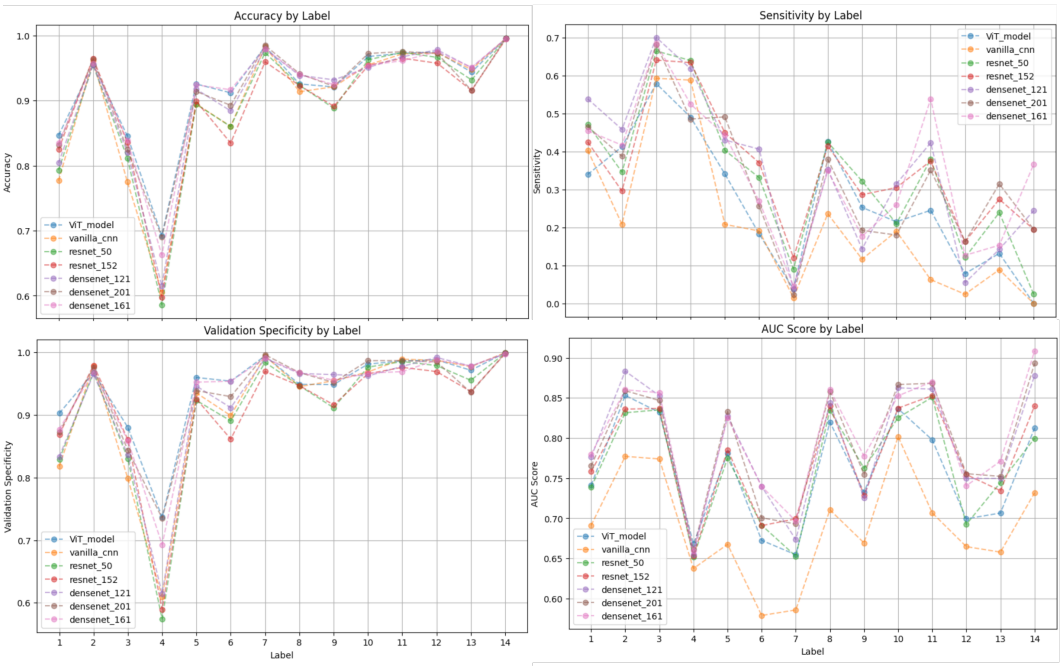

Figure 3: **Validation metrics per label in the validation data set.** Metrics are plotted for models trained with 10-to-1 positive label to negative label modified weight in objective function. Top left: Accuracy. Top right: Sensitivity. Bottom left: Specificity. Bottom right: AUC Score.

lowest specificity and accuracy. It is possible that increasing the number of epochs could mitigate some of this observed trade-off. Classes with more severe imbalance were less affected by the modified objective function and all the models have learned to be more likely to classify these classes as negative (considering the high specificity and accuracy).

The AUC scores per label (Fig. 3) do not exhibit an explicit dependence on the degree of per-label class imbalance.

4.6 ViT: Full Fine-Tuning vs. Efficient LoRA Fine-Tuning

As the ViT model is the biggest model trained in this work, it was initially suggested to perform LoRA fine-tuning to make the process more efficient. A LoRA rank of 8 and a scaling factor of 16 were chosen for the LoRA, which are standard values for LoRA. This configuration brought the number of trainable model parameters in the ViT model from 86M in the base ViT model, to 12M, a reduction of roughly 86% in trainable parameters, which was expected to bring down training time by a similarly significant proportion. Nevertheless, the LoRA-reduced ViT model exhibited slightly longer training times than the full ViT model, to much surprise (278 minutes for full ViT model, and 303 minutes for LoRA model, for 10 training epochs). Furthermore, the LoRA-reduced model exhibited worse validation loss and accuracy (1.08 and 0.87, compared to 0.97 and 0.92 for the full ViT model, see table 1), but significantly higher validation sensitivity (0.44 compared to 0.27 in the base ViT model). This could mean that the learning rate, which was not changed for the LoRA fine-tuning, was too low, and the model was slower to learn, with the loss and sensitivity not decreasing as much as the full fine-tuned model (recall sensitivity degradation as function of training epoch in Fig. 2).

4.7 Comparison with Literature and Human Performance

As briefly discussed in section 4.3, our best performing model (DenseNet-161) achieved similar results to Yang et al. 2023 in terms of prediction accuracy and AUC using a ResNet-50 model and a black-box Google AutoML Vision model. Rajpurkar et al. 2017 propose a 121-layer deep DenseNet model, trained on the same data set as in our work, with comparison to 4 experienced radiologists.

Their model achieved an average AUC score of 0.84 across all 14 labels (compared to AUC of 0.80 for our best model), and outperformed all radiologists with an F1-score of 0.435, compared to the average F1-score across all radiologists of 0.387. Rajpurkar et al. conclude that deep learning models are able to outperform trained radiologists on chest X-ray classification tasks. In more recent work, Cacciamani et. al 2022, explored performance of radiologists again modern computer-aided diagnosis AI systems in classification of prostate MRI images. They concluded that whereas the AI systems reviewed outperformed the trained radiologists in classification sensitivity, when classification was combined based on the AI system and the radiologist, best classification results (sensitivity and specificity) were obtained, concluding that a combination of domain knowledge together with an AI model yield the best results.

5 Conclusion

5.1 Summary of Findings

Deep learning models including deep CNN's and Vision Transformers show great promise in chest X-ray image analysis (and medical image analysis as a whole), but this comes with a unique set of challenges, including trade-off between model complexity and interpretability, performance, and clinical use adoption.

Among the different models of various complexities trained, the DenseNet architecture was found to yield consistently solid performance (across the various depths), moderate training time compared to other models, and achieve comparable performance to benchmarks found in literature.

5.2 Limitations and Future Work

One of the main constraints in this work has been computational resources, especially when compared to the computational costs of model training. Furthermore, the choice of using 224 x 224 resolution images rather than smaller sizes (smallest resolution available directly from data source is 28 x 28) contributed to the long training times, but allowed models to learn from images which contain more information. This has also contributed to a suboptimal hyperparameter tuning effort.

Future work would include performing more robust hyperparameter tuning for each model architecture in order to find the optimal set of parameters required for best training. Likewise, effects of increased training epochs on performance would be an interesting key aspect to study, with the hopes of increasing all of accuracy, sensitivity, and specificity, in order to yield better overall predictive capabilities.

References

[1] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, Bingbing Ni. Yang, Jiancheng, et al. "MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification." Scientific Data, 2023.

[2] Dosovitskiy, Alexey, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." International Conference on Learning Representations, 2021.

[3] He, Kaiming Zhang, Xiangyu Ren, Shaoqing Sun, Jian. (2016). Deep Residual Learning for Image Recognition. 770-778. 10.1109/CVPR.2016.90.

[4] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 2261-2269, doi: 10.1109/CVPR.2017.243.

[5] Wolf, Thomas, et al. (2020). "Transformers: State-of-the-Art Natural Language Processing". 38-45. 10.18653/v1/2020.emnlp-demos.6, 2020.

[6] Paszke, Adam, et al. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". 10.48550/arXiv.1912.01703.

[7] Hu, Edward J., et al. "LoRA: Low-Rank Adaptation of Large Language Models." arXiv preprint arXiv:2106.09685, 2021.

[8] Loshchilov, Ilya, and Frank Hutter. "Decoupled Weight Decay Regularization." arXiv preprint arXiv:1711.05101, 2019.

[9] OpenAI. "ChatGPT: Optimizing Language Models for Dialogue." OpenAI, 2023, https://openai.com/chatgpt.

[10] Rajpurkar, Pranav, et al. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning." arXiv preprint arXiv:1711.05225, 2017.

[11] Cacciamani, Giovanni E., et al. "Is Artificial Intelligence Replacing Our Radiology Stars? Not Yet!" European Urology Open Science, vol. 48, 2023, pp. 14-16. ISSN 2666-1683. https://doi.org/10.1016/j.euros.2022.09.024.

[12] Reyes, Mauricio et al. "On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities." Radiology. Artificial intelligence vol. 2,3 e190043. 27 May. 2020, doi:10.1148/ryai.2020190043

[13] Azad, Reza et al. "Advances in medical image analysis with vision Transformers: A comprehensive review." Medical image analysis vol. 91 (2024): 103000. doi:10.1016/j.media.2023.103000

[14] Shamshad, Fahad et al. "Transformers in medical imaging: A survey." Medical image analysis vol. 88 (2023): 102802. doi:10.1016/j.media.2023.102802

[15] Chan S, et al. "Will machine learning end the viability of radiology as a thriving medical specialty?" Br J Radiol. 2019 Feb;92(1094):20180416. doi: 10.1259/bjr.20180416. Epub 2018 Nov 1. PMID: 30325645; PMCID: PMC6404816.

A Appendix / supplemental material

As part of initial exploration of python libraries and existing resources to be used for this work, ChatGPT-4o, a generative AI model developed by OpenAI, 2023, was used to scope for a starting point for the technical implementation.

The input and corresponding output can be found here:

`https://chatgpt.com/share/675b2136-5180-800b-a010-bb0d37a8ae22`

The output was used to aid initial brainstorming and provide a skeleton for our own implementation.