Reichman University

Efi Arazi School of Computer Science
Machine Learning and Data Science M.Sc. Program

Implementation of Masked Autoencoders Vision Learners on a medical chest X-Ray Images

Date: August 2023

Students:

Itamar Fradkin
Ron Boxer

Mentors:

Dr. Eyal Klang, MD, Head of Big Data and AI Hub, ARC, Sheba Medical Center
Sigal Sina, Ph.D., Data Scientist, Team Leader, ARC, Sheba Medical Center
Alon Oring, Machine Learning and Data Science Program, Reichman University

Table of Contents

1. Abstract

This paper aims to evaluate the potential of Masked Autoencoder Vision Learners (MAE ViT)[1] applied to chest X-ray images with MIMIC-CXR-JPG v2.0.0[2] dataset in the context of image classification. The central goal is to assess if this model can yield significant results for X-ray image classification, especially when compared to established CNN models like ResNet[12]. To test the models, we conducted multiple experiments using different dataset sizes and diverse data augmentation techniques. The code implemented for this study can be found in our Git repository.

We observed that both models performed similarly when trained on a full dataset of approximately 250,000 samples. However, when the dataset size was reduced to 1,000 or 10,000 samples, the MAE ViT model exhibited a marginal but noticeable improvement over ResNet across a range of evaluation metrics, indicating its potential for greater efficacy in limited data availability situations.

The results highlight the lack of a universal deep-learning model for classifying chest X-ray images and emphasize the need to choose a model depending on factors including dataset size, available computing power, and the required level of accuracy.

2. Background

General background

A chest X-ray is the most prevalent type of medical imaging investigation worldwide and one that is frequently used to evaluate the thorax. Chest radiographs are used to diagnose both acute and chronic cardiopulmonary diseases, to confirm the proper positioning of devices including pacemakers, central lines, and tubes, and to aid in associated diagnostic procedures.[2]
The proportion of radiologists in the physician workforce is falling[3] in the United States.
Also, the geographic distribution of radiologists favors larger, more urban counties[4].
In huge healthcare organizations like the U.S. Department of Veterans Affairs[5] and the U.K. National Health Service[6], delays, and backlogs in timely medical imaging interpretation have significantly decreased the quality of care.
In locations with few resources and a severe lack of radiological services, the situation is considerably worse. As of 2015, only 11 radiologists served the 12 million people of Rwanda[7], while the entire country of Liberia, with a population of four million, had only two practicing radiologists[8]. The effectiveness of radiological services can be increased by using accurate automated or computer-assisted radiograph analysis. By using computer vision technologies, radiological services can also potentially improve in underserved areas.

The chief obstacles to the development and clinical implementation of AI algorithms include the availability of sufficiently large, curated, and representative training data that includes expert labeling. Current supervised AI methods require a curation process for data to optimally train, validate, and test algorithms. Most research groups and industries have limited access to data based on small sample sizes from small geographic areas. In addition, the preparation of data labeling is a costly and time-intensive process. This limitation of the process resulted in algorithms with limited utility and poor generalization[9].

Machine Learning background

Self-Supervised Learning (SSL) is a machine learning paradigm in which a model autonomously creates data labels when given unstructured data as input[10]. These labels are then used as ground truth labels in the following rounds. The core concept behind self-supervised learning is to provide supervisory signals on the unlabeled data that is presented to it on the initial iteration in an unsupervised manner.

3. Data

In our research, we used MIMIC-CXR-JPG v2.0.0[2], which is a comprehensive dataset that contains 239,921 chest X-ray images derived from imaging studies. This dataset, accompanied by 14 labels produced from natural language processing tools, is freely available with the aim of boosting research in medical computer vision. To ensure patient privacy, all images in this dataset have been thoroughly de-identified.

The dataset labeling was a key step, we used NegBio[11] open-source labeler tool for this research. NegBio is an open-source rule-based tool detecting negation and uncertainty in radiological reports. Another important part of our preprocessing was to use only images showing the “frontal” view. This helped establish an unambiguous evaluation of the models and minimize the "noise" originating from the images, a choice made after discussions with our expert mentors.

For labeling, we systematically tagged the “finding” and “no finding” labels.
We consider the following labels given as part of the data: Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardio mediastinum, Fracture, Lung Lesion, Lung Opacity, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax, Support Devices. **Images with at least one of these were labeled as a “finding” and assigned a value of 1,** and images **labeled “no finding” in the metadata were assigned a value of 0.** To reconstruct our data you can use this file from our repo.

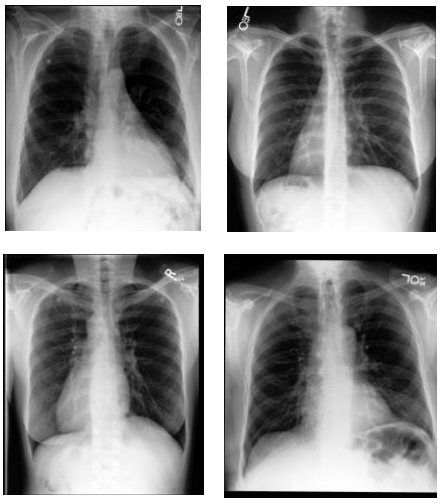
Data labels size and distribution

		Full Data		Sub-sample	
Split	**Label**	**Number Of Images**	**Fraction**	**Number Of Images**	**Fraction**
Train	0	96,694	40.63%	4,126	41.26%
	1	141,268	59.37%	5,875	58.74%
Validation	0	813	41.50%	767	41.87%
	1	1,146	58.50%	1,065	58.13%
Total		**239,921**		**11,833**	

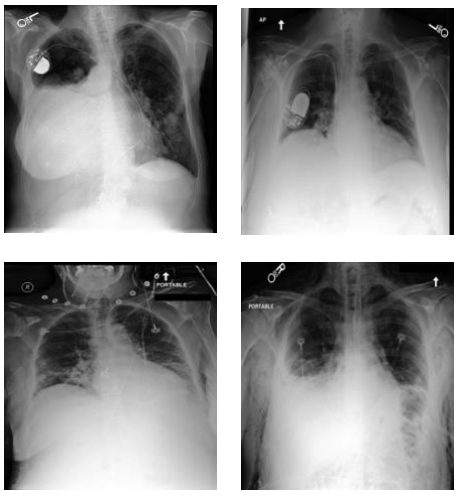
More statistical information on the labels distributions is available in the "generate-stats-from-paper.ipynb" notebook inside the Git repository. Although it doesn't affect our research since we used a binary classification.

We further created a CSV file called "dicoms_with_labels_and_splits.csv". This file was instrumental in our experiment, not only enabling easy image loading during the training phase but also helping us select the 10,000 and 1,000 instance samples for further experiments. As a result of this workflow, we were able to execute our study smoothly and ensure that our research findings were reproducible. To reconstruct this csv you can use this file from our repo.

Examples of the labels "No Finding" vs "Finding":



"No Finding"



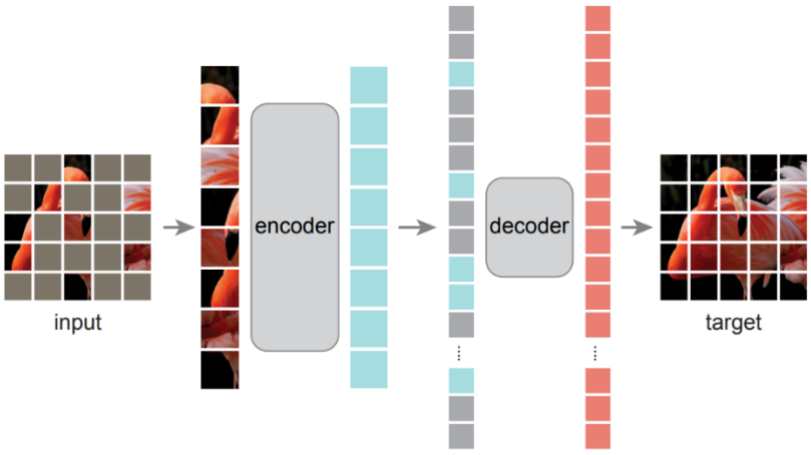
"Finding"

4. Methods

We used ResNet as our benchmark algorithm alongside the MAE ViT as a self-supervised model and compared the model's performance with different augmentations on three binary labeled sets from the MIMIC dataset: full dataset training on "pleural effusion" and "no finding" labels, full dataset training on "finding" and "no finding" labels and sub-sample training on "finding" and "no finding".

The Masked Autoencoder (MAE ViT) model, introduced in the "Masked Autoencoders Are Scalable Vision Learners" paper[1], is a Transformer-based architecture for vision-based machine learning self-supervised tasks. The MAE ViT model consists of an encoder that downscales input to a compact representation and a decoder that reconstructs the original image from this representation.

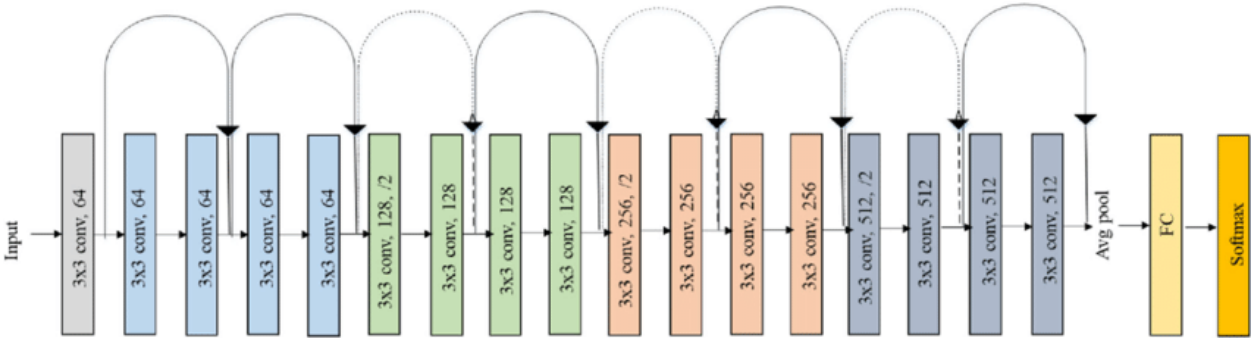
MAE architecture



Crucially, the MAE ViT employs a masking technique during training, hiding certain parts of the input and thereby forcing the model to fill in the missing information based on its understanding of the rest of the image. This process enables the model to learn meaningful representations and generalize better.

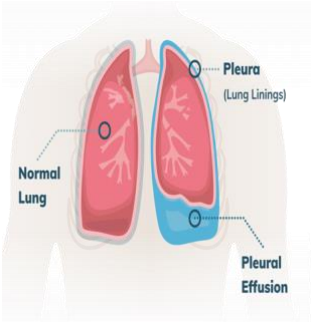
The Residual Network (ResNet) is a deep learning model that has significantly influenced the field of computer vision. Introduced by He et al. in 2015[12], ResNet addresses the problem of vanishing gradients and training difficulty in deep neural networks. The innovative architecture provided by the authors enables the training of networks that are much deeper than was previously feasible. ResNet's performance on the ImageNet dataset, where it won 1st place on the ILSVRC 2015 classification task[13], has established it as a common benchmark in the field.

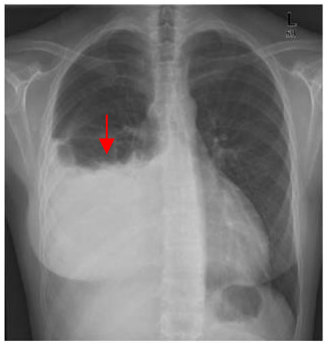
ResNet 18 architecture:



Initial experiment involved training these models on "pleural effusion" and "no finding" labels. We selected these classes due to the relative ease of clinical classification associated with "pleural effusion," especially when contrasted against the "no finding" label. A pleural effusion is a buildup of fluid between the layers of tissue that line the lungs and chest cavity.

Plural Effusion example





Source

Chest X-ray showing right pleural effusion

These experiments exhibited an encouraging decrease in loss during training, validating their ability to learn effectively. However, despite the promising results, we recognized the need for a more clinically relevant scenario - It is more important to understand whether a finding exists in the image than to identify a specific case (pleural effusion). This decision is based on the reasoning that, anyway, a clinical expert (a radiologist) must confirm that a finding has been made. As a result, we reoriented our approach and adopted a binary model, contrasting the

"finding" category with "no finding". where the "finding" label includes all instances with any clinical finding. This experiment used the full dataset. We tested two different training methods: full network training and last-layer training. The first proved to be more effective, but for both models, the performance comparison revealed minimal distinctions.

Following the results, we proceeded to assess the performance of these models on smaller datasets, specifically subsets containing 10,000 and 1,000 samples. We incorporated different data augmentation strategies into these experiments: no augmentation, classic augmentation, and an augmentation technique used in the MAE ViT model. Additionally, we tested various masking percentages (25%, 50%, and 75%) on model performance for the MAE ViT models.

5. Implementation

We implemented our methods using two machine-learning models, which were loaded from Hugging Face[14]. The models used include the Resnet 18[15] model and the MAE ViT base[16] model.

The hyperparameters for all models were set as follows: a learning rate of 1e-06, 20 epochs, a batch size of 32, and with Adamw optimizer. The differences in other hyperparameters were also considered during the implementation.

For data augmentation, we used different approaches for each model. For the Resnet model, we experimented with no augmentation and classic augmentation: Flipping, rotation, and distortion. For the MAE ViT model, we utilized the built-in augmentation, which is based on a well-known and widely used paper[17] in image classification.

In terms of masking, we applied a variety of masking ratios (25%, 50%, and 75%) for the MAE ViT model. However, no masking was used for the Resnet model in this research.

The implementation was carried out on a computer at SHEBA hospital, which runs on Windows and is equipped with an Nvidia A5000 graphics card.

During the implementation, we encountered several challenges. One of these was to fit the images to the models, from 2D to 3D images. Another challenge was the large size of the dataset, which was very spread out across numerous folders. To address this, we conducted the full training using links instead of the original paths and created some helper files as described in the Data section.

We also experimented with different augmentations, as explained above.
Finally, we found nearly identical results across the entire dataset. To probe this, we trained the models on smaller datasets (10k and 1k), examining if reduced data might improve performance.

prompting us to pivot our focus toward the third pillar for further exploration. This shift was guided by a confluence of our initial observations and mentor consultations.

6. Results

In our study, we evaluated the models using the complete dataset labeled with "plural effusion" and "findings". Remarkably, both tests yielded comparable outcomes. These findings align with “Comparing different deep learning architectures for classification of chest radiographs”[18]. A primary insight from it, is that more complex networks don't necessarily outperform simpler ones. This observation, and the results guided us to our third experiment involving a reduced dataset and various augmentations as you can see in The results tabels. Following the methods outlined in the Implementation section, we trained these models using specific hyperparameters. Subsequently, we examined diverse augmentations and masking ratios for the MAE ViT transformer.

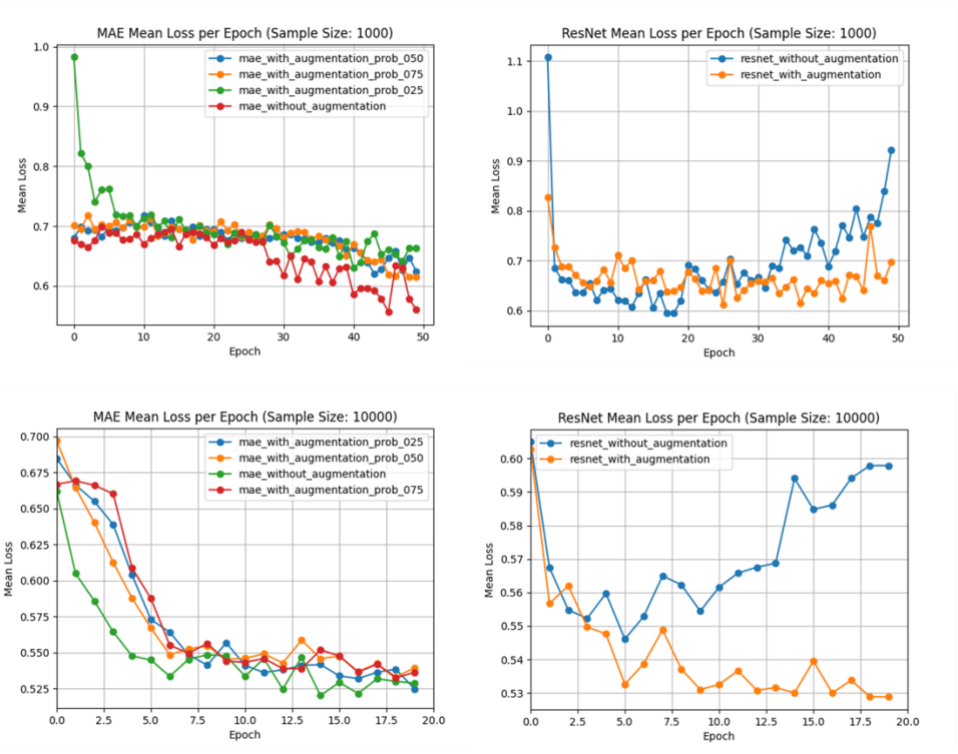
Results Tables

Sample size: 1000							
model	**transformer**	**% masking**	**Accuracy**	**Precision**	**F1**	**Recall**	**AUC**
MAE ViT	augmented	0.25	0.645	0.612	0.717	**0.865**	0.675
MAE ViT	augmented	0.50	0.660	0.670	0.676	0.683	0.715
MAE ViT	augmented	0.75	**0.710**	**0.695**	**0.739**	0.788	0.729
MAE ViT	not augmented		0.700	0.686	0.730	0.779	**0.752**
ResNet	augmented		0.660	0.638	0.709	0.798	0.726
ResNet	not augmented		0.685	0.667	0.722	0.788	0.734

Sample size: 10,000							
model	**transformer**	**% masking**	**Accuracy**	**Precision**	**F1**	**Recall**	**AUC**
MAE ViT	augmented	0.25	0.749	0.738	0.803	0.882	**0.809**
MAE ViT	augmented	0.50	0.747	0.729	0.805	**0.899**	0.796
MAE ViT	augmented	0.75	0.750	0.737	**0.805**	0.888	0.799
MAE ViT	not augmented		**0.755**	**0.775**	0.795	0.817	0.802
ResNet	augmented		0.750	0.755	0.797	0.844	0.807
ResNet	not augmented		0.743	0.759	0.788	0.819	0.780

We can observe that for a sample size of 1000, the MAE ViT model performs better than the ResNet model in terms of accuracy, precision, recall, and AUC, while the F1 score is the same for both models.
As for a sample size of 10,000, the **MAE ViT model performs better across the board**.

Mean Loss plots



7. Discussion and Conclusion

Our experiments revealed several interesting insights into the performance of two deep-learning models for chest X-ray image classification. Firstly, we observed that training on the full dataset (approximately 250K samples) did not significantly differ between the models, indicating that the models were able to learn from a large amount of data and generalize well to new examples. However, when training on smaller datasets (1K and 10K), the MAE ViT model performed better than the ResNet model. Although the improvement was not big in absolute terms it was significant and on several setup configurations. This suggests that the MAE ViT model may be more effective at handling smaller datasets, which is an important consideration in medical imaging, where dataset sizes are often limited by factors such as cost, time, and availability of annotated data.

In addition, our findings align with those of previous studies that have investigated the use of deep-learning models for medical image classification. For example, the authors of "Comparing different deep learning architectures for classification of chest radiographs"[18] found that deeper models do not always yield better results in the healthcare domain. Similarly, the authors of "Delving into Masked Autoencoders for Multi-Label Thorax Disease Classification"[19] demonstrated that pre-trained MAE ViT models can perform comparably, and sometimes better, than state-of-the-art CNNs. Our study adds to these findings by providing further evidence that models like MAE ViT can also achieve better performance only in certain scenarios.
Overall, our results suggest that there is no one-size-fits-all solution for selecting a deep-learning model for chest X-ray image classification. Instead, the choice of model depends on various factors such as dataset size, computational resources, and desired level of accuracy.

In conclusion, we presented a comprehensive evaluation of two deep learning models for X-RAY image classification, including a novel application of MAE ViT to this task. MAE ViT model performing slightly better on smaller datasets.

8. Possible future work

During this research, certain constraints, particularly those related to time, precluded the exploration of larger models such as ViT-Large, ViT-Huge, ResNet 34, ResNet 50, and ResNet 152
We believe that the extension of our work to bigger variants of the models could potentially yield significant insights and contribute to this research.

Another aspect worth researching is experimenting with different X-ray angles. Currently, we have only used the "frontal" view.

9. Acknowledges

We would like to express our sincere thanks to our mentors at the ARC Innovation Hub at Sheba Hospital for their essential assistance and support during this project. Their expertise in the healthcare domain and technical advice provided us with unique insights and perspectives that greatly enhanced our work. Special thanks to Dr. Eyal Klang, MD, Head of Big Data and AI Hub, and Sigal Sina, Ph.D., Data Science Team Leader.
We are also grateful for the computational resources they made available to us, which enabled us to conduct our experiments and simulations efficiently and effectively.

In addition, we would like to thank Alon Oring, our mentor from the Reichman University, for his technical assistance and advice during the project. His contributions were instrumental in helping us overcome various challenges and obstacles, and his input was crucial in ensuring the success of our research.
We would like to thank all our mentors and collaborators for their time, effort, and dedication to our project. Without their support, this work would not have been possible.

10. References

[1] *He, Kaiming, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. "Masked Autoencoders Are Scalable Vision Learners." arXiv.org, November 11, 2021.* https://arxiv.org/abs/2111.06377v3.

[2] *W. Johnson, Alistair E., Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. "MIMIC-CXR-JPG, a Large Publicly Available Database of Labeled Chest Radiographs." arXiv.org, January 21, 2019.* https://arxiv.org/abs/1901.07042v5.

[3] *Rosenkrantz, A. B., Hughes, D. R., & Duszak, R., Jr (2016). The U.S. Radiologist Workforce: An Analysis of Temporal and Geographic Variation by Using Large National Datasets. Radiology, 279(1), 175–184.* https://doi.org/10.1148/radiol.2015150921

[4] *Rosenkrantz, A. B., Wang, W., Hughes, D. R., & Duszak, R., Jr (2018). A County-Level Analysis of the US Radiologist Workforce: Physician Supply and Subspecialty Characteristics. Journal of the American College of Radiology : JACR, 15(4), 601–606.* https://doi.org/10.1016/j.jacr.2017.11.007

[5] *Bastawrous, S., & Carney, B. (2017). Improving Patient Safety: Avoiding Unread Imaging Exams in the National VA Enterprise Electronic Health Record. Journal of digital imaging, 30(3), 309–313.* https://doi.org/10.1007/s10278-016-9937-2

[6] *Rimmer A. (2017). Radiologist shortage leaves patient care at risk, warns royal college. BMJ (Clinical research ed.), 359, j4683.* https://doi.org/10.1136/bmj.j4683

[7] *Rosman, David & Nshizirungu, Jean & Rudakemwa, Emmanuel & Moshi, Crispin & Tuyisenge, Jean & Uwimana, Etienne & Kalisa, Louise. (2015). Imaging in the Land of 1000 Hills: Rwanda Radiology Country Report. Journal of Global Radiology. 1. 10.7191/jgr.2015.1004.* https://www.researchgate.net/publication/277641244_Imaging_in_the_Land_of_1000_Hills_Rwanda_Radiology_Country_Report

[8] *Farah S Ali, Samantha G Harrington, Stephen B Kennedy, and Sarwat Hussain. Diagnostic radiology in Liberia: a country report. Journal of Global Radiology, 1(2):6, 2015* https://doi.org/10.7191/jgr.2015.1020

[9] *Willemink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L. R., Summers, R. M., Rubin, D. L., & Lungren, M. P. (2020). Preparing Medical Imaging Data for Machine Learning. Radiology, 295(1), 4–15.* https://doi.org/10.1148/radiol.2020192224

[10] *What is Self-Supervised Learning*

[11] *Peng Y, Wang X, Lu L, Bagheri M, Summers R, Lu Z. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. AMIA Summits on Translational Science Proceedings. 2018;2017:188.* https://pubmed.ncbi.nlm.nih.gov/29888070/

[12] *He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition." arXiv.org, December 10, 2015. https://arxiv.org/abs/1512.03385v1*

[13] *https://huggingface.co/docs/transformers/main/model_doc/resnet*

[14] *https://huggingface.co/models*

[15] *https://huggingface.co/microsoft/resnet-18*

[16] *https://huggingface.co/facebook/vit-mae-base*

[17] *Cubuk, Ekin D., Barret Zoph, Jonathon Shlens, and Quoc V. Le. "RandAugment: Practical Automated Data Augmentation with a Reduced Search Space." arXiv.org, September 30, 2019. https://arxiv.org/abs/1909.13719v2.*

[18] *Bressem, K.K., Adams, L.C., Erxleben, C. et al. Comparing different deep learning architectures for classification of chest radiographs. Sci Rep 10, 13590 (2020). https://doi.org/10.1038/s41598-020-70479-z*

[19] *Xiao, Junfei, Yutong Bai, Alan Yuille, and Zongwei Zhou. "Delving into Masked Autoencoders for Multi-Label Thorax Disease Classification." arXiv.org, October 23, 2022. https://arxiv.org/abs/2210.12843v1*