2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-7358-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISBI53787.2023.10230504

DOMAIN GENERALIZATION FOR DIAGNOSIS OF PULMONARY FIBROSIS USING DOSE-INVARIANT FEATURE SELECTION

João B. S. Carvalho[†], *Carlos Cotrini*[†], *Fabian Laumer*[†], *André Euler*[⋆], *Katharina Martini*[⋆], *Thomas Frauenfelder*[⋆], *Joachim M. Buhmann*[†]

[†] Institute for Machine Learning, Department of Computer Science, ETH Zürich
[⋆] Institute of Diagnostic and Interventional Radiology, University Hospital Zürich

ABSTRACT

Automated methods for diagnosing pulmonary fibrosis based on deep learning have achieved promising results in recent times. However, their accuracy depends on the radiation dose used to generate the CT scans and the performance of popular models decreases when evaluated on CT scans with doses different than those of CT scans used for training. We propose a new method for ensuring that the representations computed by these networks are invariant to the dose, without retraining the entire network. Our method improves upon the F1 score of standard methods by 6% to 15% when evaluated on unseen samples recorded with a different radiation dose.

Index Terms— Pulmonary fibrosis, deep learning, domain generalization, feature selection.

1. INTRODUCTION

Computed tomography (CT) scans play an essential role in the diagnosis of idiopathic pulmonary fibrosis, a condition affecting more than 3 million people worldwide [1] and leading to a median survival of 3-5 years if untreated [2]. However, the CT supported diagnosis mostly involves manual procedures by experts and it suffers from high inter-observer variability [3, 4]. As a result, accurate diagnosis of pulmonary fibrosis requires subjecting patients to several procedures, some of them invasive like transbronchial or surgical biopsies, which take time and exhaust health-care resources.

In response to the need for timely and accurate diagnoses, automatic methods for diagnosing pulmonary fibrosis have been proposed. Deep learning based methods, in particular, have been successful in diagnosing fibrosis only from CT scans [5, 6, 7]. The main strategy behind these works is to train a neural network that can be subdivided into a feature extractor network that computes a numerical representation of a CT scan and a classifier network that predicts the degree of fibrosis from this numerical representation. The representation describes the main features of the CT scan. CT scans are computed on a spectrum of X-ray doses, which aim for a balance between low doses for patient safety and high doses for image quality. This raises a problem. Deep-learning models trained on a particular set of radiation doses may have feature extractors that produce representations dependent on the radiation dose, thus cascading to classifiers with a dependence on the CT dose (Fig. 1). As a result, these models may not provide accurate diagnoses on CT scans from different sets of radiation doses, as the representations associated to those scans were not seen by the model during training. This phenomenon can be characterized as an unseen data-shift [8], which entails models being exposed to a shift on the distribution of the data sampling process from training to evaluation - in our case of study, the CT radiation dose.

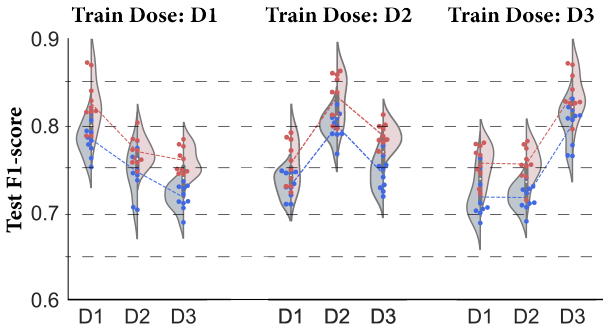

Fig. 1. We plot the F1-score performance for models trained in different doses. Models lose performance when training dose differs from the test dose (x-axis). Doses increase from $D1$ to $D3$. (Red:ViT, Blue: ResNet50)

The variable responsible for the distribution shift can often be tied to the target label through a spurious correlation. In this case, the model infers this correlation, resulting in a more significant loss of performance after training. This performance deterioration commonly occurs in the medical domain due to biased sampling of pathologies from specific hospitals [9], leading to a consistent failure when models are employed in a different clinical setting.

Contribution We propose a novel methodology for selecting deep-learning models that overcomes dose related sample distribution shifts and spurious correlations without relying on model retraining or computational intensive train-

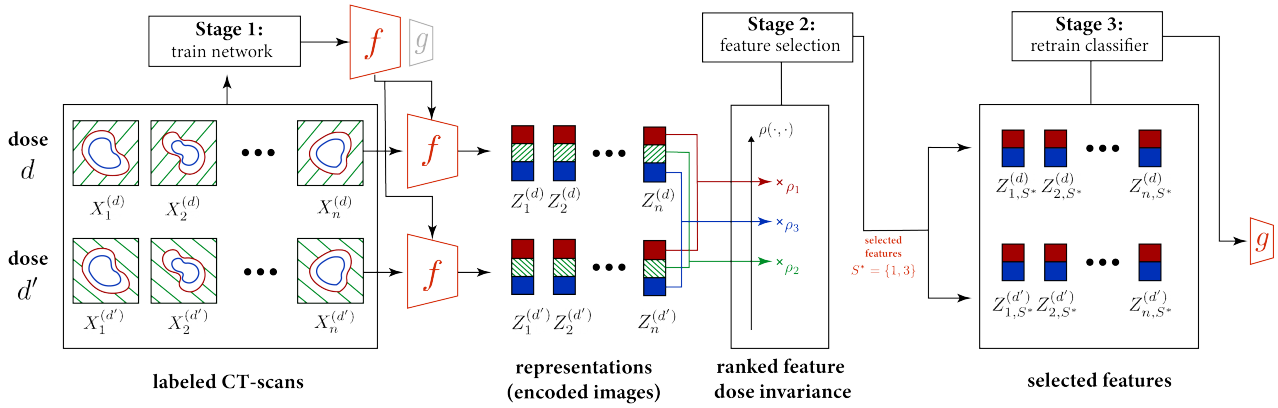

Fig. 2. Model selection pipeline. In this example the green feature differs from dose d to dose d'. The pipeline is as follows: (1) We train a neural network on CT scan slices generated with two different doses. Using the trained network's feature extractor, f, we extract representations from all the CT-scan slices. (2) We rank and select the top features based on their dose invariance using a similarity function ρ. (3) We retrain the network's classifier, g, using only the selected features.

ing schemes. The key method to achieve this robustness is to select only those features from the representations of CT scans that are *invariant* to the dose. The resulting representations of CT scans recorded with different doses must resemble samples originating from the same distribution. As a result, representations are dose-independent and the classifier is no longer substantially influenced by the dose used for the CT scan. We extensively validate our contributions on a retrospective study of 230 patients. We demonstrate that our method generalizes over unseen doses even in the presence of spurious correlations between the CT dose and the type of fibrosis. In all our settings, we beat the state of the art by at least 6% in the F1 score when the training and test dose differ, demonstrating that our method is capable of diagnosing fibrosis on CT scans generated from unseen radiation doses.

2. METHOD

We now explain our pipeline to select models that produce dose-invariant representations. It consists of three stages and is illustrated in Figure 2: ***(1)*** We train a neural network (Section 2.1) on two samples of CT scan produced by two different radiation doses. The network is the composition of a feature extractor and a classifier. ***(2)*** We perform feature selection (Section 2.2). For each dose we generate a set of representations by extracting the features of the CT scans of all patients using the feature extractor network. We then select a subset of features from these representations so that the sets of representations coming from different doses appear as samples originating from the same distribution. ***(3)*** Finally, we retrain a classifier using only the selected features. The final network is the composition of the feature extractor and the new classifier.

2.1. Data and network training

Given a set of manually labeled CT scan slices $\mathcal{D}^{(d)} = \left\{(X_1^{(d)}, Y_1), (X_2^{(d)}, Y_2) \ldots, (X_n^{(d)}, Y_n)\right\} \subseteq \mathbb{R}^N \times \mathcal{Y}$, where $d \in \mathbb{R}^+$ denotes the X-ray dose used to produce each CT and $\mathcal{Y}$ are the labels assigned to a CT scan. For each patient i, a set of three CT scans was generated through a dual-energy CT scanner following [10]. We let $X^{(d)} = \left\{X_i^{(d)} \mid i \leq n\right\}$ with n being the number of patients. A set containing two doses is used for training with the other used for evaluation. The label of a CT scan slice, $Y_i^{(d)} \in \mathcal{Y}$, follows the staging system of [11] with $\mathcal{Y} = \{0, 1, 2\}$, where 0 corresponds to absence of fibrosis patterns, 1 to mild fibrosis characterized by CT scans with observable reticular patterns, and 2 by advanced fibrosis characterized by honeycombing of the lung tissue.

For a set Γ of doses consisting of two out of the total three doses, and a set $\{\mathcal{D}^{(d)} : d \in \Gamma\}$ of labeled CT scans, we train a neural network on the union of the two sets of samples coming from different doses, $\bigcup_{d\in\Gamma} \mathcal{D}^{(d)}$. We denote such a network as a pair (f, g). Here, $f : \mathbb{R}^N \to \mathbb{R}^m$ is a feature extractor that from a CT scan slice $X_i^{(d)}$, for some dose d and $i \leq n$, computes a representation $Z_i^{(d)}$. The function $g : \mathbb{R}^m \to \mathcal{Y}$ is a classifier that diagnoses fibrosis from a representation. For a sample $X^{(d)}$ of CT scans, we define the *representation sample* $Z^{(d)} := \{f(X) \mid X \in X^{(d)}\}$ as the set of representations computed by applying the feature extractor to each CT scan slice with dose d.

2.2. Feature selection

For $S \subseteq \{1, \ldots, m\}$ and $i \leq n$, let $Z_{i,S}^{(d)}$ be the projection of $Z_i^{(d)}$ onto the features in S. We now select a subset $S \subseteq \{1, \ldots, m\}$ of features so that $Z_S^{(d)} := \left\{Z_{i,S}^{(d)}\right\}_{i\leq n}$

and $Z_S^{(d')} := \left\{ Z_{i,S}^{(d')} \right\}_{i \leq n}$, for $d \neq d'$, look "similar". Similarity means comparable high posteriors, i.e., the samples should look like they have been drawn from the same distribution. This procedure ensures that the representations $Z_S^{(d)}$ and $Z_S^{(d')}$ are invariant to the dose.

To measure the similarity between $Z_S^{(d)}$ and $Z_S^{(d')}$, we need to choose a metric $\rho(Z_S^{(d)}, Z_S^{(d')})$. Our goal is then to compute the subset S that maximizes this metric. As m is large, computing this metric for all possible values for S is intractable. We instead follow a greedy approach and compute $\rho_j := \rho(Z_{\{j\}}^{(d)}, Z_{\{j\}}^{(d')})$, for each $j \leq m$. Then we fix a threshold $k \leq m$ and select the k features with the largest ρ_j.

2.3. Similarity function

We consider the following candidates for ρ.

Pearson's sample correlation coefficient (PC). This is a standard measure of correlation between two sets Z, Z' using the means of Z and Z', respectively, $\bar{Z}$ and $\bar{Z}'$:

$$\frac{\sum_{i \leq n} \left(Z_i - \bar{Z}\right)\left(Z_i' - \bar{Z}'\right)}{\sqrt{\sum_{i \leq n} \left(Z_i - \bar{Z}\right)^2 \sum_{i \leq n} \left(Z_i' - \bar{Z}'\right)^2}},$$

Kullback-Leibler divergence (KLd). This is a metric to evaluate the divergence between two distributions p and p' and is defined as $\mathbb{E}\left[\log p(\mathbf{z}) - \log p'(\mathbf{z})\right]$, where the expectation is taken with respect to a random variable $\mathbf{z}$ whose pdf is p. In our case, p and p' are unknown and we only have two samples, Z and Z', from them. Hence, we approximate p and p' from Z and Z', respectively, using kernel density with a standard Gaussian kernel[12]. The kernel's bandwidth is $h = \min\left(\sigma, \frac{IQR}{1.34}\right)$, where σ is the empirical standard deviation of the sample, and IQR is the interquartile range.

Kolmogorov-Smirnov test (KS) [13] This is a non-parametric test to decide if two samples Z and Z' come from a same distribution. It measures $\sup_{z \in \mathbb{R}} |F_Z(z) - F_{Z'}(z)|$, with $F_Z, F_{Z'}$ as the empirical cdf of Z and Z', respectively.

Empirical posterior agreement kernel (ePA) [14, 15]: This is defined as $\mathbb{E}\left[p'(\mathbf{z})\right]$ and measures the overlap between p and p'. As before, p and p' are unknown and we only have the samples Z and Z'. We can approximate the posterior agreement with the following estimate $\frac{1}{n}\sum_{i \leq n} p'(Z_i \mid Z')$ where $p'(\cdot \mid Z')$ is a distribution fitted to Z' through kernel density estimation as before. The resulting estimate, which we denote as ePA, becomes $\frac{1}{n^2 h}\sum_{i,j} \Phi\left(h^{-1}(Z_i - Z_j')\right)$, where Φ is the standard normal pdf.

Random feature selection We also include the random sampling of features as a comparison metric.

2.4. Ranking and selecting features

For any of these candidates ρ, $j \leq m$, and two representation samples $Z^{(d)}$ and $Z^{(d')}$, with d, d' two different doses, we define its *dose invariance* ρ_j as

$$\rho_j = \frac{1}{|\mathcal{Y}|}\sum_{y \in \mathcal{Y}} \rho\left(\begin{array}{l} \{Z_{i,j}^{(d)} \mid i \leq n, Y_i^{(d)} = y\}, \\ \{Z_{i,j}^{(d')} \mid i \leq n, Y_i^{(d)} = y\} \end{array}\right), \quad (1)$$

where $Z_{i,j}^{(d)}$ is the j-th entry of $Z_i^{(d)}$. Note that we evaluate ρ conditioned on the label. One could, alternatively, condition on the patients[1] or use no conditioning at all. However, these alternatives yielded worse results in our experiments.

We now rank the features according to their dose invariance and let S^* be the set of $k \leq m$ features with highest dose invariance. Finally, we train a new classifier $g_{S^*} : \mathbb{R}^{|S^*|} \to \{0, 1, 2\}$ using only those features in S^*. As training set we use $\left\{\left(Z_{1,S^*}^{(d)}, Y_1^{(d)}\right), \ldots, \left((Z_{n,S^*}^{(d)}, Y_n^{(d)}\right)\right\} \cup \left\{\left((Z_{1,S^*}^{(d')}, Y_1^{(d')}\right), \ldots, \left(Z_{n,S^*}^{(d')}, Y_n^{(d')}\right)\right\}$. The resulting neural network is the composition of f and g_{S^*}. We find S^* by greedy search and automatically choose k based on performance in the validation set. By choosing features with the highest dose invariance we aim at making representations invariant to the dose, thus resulting in a network that makes predictions without significant influence from the CT scan dose.

3. EXPERIMENTS

	Exp 1		Exp 2		Exp 3	
$p(Y\|D)$	$D = d$	$D = d'$	$D = d$	$D = d'$	$D = d$	$D = d'$
$Y = 0$	$1/3$	$1/3$	$1/6$	$1/2$	0	$2/3$
$Y = 1$	$1/3$	$1/3$	$1/3$	$1/3$	$1/3$	$1/3$
$Y = 2$	$1/3$	$1/3$	$1/2$	$1/6$	$2/3$	0

Table 1. Different label distributions over the experiments.

To evaluate the ability of our method to generalize over unseen doses not present in Γ we performed a series of three experiments, all following a 10-fold cross-evaluation setup. To avoid intra-sample bias the partitioning of training, validation, and test set was performed at a patient level. The test set has balanced labels. Our baseline is the network without any feature selection, which we denote with "bas.". In our results we present the out-of-distribution evaluation of the models, and an in-distribution baseline, i.e., with test doses being equal to the train dose. The first experiment follows a uniformly distributed unseen data shift, i.e., with two doses used for training and a uniform distribution over labels. The second and third experiments result from an induced spurious correlation between the CT radiation dose d, and the label variable Y, with dose changes being correlated with an increase/decrease of pulmonary fibrosis severeness. This is achieved through a biased sampling of $(Y_1^{(d)}, \ldots, Y_n^{(d)})$ and $(Y_1^{(d')}, \ldots, Y_n^{(d')})$ as described in Table 1.

[1]Note that, as each patient CT scan yields multiple slices, p and p' are possible to estimate.

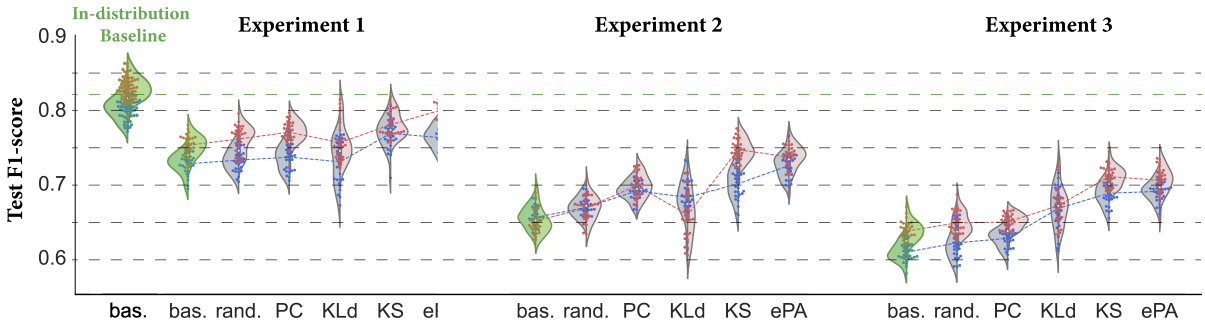

Fig. 3. For each experiment we plot the out-of-distribution F1-score attained by our method paired with each possible similarity function. In green, we present the models without any feature selection as a baseline. Red: ViT. Blue: ResNet50.

Dataset details. The in-house dataset used for this work resulted from a retrospective study of 230 patients from the University Hospital of Zurich with systemic scleroderma and different levels of changes to the lung connective tissue. Each patient was scanned through the dual-split mode of a dual-energy CT scan (SOMATOM Force), resulting in a total of 690 images split into three CT doses. The desired radiation dose (100%) was distributed so that the first X-ray source contributed 1/3 and the second x-ray source 2/3 to the final dose. From each CT scan, a set of between 6 and 12 axial slices was extracted around a primary region of interest identified by expert radiologists, with the number of slices dependent on the extent of the potential lesion. The final image size was 512×512 voxels. Each image was clipped to the $[1\%, 99\%]$ intensity spectrum and normalized to the range of $[0, 1]$.

Implementation details. For the feature extractor network, we chose the commonly used state of the art neural networks ResNet50[16] and Vision Transformer (ViT)[17], taking all layers up to the first fully connected layer. The classifier was a one hidden layer multilayer perceptron. All models were trained with a batch size of 64, using Adam with weight decay set to 10^{-3} and a learning rate schedule starting from 10^{-3} and stopping at 10^{-5}. Model selection followed early stopping. The Cohen's kappa was used for monitoring the performance in the validation set due to the unbalanced label distribution that some experiments had. The source code is available at `https://github.com/JoaoCarv/invariant-dose-fselect`.

3.1. Results and discussion

Our method improves domain generalization. Observe in Figure 3 that the F1 score, when using the KS or the ePA functions, improve by at least 6% with respect to the state of the art. In Experiment 3, where we have the strongest spurious correlation, we even improved the F1 score by 15%. These results, combined with those from Experiment 2, show that our method is even robust against spurious correlations. Additionally, we report only a marginal decrease in performance ($< 2\%$ F1-score averaged across all similarity functions) when evaluated in the in-distribution setting.

Our proposed similarity functions improve upon the state of the art [18]. The results from Experiments 2 and 3 show that by using ePA and KS over the standard PC, we get an improvement in the F1-score of around 5%. In addition, in Experiment 1, by using ePA, we were in some cases close to the ideal in-distribution baseline. As a minor remark, note that the popular KLd attained the largest standard deviation and negligible improvements comparing to PC, confirming its unstability as a similarity metric.

4. RELATED WORK

In spite of reaching a performance comparable to human experts, current methods for deep-learning based diagnosis of idiopathic pulmonary fibrosis [5, 7, 6] rely on particular doses to make a correct diagnosis. As a result, they fail to diagnose fibrosis in CT-scans that use a different dose [19].

Domain generalization via feature selection has been investigated before, but the only candidate studied is the Pearson's correlation coefficient [18]. Our work demonstrates that there are alternatives for ρ like the Kolmogorov-Smirnov test and posterior agreement which attain a better performance than Pearson's coefficient when selecting features.

5. CONCLUSION

This work demonstrates how to construct neural networks that overcome the inability of previous deep learning methods to generalize over CT-scans taken with different doses. By selecting features from the intermediate representations computed by these networks we show that our models generalize better to different doses when diagnosing idiopathic pulmonary fibrosis. We evaluated several similarity functions for measuring dose invariance in these features and achieved an increase between 6% and 15% in the F1 score. Finally, we also see our method being extended beyond pulmonary fibrosis to other medical settings, and more generally, to other distribution shifts besides radiation dose.

6. ACKNOWLEDGMENTS

The project was partially supported by the Foundation for Pneumoconiosis Research, Switzerland.

7. COMPLIANCE WITH ETHICAL STANDARDS

This retrospective study was performed in line with the principles of the Declaration of Helsinki. It received institutional review board and local ethics committee approval. All patients provided written informed consent.

8. REFERENCES

[1] Fernando J Martinez et al., "Idiopathic pulmonary fibrosis," *Nature reviews Disease primers*, vol. 3, no. 1, pp. 1–19, 2017.

[2] Ganesh Raghu et al., "An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management," *American journal of respiratory and critical care medicine*, vol. 183, no. 6, pp. 788–824, 2011.

[3] Simon LF Walsh et al., "Interobserver agreement for the ATS/ERS/JRS/ALAT criteria for a UIP pattern on CT," *Thorax*, vol. 71, no. 1, pp. 45–51, 2016.

[4] Junya Tominaga et al., "Diagnostic certainty of idiopathic pulmonary fibrosis/usual interstitial pneumonia: the effect of the integrated clinico-radiological assessment," *European Journal of Radiology*, vol. 84, no. 12, pp. 2640–2645, 2015.

[5] Simon LF Walsh, Lucio Calandriello, Mario Silva, and Nicola Sverzellati, "Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study," *The Lancet Respiratory Medicine*, vol. 6, no. 11, pp. 837–845, 2018.

[6] Andreas Christe et al., "Computer-aided diagnosis of pulmonary fibrosis using deep learning and CT images," *Investigative radiology*, vol. 54, no. 10, pp. 627, 2019.

[7] Anju Yadav et al., "FVC-NET: An automated diagnosis of pulmonary fibrosis progression prediction using honeycombing and deep learning," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[8] Olivia Wiles et al., "A fine-grained analysis of robustness to distribution shifts," in *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

[9] Alexander D'Amour et al., "Underspecification presents challenges for credibility in modern machine learning," *Journal of Machine Learning Research*, 2020.

[10] Davide Bellini et al., "Dual-source single-energy multidetector CT used to obtain multiple radiation exposure levels within the same patient: phantom development and clinical validation," *Radiology*, vol. 283, no. 2, pp. 526, 2017.

[11] Nicole SL Goh et al., "Interstitial lung disease in systemic sclerosis: a simple staging system," *American journal of respiratory and critical care medicine*, vol. 177, no. 11, pp. 1248–1254, 2008.

[12] Bernard W Silverman, *Density estimation for statistics and data analysis*, Routledge, 2018.

[13] Gregory W Corder and Dale I Foreman, *Nonparametric statistics: A step-by-step approach*, John Wiley & Sons, 2014.

[14] Joachim M Buhmann, "Context sensitive information: Model validation by information theory," in *Mexican Conference on Pattern Recognition*. Springer, 2011, pp. 12–21.

[15] Morteza Haghir Chehreghani, Alberto Giovanni Busetto, and Joachim M Buhmann, "Information theoretic model validation for spectral clustering," in *Artificial Intelligence and Statistics*. PMLR, 2012, pp. 495–503.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[18] Vikas Garg, Adam Tauman Kalai, Katrina Ligett, and Steven Wu, "Learn to expect the unexpected: Probably approximately correct domain generalization," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 3574–3582.

[19] Rongping Zeng et al., "Performance of a deep learning-based CT image denoising method: Generalizability over dose, reconstruction kernel, and slice thickness," *Medical Physics*, vol. 49, no. 2, pp. 836–853, 2022.