Using Distributionally Robust Optimization to improve robustness in cancer pathology

Surya Narayanan Hari **Jackson Nyman** **Nicita Mehta** **Haitham Elmarakeby**

Bowen Jiang **Felix Dietlein** **Jacob Rosenthal** **Eshna Sengupta** **Renato Umeton**

Eliezer M. Van Allen*
Dana Farber Cancer Institute
Harvard Medical School
Boston, MA 02115
eliezerm_vanallen@dfci.harvard.edu

Abstract

Computer vision (CV) approaches applied to digital pathology have informed biological discovery and clinical decision-making. However, batch effects in images represent a major challenge to effective analysis. A CV model trained using Empirical Risk Minimization (ERM) risks learning batch-effects when they may align with the labels and serve as spurious correlates. The standard methods to circumvent learning such confounders include (i) application of image augmentation techniques and (ii) examination of the learning process by evaluating through external validation (e.g., unseen data coming from a comparable dataset collected at another hospital). The latter approach is data-hungry and the former, risks occluding biological signal. Here, we suggest two solutions from the Distributionally Robust Optimization (DRO) families. Our contributions are i) a DRO algorithm using abstention which is a slight variation over existing abstention-based DRO algorithms and ii) a group-DRO method where groups are defined as hospitals from which data are collected. We find that the model trained using abstention-based DRO outperforms a model trained using ERM by 9.9% F1 in identifying tumor vs. normal tissue in lung adenocarcinoma (LUAD) at the expense of coverage. Further, by examining the areas abstained by the model with a pathologist, we find that the model trained using a DRO method is more robust to heterogeneity and artifacts in the tissue. Together, we propose selecting models that are more robust to spurious features for translational discovery and clinical decision support. [1]

1 Introduction

Computer vision (CV) approaches applied to cancer histopathology image data have demonstrated potential for biological discovery, precision diagnostics, and as predictive biomarkers [1–5]. Previous work has shown that models trained on one hospital and tested on another show varying levels of performance [6]. This outcome could potentially result from the model learning spurious correlates in the data, such as batch effects, which are artifacts introduced as a result of the Whole Slide Image (WSI) preparation process, and induce a signal that is readily learnable, but not biologically relevant.

Mitigating batch effects parametrically incurs challenges, as they may arise from different parts of the tissue pre-processing pipeline [7]. Further, large models are likely to learn spurious correlates

[1] Extended Abstract, link to full version here

Workshop on Distribution Shifts, 35th Conference on Neural Information Processing Systems (NeurIPS 2021).

when trained to near-zero training error [8], resulting in poor test performance in sub-populations of the data, especially those that are under-represented in the training set [9].

This problem is especially exacerbated in the histopathology domain owing to the giga-pixel nature of the images, which warrant the image to be processed in smaller patches whist attributing the label given to the whole gigapixel image to all its patches - a phenomenon called weak labels [10]. A model risks learning spurious correlates on patches that don't present the phenotype given to the giga-pixel image it was taken from. Here, we propose circumventing this problem using DRO methods [9, 11, 12] in a histopathology task with clinical relevance.

2 Experimental Setup

2.1 Network Architecture

2.1.1 ERM

We use a pretrained ResNet-50 convolutional neural network (CNN) [13] pre-trained on the ImageNet dataset [14] to train an image classification model, herefore referred to as 'the ERM model'. We used a cross-entropy loss function where the loss is computed and aggregated over all samples of a batch.

2.1.2 group-DRO

In our implementation of a group-DRO method, we defined the groups as hospitals from which the WSIs were taken. We trained an algorithm by backpropagating the loss over the tiles from the worst performing hospital, measured by cumulative loss in a batch. However, the reported statistics, such as F1, are reported over the whole validation / testing dataset, and not the worst performing hospital.

2.1.3 Model trained using Abstention

```
Input: abstention threshold $p$, forward function $f$, optimizer $g$, loss function $\mathcal{L}$
Output: $\theta$, the parameters of the model
Initialize $\theta$;
for $i \leftarrow 1$ to $n$ do
    $\tilde{y} = f_{\theta_i}(x)$;
    $\tilde{y}' = \{\tilde{y}_i \| \tilde{y}_i < p \vee \tilde{y}_i > 1 - p\}$;
    $l = \mathcal{L}(\tilde{y}, y)$;
    $\theta_{i+1} \leftarrow g(\theta_i, l)$;
end
```

Algorithm 1: Proposed DRO abstention algorithm

Models were trained using an abstention algorithm (Algorithm 1) whereby we only backpropagated the losses on images for which the predicted softmax scores were greater a threshold p for the predicted class. We first normalized the softmax scores using temperature scaling [15]. We ablated p and measured the coverage (defined by [12] as the number of samples not abstained on) as well as the F1 performance of the model on the samples not abstained on.

2.2 Lung Adenocarcinoma (LUAD)

We evaluated a DRO method on the task of detecting tumor tissue in LUAD WSIs from the Cancer Genome Atlas (TCGA) ($n = 522$). We trained a binary classifier using slide-level labels to classify tiles into tumor or normal tissue.

LUAD is one of the two major histologic subtypes of Non-Small Cell Lung Cancers (NSCLC). Identification of the tumor in a WSI can help guide pathologic assessment and guide treatment decisions [3, 16, 17]. However, identification of tumor may be confounded by scarring tissue from the effects of smoking on lung tissue, amongst other features.

In one set of experiments, we compared ERM against DRO methods trained on data from one hospital and validated on an external validation set consisting of unseen data coming from a comparable dataset collected at another hospital. We also compared the use of image augmentation via recoloring

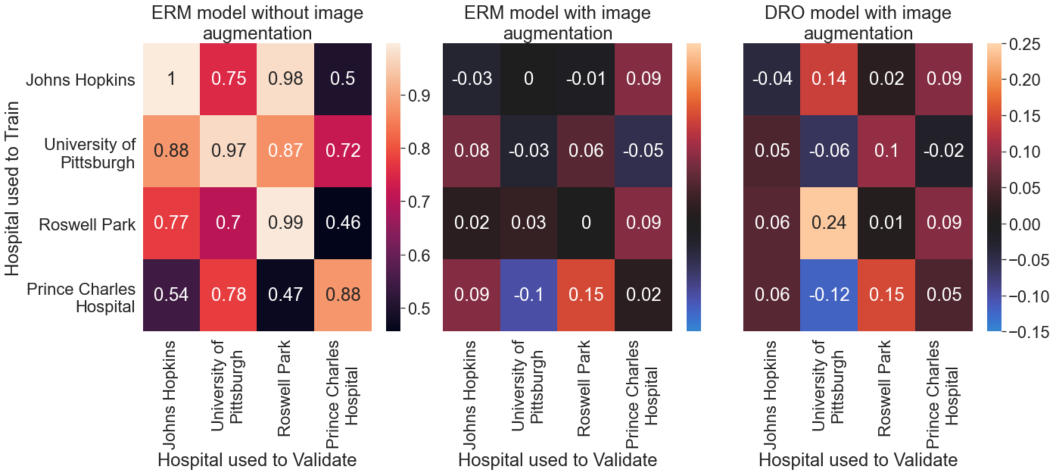

Figure 1: High Amount of heterogeneity in performance depending on which hospital's data are used to train (left). DRO with image augmentation (right) shows improvement over ERM with image augmentation (middle). Middle and right grids shown are differences over grid on the left.

to help reduce the batch effects. In another set of experiments, we combined data from multiple hospitals in our training set and ablated the number of hospitals contributing the external validation dataset to measure the robustness of the ERM and DRO methods.

3 Results

i	ERM	Abstention threshold				gDRO
		0.6	0.7	0.8	0.9	
1	91.1	93.1	96.6	98.4	97.2	92.6
2	88.3	88.6	89.1	93.3	95.2	85.5
3	78.6	77.5	78.9	78.6	82.0	78.7
4	81.1	78.9	80.4	84.0	91.0	82.6
5	72.2	73.4	71.9	76.3	79.7	72.8

i	Abstention threshold			
	0.6	0.7	0.8	0.9
1	.96	.93	.85	.62
2	.95	.87	.79	.54
3	.95	.87	.84	.65
4	.94	.85	.74	.55
5	.96	.90	.84	.60

Table 1: Measuring the performance of ERM, abstention-based DRO and group-DRO (gDRO) models while validating on data combined from from i hospitals; showing macro - F1 (left) and coverage of the abstention models (right).

First, we evaluated models trained on a single source site and validated on either the same or different single source site on a task of LUAD identification. Overall, we found significant heterogeneity in model performance based on the hospital whose data were used to train and validate the model (Figure 1). We also found that image augmentation produced only up to 0.15 improvement in F1, and using our abstention model in addition to image augmentation produced up to 0.24 improvement in F1.

When trained on data from multiple hospitals, we found that the DRO model outperformed a conventional convolutional neural network trained using ERM for the task of detecting tissue with LUAD under all numbers of hospitals held out, at the expense of coverage. (Table 1)

We subsequently conducted statistical testing on performances of models that were thresholded post-hoc at the same thresholds that the models trained with abstention were trained with. We found no statistically significant advantage to using models trained with abstention. However, as we will discuss below, similarly performing models can learn drastically different features, particularly spurious ones in the case of ERM models.

Upon investigation of the tiles that the DRO model with abstention abstained on, we noted that a DRO model abstained on tiles that an ERM model predicted incorrectly (Figure 2, top). DRO models

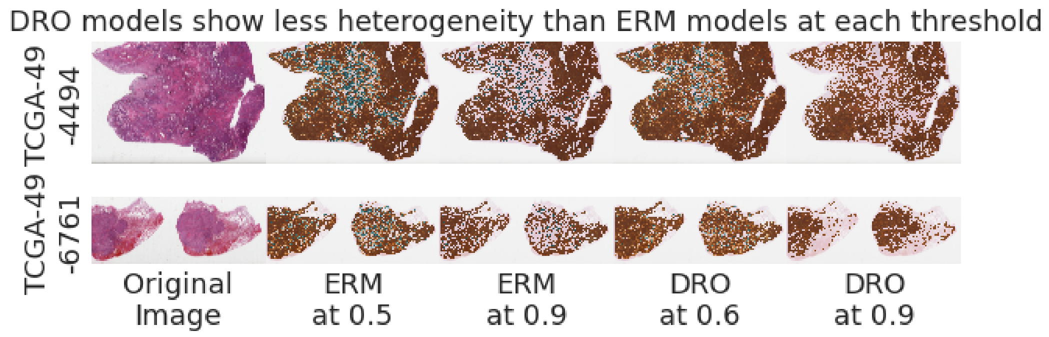

Figure 2: Showing methods thresholded at various confidences. First row: DRO models with higher confidence thresholds abstain on tiles that an ERM model predicts as normal tissue (blue) in a slide labeled as tumorous, thus avoiding learning contradictory features. Second row: ERM methods call non-tumor region on the right hand side of the tissue as tumor, even at high confidence thresholds, whereas DRO methods abstain on these tiles where the tissue does not bear tumor.

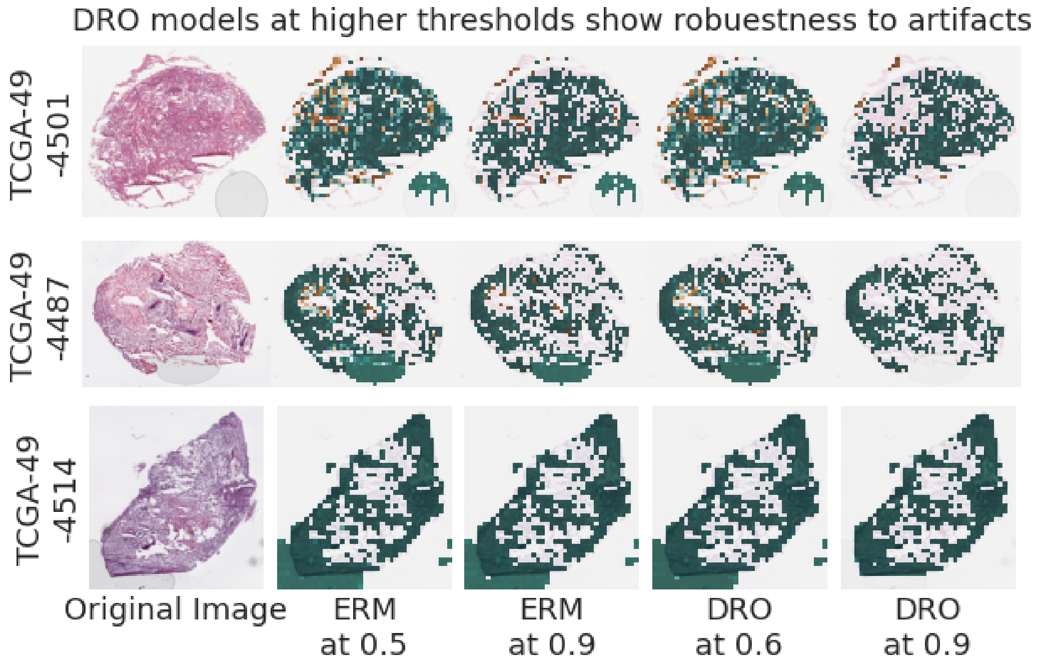

Figure 3: Showing methods thresholded at various confidences. ERM methods predict bubble artifacts as healthy surrounding tissue. DRO methods at higher confidence thresholds abstain from making predictions on artifacts.

could have presented better performance by abstaining from learning potentially conflicting features, since these tiles might present contradictory features to their ground truth label. Further, the ERM methods incorrectly predicted regions of a WSI as tumorous even at higher confidences, that were confirmed by a pathologist as non-tumorous (Figure 2, bottom)). On the other hand the abstention method abstained on these tiles. This could be because these tiles presented spuriously correlated features of that were exclusively correlated with the label of tumor.

Also, ERM models predicted tiles covered by slide-preparation artifacts that leaked through the QC pipeline such as air bubbles (Figure 3) as healthy surrounding tissue in three different slides, implying that these air bubbles might have introduced a spurious correlate. DRO models trained at high confidence thresholds abstained from making predictions on these regions of the WSI. By its abstention from artifacts, the abstention method is less likely to learn spurious correlates.

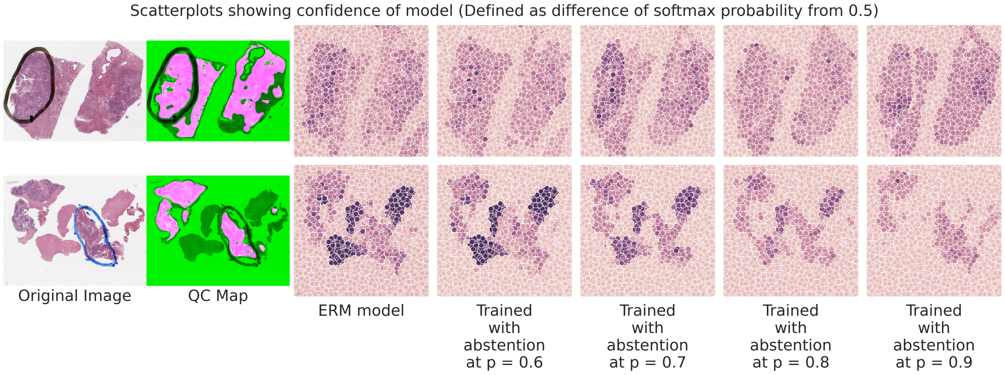
Scatterplots showing confidence of model (Defined as difference of softmax probability from 0.5)
Original Image
QC Map
ERM model
Trained with abstention at p = 0.6
Trained with abstention at p = 0.7
Trained with abstention at p = 0.8
Trained with abstention at p = 0.9

Figure 4: Comparison of Model Confidence Scores (prediction value distance from 0.5) for ERM and DRO models. QC Map: included (pink) vs excluded (green) slide area from HistoQC standard pipelines; Heatmaps display the average confidence of regions determined by watershed segmentation (greater confidence indicated by darker color).

However, the above results might have been produced at the expense of coverage. Even at full coverage, abstention-trained models show utility in the content of the signal used by each model, in two examples in a task of distinguishing between the subtypes of Lung cancer (Figure 4). Models trained with abstention, avoid making predictions on pen marks, while ERM allocates high confidence (darker color) to these artifacts (top row). In a second example taken from a brain biopsy of a metastatic lung cancer, we observed i) the ERM model placed importance on surrounding brain tissue which was confirmed by a pathologist to not bear any tumor, and thus has learned spurious signal; and ii) the model trained with stringent abstention in contrast completely disregards the brain tissue, while placing modest confidence in the verified lung tumor tissue. We would like to highlight that no coverage was lost to thresholding in visualizing these examples and all tiles of Whole Slide Image (WSI) are shown. However, owing to the different training processes, the models learned different features.

To further demonstrate the differences in features learned in each model type, we used each model to separately produce "pruned" datasets: datasets consisting of tiles with maximum softmax value above a certain threshold. We subsequently used these datasets to train a further set of ERM models to distinguish lung subtypes. At higher confidence levels (0.8 and 0.9), models trained on DRO-pruned data offered better performance than those trained on ERM-pruned data (0.61 ± 0.12 vs 0.42 ± 0.11 F1 $[p = 0.10]$ at threshold 0.8; 0.81 ± 0.12 vs 0.53 ± 0.12 F1 $[p = 0.047]$ at threshold 0.9).

4 Conclusion

Here, we evaluated the impact of batch effects and developed approaches to mitigate these fundamental challenges to digital pathology. We suggested potential causes of heterogeneity in model performance that can impact downstream analyses and proposed models that are robust to the distributional shifts between training and held-out test sets. Prospectively, consideration of batch effects in CV histopathology analysis will guide successful biological investigations.

5 Discussion

Ultimately, we found that DRO methods that aim to either optimize the model's performance on a previously defined subgroup or a learned subgroup, defined in our case by the training samples that the model performed well on, were able to provide better performances on an external validation set. The latter approach is aligned with potential clinical support use cases, whereby a model can be allowed to abstain if it is not at least $p\%$ confident that the data are not sampled from the same distribution it has been trained on. We make the assumption that examples that a model predicts with low confidence are OOD. However, this assumption needs further validation studies to confirm.

References

[1] Y. Fu, A. W. Jung, R. V. Torne, S. Gonzalez, H. Vöhringer, A. Shmatko, L. R. Yates, M. Jimenez-Linan, L. Moore, and M. Gerstung, "Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis," *Nature Cancer*, vol. 1, no. 8, pp. 800–810, Aug. 2020, number: 8 Publisher: Nature Publishing Group. [Online]. Available: http://www.nature.com/articles/s43018-020-0085-8

[2] M. Y. Lu, M. Zhao, M. Shady, J. Lipkova, T. Y. Chen, D. F. K. Williamson, and F. Mahmood, "Deep Learning-based Computational Pathology Predicts Origins for Cancers of Unknown Primary," *Nature*, vol. 594, no. 7861, pp. 106–110, Jun. 2021, arXiv: 2006.13932. [Online]. Available: http://arxiv.org/abs/2006.13932

[3] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, and A. Tsirigos, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature Medicine*, vol. 24, no. 10, pp. 1559–1567, Oct. 2018. [Online]. Available: http://www.nature.com/articles/s41591-018-0177-5

[4] W. Bulten, M. Balkenhol, J.-J. A. Belinga, A. Brilhante, A. Çakır, L. Egevad, M. Eklund, X. Farré, K. Geronatsiou, V. Molinié, G. Pereira, P. Roy, G. Saile, P. Salles, E. Schaafsma, J. Tschui, A.-M. Vos, ISUP Pathology Imagebase Expert Panel, B. Delahunt, H. Samaratunga, D. J. Grignon, A. J. Evans, D. M. Berney, C.-C. Pan, G. Kristiansen, J. G. Kench, J. Oxley, K. R. M. Leite, J. K. McKenney, P. A. Humphrey, S. W. Fine, T. Tsuzuki, M. Varma, M. Zhou, E. Comperat, D. G. Bostwick, K. A. Iczkowski, C. Magi-Galluzzi, J. R. Srigley, H. Takahashi, T. van der Kwast, H. van Boven, R. Vink, J. van der Laak, C. Hulsbergen-van der Kaa, and G. Litjens, "Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists," *Modern Pathology*, vol. 34, no. 3, pp. 660–671, Mar. 2021. [Online]. Available: https://www.nature.com/articles/s41379-020-0640-y

[5] J. A. Diao, J. K. Wang, W. F. Chui, V. Mountain, S. C. Gullapally, R. Srinivasan, R. N. Mitchell, B. Glass, S. Hoffman, S. K. Rao, C. Maheshwari, A. Lahiri, A. Prakash, R. McLoughlin, J. K. Kerner, M. B. Resnick, M. C. Montalto, A. Khosla, I. N. Wapinski, A. H. Beck, H. L. Elliott, and A. Taylor-Weiner, "Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes," *Nature Communications*, vol. 12, no. 1, p. 1613, Dec. 2021. [Online]. Available: http://www.nature.com/articles/s41467-021-21896-9

[6] E. Wu, K. Wu, R. Daneshjou, D. Ouyang, D. E. Ho, and J. Zou, "How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals," *Nature Medicine*, vol. 27, no. 4, pp. 582–584, Apr. 2021. [Online]. Available: http://www.nature.com/articles/s41591-021-01312-x

[7] M. Salvi, U. R. Acharya, F. Molinari, and K. M. Meiburger, "The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis," *Computers in Biology and Medicine*, vol. 128, p. 104129, Jan. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0010482520304601

[8] S. Sagawa, A. Raghunathan, P. W. Koh, and P. Liang, "An Investigation of Why Overparameterization Exacerbates Spurious Correlations," *arXiv:2005.04345 [cs, stat]*, Aug. 2020, arXiv: 2005.04345. [Online]. Available: http://arxiv.org/abs/2005.04345

[9] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization," *arXiv:1911.08731 [cs, stat]*, Apr. 2020, arXiv: 1911.08731. [Online]. Available: http://arxiv.org/abs/1911.08731

[10] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine*, vol. 25, no. 8, pp. 1301–1309, Aug. 2019. [Online]. Available: http://www.nature.com/articles/s41591-019-0508-1

[11] A. Kamath, R. Jia, and P. Liang, "Selective Question Answering under Domain Shift," *arXiv:2006.09462 [cs]*, Jun. 2020, arXiv: 2006.09462. [Online]. Available: http://arxiv.org/abs/2006.09462

[12] E. Jones, S. Sagawa, P. W. Koh, A. Kumar, and P. Liang, “Selective Classification Can Magnify Disparities Across Groups,” *arXiv:2010.14134 [cs, stat]*, Apr. 2021, arXiv: 2010.14134. [Online]. Available: http://arxiv.org/abs/2010.14134

[13] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *arXiv:1512.03385 [cs]*, Dec. 2015, arXiv: 1512.03385. [Online]. Available: http://arxiv.org/abs/1512.03385

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” in *CVPR09*, 2009.

[15] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On Calibration of Modern Neural Networks,” *arXiv:1706.04599 [cs]*, Aug. 2017, arXiv: 1706.04599. [Online]. Available: http://arxiv.org/abs/1706.04599

[16] X. Luo, X. Zang, L. Yang, J. Huang, F. Liang, J. Rodriguez-Canales, I. I. Wistuba, A. Gazdar, Y. Xie, and G. Xiao, “Comprehensive Computational Pathological Image Analysis Predicts Lung Cancer Prognosis,” *Journal of Thoracic Oncology*, vol. 12, no. 3, pp. 501–509, Mar. 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1556086416312369

[17] K.-H. Yu, C. Zhang, G. J. Berry, R. B. Altman, C. Ré, D. L. Rubin, and M. Snyder, “Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features,” *Nature Communications*, vol. 7, no. 1, p. 12474, Nov. 2016. [Online]. Available: http://www.nature.com/articles/ncomms12474

[18] J. Brownlee, “A Gentle Introduction to Early Stopping to Avoid Overtraining Neural Networks,” Dec. 2018. [Online]. Available: https://machinelearningmastery.com/early-stopping-to-avoid-overtraining-neural-network-models/

[19] T. Pytorch, “Illustration of transforms — Torchvision master documentation.” [Online]. Available: https://pytorch.org/vision/master/auto_examples/plot_transforms.html

A Appendix

A.1 Measurement details of task training on one hospital and validating on another

In one set of experiments done on TCGA-LUAD, we trained the model on data taken from one hospital and validated it on data taken from another, to mimic a real-world setting where data is private and cannot be shared between institutions. In a resource scarce setting, a model trained on one hospital, cannot be re-trained on data from another. In order to study the effect of the preprocessing steps employed by a singular hospital, we were limited in our analysis to data from hospitals that have both tumor samples and surrounding normal tissue. In this set of experiments, we only performed three cross validation trials, owing to resource limitations.

A.2 Training details

We train our models to minimize error and stop training if the error does not improve on the validation set over five consecutive measurements [18]. The validation performance was measured four times per epoch. We used image augmentation via jittering the RGB pixel values in the RGB space to prevent overfitting to the color distribution by inducing random changes in the brightness, saturation, and other properties of an image, also known as color jitter [19]. We used a random-crop size of 224 pixels within the 512 pixel patch during our training process as a method to prevent overfitting. We performed 5-fold cross validation on all of our experiments. However, each fold of the cross-validation was not forced to be non-overlapping, owing to data availablility constraints.

Color Jitter We used image augmentation via jittering the RGB pixel values in the RGB space to prevent overfitting to the color distribution by inducing random changes in the brightness, saturation, and other properties of an image, also known as color jitter [19]. To discretize the color jitter, we defined a light version of the color jitter that allowed the brightness to be chosen uniformly at random between [0.875, 1.125], the contrast to be chosen uniformly at random between [0.5, 1.5], the saturation to be chosen uniformly at random between [0.5, 1.5] and the hue to be chosen between [-0.1, 0.1]. We similarly defined a heavy version of the color jitter to be four times proportionally higher. We allowed the brightness to be chosen uniformly at random between [0.5, 1.5], the contrast to be chosen uniformly at random between [0, 3], the saturation to be chosen uniformly at random between [0, 3] and the hue to be chosen between [-0.4, 0.4]. The limit on the color jitter we could introduce was placed by the hue, which was forced to be between [-0.5, 0.5]. We chose the heavy color jitter to be less than the maximum allowed by the hue, to preserve some of the color signal presented by the tile.