A trustworthy framework for medical image analysis with deep learning

Ma, Kai; He, Siyuan; Xi, Pengcheng; Ebadi, Ashkan; Tremblay, Stéphane; Wong, Alexander

Publisher's version / Version de l'éditeur:

Journal of Computational Vision and Imaging Systems, 8, 1, pp. 51-54, 2023-05-10

A Trustworthy Framework for Medical Image Analysis with Deep Learning

Kai Ma — University of Waterloo
Siyuan He — University of Waterloo
Pengcheng Xi — National Research Council Canada
Ashkan Ebadi — National Research Council Canada
Stephane Tremblay — National Research Council Canada
Alexander Wong — University of Waterloo
Email: {k78ma, sy4he, a28wong}@uwaterloo.ca, {pengcheng.xi, ashkan.ebadi, stephane.tremblay}@nrc-cnrc.gc.ca

Abstract

Computer vision and machine learning are playing an increasingly important role in computer-assisted diagnosis; however, the application of deep learning to medical imaging has challenges in data availability and data imbalance, and it is especially important that models for medical imaging are built to be trustworthy. Therefore, we propose TRUDLMIA, a trustworthy deep learning framework for medical image analysis, which adopts a modular design, leverages self-supervised pre-training, and utilizes a novel surrogate loss function. Experimental evaluations indicate that models generated from the framework are both trustworthy and high-performing. It is anticipated that the framework will support researchers and clinicians in advancing the use of deep learning for dealing with public health crises including COVID-19.

1 Introduction

The COVID-19 pandemic continues to affect lives around the world. Medical imaging, including chest X-rays, plays a key role in diagnosis. Using computer vision and deep learning, computer-assisted diagnosis has shown great potential in this field [1, 2]. Moreover, improving on the task of COVID-19 chest X-ray classification may lead to advancements for similar tasks. However, three main issues have been identified in this field.

Limited data. Datasets for medical imaging, especially for novel diseases such as COVID-19, are typically small compared to natural image datasets, making model training more difficult. As such, medical feature learning is commonly conducted through transfer learning [3], with large-scale supervised learning followed by down-stream fine-tuning. This approach, however, is limited by the relevance of the large-scale data to the downstream task, as well as label quality. Consequently, self-supervised learning (SSL) has been proposed and has shown comparable performance to state-of-the-art supervised models [4, 5]. The methodology behind *contrastive* self-supervised learning, which aims to learn feature representations by comparing closely related data samples to each other, is especially relevant, as medical images are extremely similar and can appear identical to untrained eyes.

Class imbalance. Aside from small datasets, another important issue for medical imaging is class imbalance, where there are significantly more negative (benign) data samples than positive (malignant) ones. Thus, models are heavily biased toward the majority negative class and exhibit poor predictive performance for the minority positive class, which is often more important for medical diagnosis. A way to combat this during the training process is to maximize the AUC (area under the ROC curve) instead of minimizing cross-entropy (CE) loss [6]. This is suitable for imbalanced data, as maximizing AUC aims to rank the prediction score of positive samples higher than negative ones. However, AUC maximization is more sensitive to model changes, making it less practical than minimizing CE loss [6].

Low trustworthiness. In the context of medical AI, trustworthiness of predictions is important to both patients and clinicians. An existing problem is that deep neural networks optimized with the standard CE loss function tend to be overly cautious for the minority class, while being overconfident for the majority class [7]. This problem is especially hard to deal with, as model trust quantification is a relatively new and undeveloped area. Existing literature typically focuses on evaluating the trust for a prediction from a single data sample [8, 9]; therefore, these approaches often suffer from various weaknesses, including low interpretability and being limited to Bayesian networks [10]. To deal with these limitations, a concept of "question-answer" trust has been introduced [11], where the trustworthiness of a model is determined by its behaviour when answering questions, such that undeserved confidence is penalized while well-placed confidence is rewarded. Through this method, a simple scalar "trust score" is introduced, such that a higher trust score indicates a more trustworthy model. To remedy the problem of low model trust, we use Deep AUC Maximization to replace traditional CE loss with the robust AUC min-max margin loss [6].

To address the aforementioned issues, we propose **TRUDLMIA**, a **simple and trustworthy deep learning framework** for medical image analysis. Both supervised and self-supervised learning are combined for effective medical image feature learning and a novel surrogate loss function is adopted to build high-performing, high-trust models. The framework adopts a model-agnostic modular design for generalization capabilities.

In summary, our contributions and findings are three-fold:

- We propose a general deep learning framework for medical image analysis which can be used to build high-performing, high-trust models;
- We show that fine-tuned models with self-supervised pre-training surpass supervised ones for COVID-19 classification, including state-of-the-art deep learning models designed specifically for the task;
- AUC maximization with margin loss leads to more effective feature learning and higher trustworthiness, effectively dealing with the problems of class imbalance and prediction under/over-confidence.

2 Literature Review

In computer vision, SSL has gained popularity for learning representations, requiring no labels unlike supervised training. The SSL approaches can be categorized into generative or contrastive/discriminative. Two mainstream contrastive approaches, *momentum contrast* (MoCo) [5] and *SimCLR* [4], learn features by comparing data samples to each other. MoCo approaches this task through a mechanism analogous to dictionary look-up. Through a contrastive loss function, a visual representation encoder is trained by matching encoded queries to a dictionary of encoded keys. SimCLR, however, focuses on the use of data augmentations on the same image. The SimCLR aims to minimize the distance between data augmentations of the same image, while maximizing the distance between different images.

Contrastive learning methods have been shown to be effective in medical contexts. The MoCo model has been trained on a chest X-ray data-set to produce the MoCo-CXR model [12]. Subsequent fine-tuning experiments show that models initialized with MoCo-CXR outperformed non-MoCo-CXR counterparts, especially on limited training data. Experiment on a dataset that is unseen during pre-training also shows that MoCo-CXR pre-training has good transferability across chest X-ray datasets and tasks.

In dealing with the COVID-19, the SSL methods exhibit advantages over supervised counterparts because of limited data with labels. In [13], the authors showed results that self-supervised pre-training using MoCo led to better results than supervised pre-training for screening COVID-19 patients. In [14], the authors pre-train the MoCo model on chest X-rays to learn more general image representations to use for prognosis tasks, differing from previous work in that existing solutions leverage supervised pre-training on non-COVID images, an approach limited by the difference between the pre-training data and the target COVID-19

patient data. It thus achieves comparable prediction accuracy to that of experienced radiologists analysing the same information. SimCLR has also been applied for medical AI. In [15], the authors propose multi-instance contrastive learning, a novel approach that generalizes contrastive learning to leverage special characteristics of medical image analysis. They observe that SSL pre-training on ImageNet, followed by additional pre-training on unlabeled domain-specific medical images, improves classifier accuracy [15].

The quantification of model trustworthiness is a new and undeveloped area, compared to other deep learning performance metrics. Existing work typically focuses on evaluating trustworthiness for a prediction made on a single data sample, either through measuring agreement with a nearest-neighbour classifier [16, 17] or estimations for model uncertainty [9, 10]. Other newer approaches employ complex frameworks such as probabilistic descriptions based on network topologies [18], and even cloud-based heuristics that rely on a large number of models [19]. However, these approaches often suffer from severe weaknesses — their trust quantification is often highly complex, hard to interpret, or limited to certain neural networks like Bayesian ones [10]. To deal with these limitations, a concept of “question-answer" trust has been introduced in [11], where the trustworthiness of a model is determined by its behaviour when answering questions correctly or incorrectly. Consequently, undeserved confidence is appropriately penalized while well-placed confidence is rewarded. Through this method, a simple scalar “trust score" is introduced to express question-answer to practice.

In medical imaging, AUC score has become a common metric to compare deep learning methods, and directly maximizing AUC score is a proven method of improving model performance. Furthermore, proponents of AUC maximization claim that it is optimal for handling imbalanced data, especially for tasks where the positive class is important, since maximizing AUC aims to rank the prediction score of any positive data higher than any negative data [6]. An ongoing area of study is the design of surrogate loss functions for AUC maximization, with the standard naive approach being a simple pairwise surrogate loss based on the definition of the AUC score [20]. However, this suffers from severe scalability issues and sensitivity to noise. To alleviate this, the authors of [6] propose AUC min-max margin loss, a novel surrogate loss function for maximizing AUC score, which uses a squared hinge function (common in margin-based SVM classifiers). Deep AUC Maximization (DAM) with AUC min-max margin loss has shown state-of-the-art results on various difficult tasks with unbalanced data, including medical imaging tasks such as CheXpert [21] and melanoma detection.

3 Methodology

3.1 Framework Design

The proposed TRUDLMIA framework comprises three main modules: i) generic learning through large-scale supervised pre-training using natural images (ImageNet [22]), ii) adapted learning through large-scale self-supervised learning using natural images or domain images without labels, and iii) targeted learning through supervised fine-tuning on downstream tasks using a labeled dataset (see Fig. 1). Conducting supervised pre-training (module i) before self-supervised pre-training (module ii) means less epochs of the latter are required, making the framework much more computationally efficient as SSL can be very costly. The three modules are built upon each other in order to build trustworthy models with optimal performance, despite the fact that training data on target tasks are limited in quantity and imbalanced. For validation, the model achieving the best validation accuracy is saved and evaluated on the unseen test split.

We study two main-stream SSL approaches, namely SimCLR and MoCo, on their performance and trustworthiness as pre-training for our framework. In the comparison below and ablation studies, we mainly use SimCLR, as we found that it performed better than MoCo for the given task (results shown in Section 4.4). We use AUC maximization with AUC min-max margin loss in module iii) of the TRUDLMIA framework, which has several benefits over traditional cross-entropy (CE) loss. Most importantly,

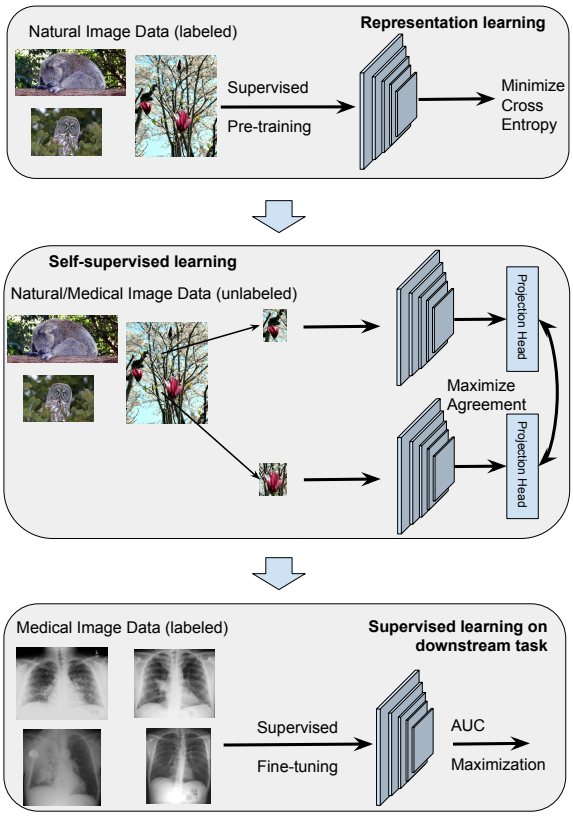

Fig. 1: Overview of the proposed deep learning framework.

AUC maximization is better at handling imbalanced data [6], being more resistant to trust issues caused by CE loss (over-confidence in the majority class and over-cautiousness in the minority class [7]).

The TRUDLMIA framework adopts a plug-in architecture in the computation of image features. In our study, we adopt main-stream deep convolutional neural network (CNN) models, i.e., ResNet [23] and DenseNet [24], which can be replaced with other network architectures. The two self-supervised learning (SSL) approaches compared in module ii) are also replaceable. Likewise, the AUC maximization used in the module iii) can be replaced with alternative loss functions.

3.2 Trust Score Computation

We compute trust scores for models based on a method introduced in [11]. Given a question x, an answer y with respect to a model M, such that $y = M(x)$, and z representing the correct answer to x, we then use $R_{y=z|M}$ to denote the space of all questions where the answer y given by model M matches the correct answer z. Likewise, we use $R_{y\neq z|M}$ to denote the space of all questions where the answer y given by model does not match the correct answer. We also define the confidence of M in an answer y to question x as $C(y \mid x)$. Thus, *question-answer trust* of an answer y given by model M of a question x, with knowledge of the correct answer z, is defined as

$$Q_z(x,y) = \begin{cases} C(y \mid x)^{\alpha}, & \text{if } x \in R_{y=z|M} \\ (1 - C(y \mid x))^{\beta}, & \text{if } x \in R_{y\neq z|M}, \end{cases}$$

with α and β denoting reward and penalty relaxation coefficients.

To integrate the trust score computation into the models trained through our framework, we first calculate an optimal threshold value by maximizing F1-score on a validation split. Data samples are classified using the threshold, and outputs are then normalized such that negative predictions are scaled between 0 and 0.5 and positive predictions are scaled between 0.5 and 1. This allows us to express model confidence, which is then used to compute the *trust score* according to the question-answer method introduced above. We use $\alpha = 1$ and $\beta = 1$, equally rewarding well-placed confidence and undeserved overconfidence. This is done for all of the positive samples in the unseen test split. Finally, an overall positive class

trust score for the model is determined by calculating the mean of the computed individual scores.

4 Experimental Results

4.1 Dataset

Table 1: Data split for COVIDx 8B

Split	Negative	Positive	Total
Train	13,793	2,158	15,951
Test	200	200	400

Various datasets are used in TRUDLMIA modules. In supervised pre-training (module i), the deep CNN models are pre-trained on the ImageNet [22] dataset. In self-supervised learning (module ii), the MoCo model is pre-trained on both ImageNet and MIMIC-CXR dataset [25]. The SimCLR model is pre-trained on the ImageNet dataset. For the downstream task (module iii), the COVIDx dataset (Version 8B) [26], a small dataset with a high class imbalance, is used for fine tuning the models end-to-end. The dataset training/testing split is shown in Table 1). All models built in our experiments are evaluated on the same test subset.

4.2 COVID-CXR-SSL

Our top-performing model, dubbed COVID-CXR-SSL, demonstrates both high performance and trust score (see Table 2). The model uses ResNet in module i) and SimCLR in module ii), followed by fine tuning on the COVIDx dataset in module iii).

We compare the COVID-CXR-SSL with COVID-Net CXR-2 [27] and COVID-Net CXR-3 [28], high-performing models that have been designed for the same dataset. COVID-Net CXR-2 uses machine driven design to automatically discover highly customized macro/micro-architecture designs. COVID-Net CXR-3 employs a self-attention mechanism (MEDUSA). Both models are state-of-the-art, consistently surpassing high-performing models on various medical imaging tasks [28]. Our COVID-CXR-SSL model outperforms both COVID-Net CXR-2 and COVID-Net CXR-3 across all metrics on COVIDx V8B. It is noted that the COVID-Net CXR models use input images at a resolution of 480×480, while our model uses a resolution of only 224×224.

Table 2: Model performance and trust scores for COVID-Net CXR-2, COVID-Net CXR-3 and COVID-CXR-SSL

Model	Precision		Sensitivity		Trust
	Pos.	Neg.	Pos.	Neg.	
COVID-Net CXR-2	0.970	0.955	0.956	0.970	-
COVID-Net CXR-3	0.990	0.975	0.975	0.990	-
COVID-CXR-SSL	**1.000**	**0.980**	**0.980**	**1.000**	**0.964**

4.3 Ablation Study

We conduct an ablation study to investigate the contribution of different modules in the TRUDLMIA framework. We start with fine-tuning a pre-trained ResNet on the COVIDx dataset as a baseline (model "SL"). The pre-trained ResNet model is also used as a backbone architecture for training with the SimCLR architecture followed by fine-tuning it using the CE loss function (model "SL+SSL"). Furthermore, the SimCLR model is also fine-tuned using the AUC maximization loss function (model "SL+SSL+AUC"). Both the "SL+SSL" and "SL+SSL+AUC" models are fine tuned for 200 epochs. Table 3 lists the performance metrics and trust scores computed on the models. We obtain an increase of about 6% on precision and sensitivity metrics and 4% on trust score from the adoption of SSL. AUC maximization further improves the performance while improving trust score slightly.

Table 3: Ablation study on model performance and trust scores for different model architectures

Architecture	Precision		Sensitivity		Trust
	Pos.	Neg.	Pos.	Neg.	
SL	1.000	0.885	0.870	1.000	0.918
SL+SSL	1.000	0.939	0.935	1.000	0.952
SL+SSL+AUC	**1.000**	**0.952**	**0.950**	**1.000**	**0.954**

4.4 Selection of SSL Plug-in

Given the choice of different SSL approaches, we conduct a comparison with MoCo architecture in the TRUDLMIA framework. After fine-tuning for 100 epochs, we select the best models for evaluation and provide their performance metrics in Table 4. The SimCLR based model outperforms the MoCo-based one across all metrics. Our results also indicate that pre-training with natural images on ImageNet is more effective than pre-training on large-scale medical image dataset, MIMIC-CXR, despite the latter being more relevant to the downstream task. [25].

Table 4: Model performance and trust scores for different SSL plug-ins

SSL plugin	Precision		Sensitivity		Trust
	Pos.	Neg.	Pos.	Neg.	
MoCo (*MIMIC-CXR*)	0.995	0.896	0.884	0.995	0.909
MoCo (*ImageNet*)	0.998	0.934	0.930	0.998	0.937
SimCLR (*ImageNet*)	**1.000**	**0.952**	**0.950**	**1.000**	**0.954**

5 Conclusion

In this work, we propose TRUDLMIA, a trustworthy and high-performing modular deep learning framework for medical image analysis, alleviating the prevalent issues of limited data, class imbalance, and low trustworthiness. The framework comprises large-scale supervised and self-supervised learning, as well as fine-tuning on downstream tasks in a supervised fashion. Through a highly successful assessment on the COVIDx dataset, TRUDLMIA framework proves its effectiveness for medical image analysis.

The proposed TRUDLMIA has shown great potential as a deep learning framework for medical image analysis due to its efficacy and simplicity. Models trained through the framework surpass traditional supervised models, including the state-of-the-art COVID-Net CXR-3. Our trustworthy model, COVID-CXR-SSL, will be made publicly available. We hope the TRUDLMIA can contribute to the ongoing fight against the pandemic and establish a viable path for future ones.

Our future work includes exploring the explainability of the models built using the framework, e.g., generating saliency maps with methods such as Grad-CAM [29], in collaboration with radiologists. We are interested in investigating the use of the latest Vision Transformers [30] to replace the CNNs used in the TRUDLMIA framework. Furthermore, the use of Generative Adversarial Networks (GANs) [31] can be explored for augmenting dataset size, as well as its impact on model trust.

6 Potential Societal Impact

Data collection for this study was conducted ethically from public and approved data. While promising, our work is by no means a production-ready solution. Thus, in all of our published files, a disclaimer is included, recommending prospective COVID-19 patients to seek help from professional medical practitioners.

References

[1] S. B. Desai, A. Pareek, and M. P. Lungren, "Deep learning and its role in covid-19 medical imaging," *Intelligence-Based Medicine*, vol. 3-4, p. 100013, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666521220300132

[2] R. Rehouma, M. Buchert, and Y. P. Chen, "Machine learning for medical imaging-based covid-19 detection and diagnosis," *International Journal of Intelligent Systems*, p. 10.1002/int.22504, May 2021, pMC8242401[pmcid]. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8242401/

[3] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (bit): General visual representation learning," 2019. [Online]. Available: https://arxiv.org/abs/1912.11370

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020. [Online]. Available: https://arxiv.org/abs/2002.05709

[5] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," 2019. [Online]. Available: https://arxiv.org/abs/1911.05722

[6] Z. Yuan, Y. Yan, M. Sonka, and T. Yang, "Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification," 2020. [Online]. Available: https://arxiv.org/abs/2012.03173

[7] C. Esposito, G. Landrum, N. Schneider, N. Stiefl, and S. Riniker, "Ghost: Adjusting the decision threshold to handle imbalanced data in machine learning," *Journal of Chemical Information and Modeling*, vol. vol. 61, no. 6, pp. 2623–2640, p. 2623–2640, 2021. [Online]. Available: https://doi.org/10.1021/acs.jcim.1c00160

[8] Y. Geifman, G. Uziel, and R. El-Yaniv, "Boosting uncertainty estimation for deep neural classifiers," *CoRR*, vol. abs/1805.08206, 2018. [Online]. Available: http://arxiv.org/abs/1805.08206

[9] J. S. Titensky, H. Jananthan, and J. Kepner, "Uncertainty propagation in deep neural networks using extended kalman filtering," *CoRR*, vol. abs/1809.06009, 2018. [Online]. Available: http://arxiv.org/abs/1809.06009

[10] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *CoRR*, vol. abs/1703.04977, 2017. [Online]. Available: http://arxiv.org/abs/1703.04977

[11] A. Wong, X. Y. Wang, and A. Hryniowski, "How much can we really trust you? towards simple, interpretable trust quantification metrics for deep neural networks," *CoRR*, vol. abs/2009.05835, 2020. [Online]. Available: https://arxiv.org/abs/2009.05835

[12] H. Sowrirajan, J. Yang, A. Y. Ng, and P. Rajpurkar, "Moco-cxr: Moco pretraining improves representation and transferability of chest x-ray models," 2020. [Online]. Available: https://arxiv.org/abs/2010.05352

[13] S. He, P. Xi, A. Ebadi, S. Tremblay, and A. Wong, "Performance or trust? why not both. deep AUC maximization with self-supervised learning for COVID-19 chest x-ray classifications," *CoRR*, vol. abs/2112.08363, 2021. [Online]. Available: https://arxiv.org/abs/2112.08363

[14] A. Sriram, M. Muckley, K. Sinha, F. Shamout, J. Pineau, K. J. Geras, L. Azour, Y. Aphinyanaphongs, N. Yakubova, and W. Moore, "Covid-19 prognosis via self-supervised representation learning and multi-image prediction," 2021. [Online]. Available: https://arxiv.org/abs/2101.04909

[15] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, V. Natarajan, and M. Norouzi, "Big self-supervised models advance medical image classification," 2021. [Online]. Available: https://arxiv.org/abs/2101.05224

[16] H. Jiang, B. Kim, M. Y. Guan, and M. Gupta, "To trust or not to trust a classifier," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 5546–5557.

[17] M. Xiong, S. Li, W. Feng, A. Deng, J. Zhang, and B. Hooi, "Birds of a feather trust together: Knowing when to trust a classifier via adaptive neighborhood aggregation," *Transactions on Machine Learning Research*, 2022. [Online]. Available: https://openreview.net/forum?id=p5V8P2J61u

[18] M. Cheng, S. Nazarian, and P. Bogdan, "There is hope after all: Quantifying opinion and trustworthiness in neural networks," *Frontiers in Artificial Intelligence*, vol. 3, 2020. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frai.2020.00054

[19] B. Qolomany, I. Mohammed, A. I. Al-Fuqaha, M. Guizani, and J. Qadir, "Trust-based cloud machine learning model selection for industrial iot and smart city services," *CoRR*, vol. abs/2008.05042, 2020. [Online]. Available: https://arxiv.org/abs/2008.05042

[20] W. Gao and Z. Zhou, "On the consistency of AUC optimization," *CoRR*, vol. abs/1208.0645, 2012. [Online]. Available: http://arxiv.org/abs/1208.0645

[21] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," 2019. [Online]. Available: https://arxiv.org/abs/1901.07031

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," pp. 248–255, 2009.

[23] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2016. [Online]. Available: https://arxiv.org/abs/1608.06993

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[25] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, no. 1, p. 317, Dec 2019. [Online]. Available: https://doi.org/10.1038/s41597-019-0322-0

[26] L. Wang, Z. Q. Lin, and A. Wong, "Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *Scientific Reports*, vol. 10, no. 1, p. 19549, Nov 2020. [Online]. Available: https://doi.org/10.1038/s41598-020-76550-z

[27] M. Pavlova, N. Terhljan, A. G. Chung, A. Zhao, S. Surana, H. Aboutalebi, H. Gunraj, A. Sabri, A. Alaref, and A. Wong, "Covid-net cxr-2: An enhanced deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," 2021. [Online]. Available: https://arxiv.org/abs/2105.06640

[28] H. Aboutalebi, M. Pavlova, H. Gunraj, M. J. Shafiee, A. Sabri, A. Alaref, and A. Wong, "Medusa: Multi-scale encoder-decoder self-attention deep neural network architecture for medical image analysis," 2021. [Online]. Available: https://arxiv.org/abs/2110.06063

[29] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization," *CoRR*, vol. abs/1610.02391, 2016. [Online]. Available: http://arxiv.org/abs/1610.02391

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[31] H. Ali and Z. Shah, "Combating covid-19 using generative adversarial networks and artificial intelligence for medical images: A scoping review," *arXiv preprint arXiv:2205.07236*, 2022.