Examining Batch Effect in Histopathology as a Distributionally Robust Optimization Problem

Surya Narayanan Hari, Jackson Nyman, Nicita Mehta, Bowen Jiang, Jacob Rosenthal,
Felix Dietlein, Eshna Sengupta, Renato Umeton, Eliezer M. Van Allen
Dana Farber Cancer Institute
Boston, MA
http://vanallenlab.dana-farber.org/

Abstract

Computer vision (CV) approaches applied to digital pathology have informed biological discovery and development of tools to help inform clinical decision-making. However, batch effects in the images represent a major challenge to effective analysis and interpretation of these data. 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). Here, we show that the source site of a histopathology slide can be learned from the image using CV algorithms in spite of image augmentation, and we explore these source site predictions using interpretability tools. A CV model trained using Empirical Risk Minimization (ERM) risks learning this signal as a spurious correlate in the weak-label regime, which we abate by using a Distributionally Robust Optimization (DRO) method with abstention. We find that the model trained using DRO outperforms a model trained using ERM by 9.9, 13 and 15% in identifying tumor versus normal tissue in Lung Adenocarcinoma, Gleason score in Prostate Adenocarcinoma, and tumor tissue grade in clear cell Renal Cell Carcinoma. Further, by examining the areas abstained by the model, we find that the model trained using a DRO method is more robust to heterogeneity and artifacts in the tissue. We believe that a DRO method trained with abstention may offer novel insights into relevant areas of the tissue contributing to a particular phenotype. Together, we suggest using data augmentation methods that help mitigate a digital pathology model's reliance on spurious visual features, as well as selecting models that are more robust to spurious features for translational discovery and clinical decision support.

1. Introduction

1.1. Heterogeneity in model outcomes arising from batch effects

Computer vision (CV) approaches applied to cancer histopathology image data have demonstrated emerging potential for biological discovery, precision diagnostics, and as predictive biomarkers [1] [2] [3] [4] [5]. Previous work has shown that models trained on one hospital and tested on another show varying levels of performance; which can even further extend to variance in performance among underserved subpopulations of each hospital [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. Such batch effects and disparity exist due to differences in slide preparation and underlying differences in patient populations served at different hospitals, among other factors. These batch effects may interfere with more biologically relevant prediction tasks by inducing spurious signal.

Further, there exists heterogeneity in the signal within different patches of a single slide as well. Existing CV models applied to histopathology data require significant computational capacity to process a single slide, and existing solutions involve sampling patches from a slide or constructing a graph from the whole slide image (WSI) [7] [8]. To circumvent the problem of choosing patches or assigning a graph structure to the WSI, we propose using an abstention method that abstains on patches of the slide irrelevant to the task to resolve the model's uncertainty in the presence of heterogeneity. We propose

a group Distributionally Robust Optimization (group-DRO) method with abstention as a solution to heterogeneous datasets due to its ability to abstain when features of the image are out of the distribution of features learned as pertinent to the label.

1.2. Using Distributionally Robust Optimization (DRO) to mitigate effect of spurious correlates

Spurious correlates in data impede efforts to deliver translational care and clinical support through artificial intelligence. Machine learning applied to data from publicly available cohorts, such as the Cancer Genome Atlas (TCGA), can learn spurious correlates while trying to analyze large amounts of digitized pathology data paired with molecular and clinical outcomes, impeding multi-hospital analyses from pan-cancer patient cohorts.

Mitigating all forms of batch effects parametrically incurs challenges since batch effects may arise from different parts of the tissue pre-processing pipeline [9]. Being able to predict artifacts of the scan, such as scanner manufacturer and acquisition protocol [10], slide preparation date, source site from which the scan was taken, [10][11] and image quality [12] can induce spurious correlates. Models are likely to learn these spurious correlates when trained to near-zero training error in the weak-label regime [13].

In the CV domain, large models that overfit to spurious correlates result in poor test performance in sub-populations of the data, especially those that are under-represented in the training set [14]. Here, we use using a group-DRO method [14] with abstention [15] across three CV histopathology tasks with clinical relevance.

1.3. Identifying tumor in Lung Adenocarcinoma

Lung Adenocarcinoma (LUAD) is one of the two major histologic subtypes of Non-Small Cell Lung Cancers (NSCLC). Its histology is identified by tumors growing from gland-like structures. Identification of the tumor in a WSI can help guide pathologic assessment, as well as potentially determine the efficacy of therapy [3] [16] [17]. However, identification of tumor may be confounded by scarring tissue from the effects of smoking on lung tissue, amongst other features.

1.4. Predicting grade in Kidney Cancer

In patients with clear cell Renal Cell Carcinoma (ccRCC), amongst pathological features classified based on cell shape and arrangement, nuclear size, nuclear irregularity and nucleolar prominence showed highest effectiveness in predicting distant metastasis, even more so than tumor size [18] and are used to grade the tumor, with a higher grade implying worse prognosis. These morphological features can be distinguished visually and offer potential for the application of CV algorithms. However, due to inter-observer variability and intra-tumoral heterogeneity, CV algorithms are susceptible to batch effects and confounding by spurious correlates.

1.5. Predicting Gleason score in Prostate Adenocarcinoma (PRAD)

Similar to ccRCC, a grading system is used to describe the patterns observed in tumor tissue in Prostate Adenocarcinoma (PRAD), ranging from 1 to 5. A Gleason score for the sample biopsy is then calculated by adding the two most prominent grades visible in the tissue. In practice, the lowest Gleason score awarded is a 6. Recent works have shown the use of CV to predict the Gleason score of a scan of biopsy tissue [19] [20]. However, whether or not Gleason scoring models are learning spurious correlates of the Gleason grade is incompletely characterized but critical for clinical use.

2. Experimental Setup

2.1. Network Architecture

2.1.1 ERM

We use a pretrained ResNet-50 convolutional neural network (CNN) [21] to embed our images. The model was pre-trained on the ImageNet dataset [22]. We replaced the final layer with a layer having a number of heads pertaining to the number of classes in our task whose weights are initialized uniformly at random [23]. We used a cross-entropy loss function where the loss is computed and aggregated over the entire dataset.

2.1.2 DRO and Abstention

Models were trained using an abstention algorithm (Algorithm 1) whereby we only accumulated and backpropagated the losses from images for which the model predicts a class with a normalized softmax logit score greater than a predefined threshold, p. We interpret this threshold as a confidence and only report losses on images for which the confidence value is

Input: abstention threshold p, forward function f, optimizer g, loss function $\mathcal{L}$
Output: θ, the parameters of the model
Initialize θ;
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: Forward Propagation of Loss in Abstention architecture

greater than p. This approach is aligned with potential clinical support use cases, whereby a model can be allowed to abstain if the data are not sampled from the same distribution it has been trained on using a confidence threshold to aid the decision making process whether the sample is out-of-distribution (OOD) or not. To rescale the outputs of the softmax function into a probability distribution for thresholding by p to select on the order of $2p$ samples, we used temperature scaling [24].

2.1.3 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 [25]. 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 [26]. 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.

We compared ERM against DRO on data from an external validation set consisting of unseen data coming from a comparable dataset collected at another hospital. We ablated the number of hospitals contributing the external validation dataset to measure the robustness of the ERM and DRO methods.

For the DRO methods, we report test statistics, such as F1 and loss, on images for which the model reports softmax logits with confidence values greater than p. We compute a macro-F1, aggregating the F1 scores of the individual classes without weighting them by the number of samples.

2.2. Tasks

2.2.1 Predicting the source site of a histopathology tissue

We used an image-classification algorithm to predict a scan's source hospital for LUAD images. After image quality control (QC) done using HistoQC [27], we used image augmentation techniques to mask the source-hospital signal and evaluate the models' performances under these distributional shifts. We cross-validated the experiment multiple times with different, random splits of training and validation sets in each iteration. However, while the choice of slides used in training and validation sets were made independently of other runs, slides were allowed to overlap between iterations, owing to data availability constraints.

Data Imbalance There was an uneven distribution of tiles across hospitals donating to TCGA. Balancing the number of WSIs and the number of QC-checked tiles from each hospital proved challenging as some hospitals contributed only a single WSI. Thus, we limited our study to the five most populous hospitals, as measured by the number of WSIs from the site, unless mentioned otherwise.

Data Splitting The data were split into training and validation sets in two different ways: 1) data from held out patients, who may be from the same hospital that was used to train; and 2) data arbitrarily split such that tiles used for the training set were taken from the left 70% of each WSI, and the remaining tiles in the slide were used for validation. In the second method, data from the same patient and hospital could be used in both the training and validation sets without reusing the same tiles since we had multiple, non-overlapping tiles from the same patient and hospital.

LUAD-i	Identifying tumor in tissue from i held out hospitals
PRAD	Identifying whether the tumor is Gleason score 6 or greater than 6
PRAD - TF	Identifying whether the tumor is Gleason score 6 or greater than 6, after filtering out non-tumor tissue
cc-RCC	Identifying whether the tumor is grade 2 or 4
cc-RCC-TF	Identifying whether the tumor is grade 2 or 4, after filtering out non-tumor tissue

Table 1: Tasks used to compare DRO and ERM models

Interpretability We leveraged Grad-CAM [28] as an initial step in interpretability. Grad-CAM outputs require examination on an image-by-image basis, which is difficult to scale when using millions of images to train. To address this issue, we used the Intersection Over Union (IOU) score between the Grad-CAM masks of different models and examined the overlap as an aggregate to quantify the performance of the model. Specifically, in the task of identifying the source site, we examined the overlap between the Grad-CAM masks produced by the different models trained on different image augmentation techniques for the same task. We hypothesized that if the features that allow a model to perform on the task transcend the color distribution of the tile and are robust to data augmentation techniques or non-stain artifacts of the tissue, such as tissue thickness, scanner quality, or other artifacts introduced by the pre-processing pipeline, there will be a high overlap in the Grad-CAM masks of the model applied to images augmented using different methods. In this way, we used the IOU score as a measure of agreement between two Grad-CAM heatmaps.

2.2.2 Comparing ERM vs. DRO

We evaluate our ERM and DRO algorithms on the tasks described in table 1. We provide relevant detail on the tasks here below.

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 tissue patches into tumor or normal tissue.

Predicting Grade of tissue in TCGA-ccRCC We classified tumor tissues from TCGA-ccRCC into Grade II or Grade IV cancer using slide-level labels. In order to prevent introducing confounders to the model, we first trained a model to detect tumor tissue in TCGA-ccRCC. This model was trained from healthy surrounding tissue, from an in-house dataset. We proceeded with subsequent analysis on tiles of the WSI that showed higher likelihood of being tumor tissue than healthy tissue. We also repeated the experiments on the whole dataset without removing non-tumor tiles for the sake of completeness.

Prostate Adenocarcinoma (PRAD) We predicted the aggregate Gleason score of a TCGA-PRAD ($n = 371$) tile using a binary classifier of low (score of $\leq$6) or intermediate/high (> 6). We first eliminated tiles that had a less than random chance of being tumor using predictions made on patch-wise labels and data from Schömig-Markiefka et al. [12]. We also performed some experiments where we used all patches.

3. Results

3.1. Heterogeneity in predicting tumor vs. normal tissue

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). For example, a model trained on data from the University of Pittsburgh achieved a validation F1 of 0.97 when validated on data from a held-out set of patients from the University of Pittsburgh, but, at best, only achieved a validation F1 of 0.72, when evaluated on data from Prince Charles Hospital.

We then consolidated the data by aggregating across hospitals whose data were used to train and validate, again observing inter-hospital validation heterogeneity (Appendix figure 7). We also found that hospitals whose data on which models achieve a higher validation F1, do not achieve comparable performance when models trained on that same site's data are validated on other hospitals, and vice versa. For example, a model trained on data from the University of Pittsburgh, achieved a median validation F1 of 0.87 when validated on other hospitals. However, models trained on data from other hospital sites and

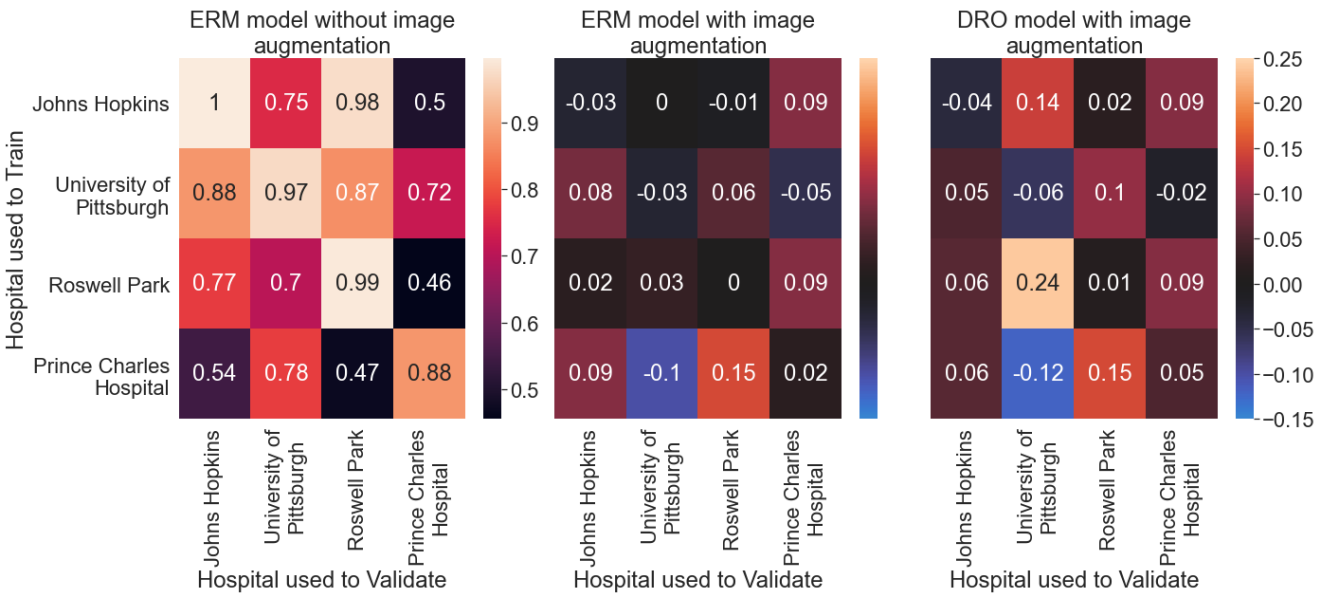


Figure 1: High Amount of heterogeneity in performance depending on which hospital's data are used to train. Middle and right grids shown are differences over grid on the left.

validated on the University of Pittsburgh cohort achieved a median F1 of 0.74. Further, for data from the hospitals at the University of North Carolina and Roswell Park, models achieved higher performance when used for validation (0.95 and 0.92 median F1, respectively) rather than for training (0.78 and 0.74 median F1, respectively). It is possible to reduce the heterogeneity that arises from the data preparation and pre-processing steps.

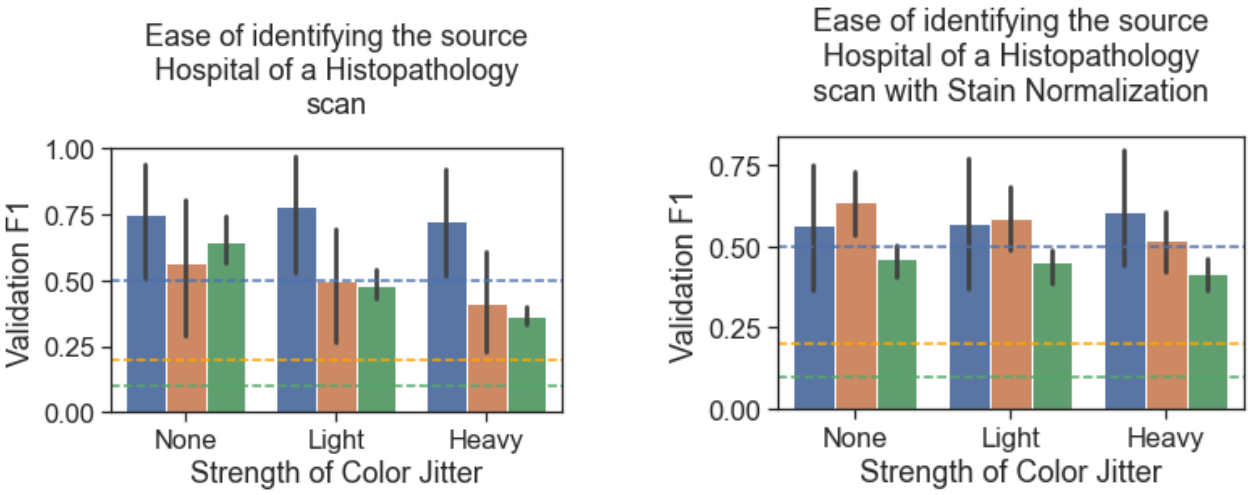
We also found that heavy color jitter produced only up to 0.15 improvement in F1 and using our abstention model produced up to 0.24 improvement in F1 when used in conjunction with heavy color jitter. To this effect, we propose using the DRO model to be more robust to the heterogeneity in training data and OOD validation data.

3.2. Impact of image augmentation on identifying the source site of an image

Given the heterogeneity in model performance, we next evaluated a possible source of this heterogeneity that arises from the data preparation and pre-processing steps. Consistent with prior reports [29], we found that a model could recognize the source site of a histopathology scan through visual features (Figure 2a). Thus, we assessed how image augmentation techniques like random changes in the brightness, saturation, and other properties of an image , also known as color jitter [26], might impact a model's performance in identifying the source site of an image. We were able to identify the source hospital of a histopathology scan without any color jitter with 0.72 validation F1 on a hold-out patient set when distinguishing between five hospitals and 0.61 validation F1 when distinguishing between ten hospitals. Reducing the color jitter strength to a light color jitter decreased the model's ability to decipher the source site of the image when distinguishing between both five and ten hospitals (a decrease of 9% and 16% validation F1 for 5 or 10 hospitals, respectively). Increasing the color jitter strength to a heavy color jitter decreased model performance even further (33% and 24% validation F1 for 5 and 10 hospitals, respectively). Given that the heavy color jitter was close to the maximum amount of color perturbation possible, we concluded that the source site signal is partially encoded in the stain profile of an image. To mask out the stain profile, we normalized the stain across the images. However, in spite of stain normalization, we were still able to distinguish the source hospital of an image among 5 source hospitals (Figure 2b) with 0.67 validation F1 under no color jitter, 0.57 validation F1 under light and 0.52 validation F1 under heavy color jitter. Thus, source hospital information is, at least, in part encoded in the stain profile of the scan, which can only be partially occluded by image augmentation techniques, such as color jitter and stain normalization.

3.3. Using Grad-CAM to identify features contributing to source-site prediction

In order to understand the features contributing to source-site signal, we used Grad-CAM (Section 2.2.1) [28]. We found that the image augmentations did not drastically alter the regions of the image highlighted by Grad-CAM (Figure 3). Further,



(a) Increasing color jitter does not eliminate a model's ability to identify the source site of a histopathology image

(b) Ability to identify source site despite normalizing the stain across all images

Figure 2: Distinguishing between 2, 5, 10 hospitals shown in blue, orange, green respectively. Random performance shown by color-coded dashed line (random performance for blue bar shown by blue dashed line, etc.)

we found that Grad-CAM segmentations did not agree with any discernible boundaries of objects in the image, making the masks hard to interpret. Thus, we could not identify additional interpretable features that contributed to source site prediction. By IOU, we found that there was a high overlap between the masks corresponding to the strategies used to prevent overfitting (median 0.94), which suggests that these methods did not mask out the signal used to identify source hospitals entirely.

3.4. Heterogeneity in source site Identification

In addition to distinguishing between individual source sites, we found that our model was able to distinguish between two arbitrary groups of source sites, G_1 and G_2 (Figure 4). However, we found that a validation set of slides containing only tumor tissue was OOD when the training set included slides with only healthy lung tissue, and vice versa. For example, when forming G_1 and G_2 of sizes two and three, a training set of images containing both tumor and healthy tissues resulted in a validation F1 of 0.76, but training on images containing only tumor tissue and validating on images of surrounding healthy tissue, or vice versa, resulted in validation F1s of 0.58 and 0.61, respectively. More so, when the training set and validation set included images with both tumor and healthy tissues, the performance of the model improved by up to 25% F1. This suggests that features learned to distinguish the sources are not limited to the stain pattern common to the source since this signal would be agnostic to whether the tissue is tumor or stroma. Thus, model performance should be validated across different subgroups of the data to be considered robust.

3.5. Using group-DRO to improve generalization in identifying tumor tissue in TCGA-LUAD

Given the multiple challenges presented by batch effects, we trained a model with group-DRO to evaluate this approach's robustness to spurious confounders. 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 (Table 2)

Upon investigation, we noted that these methods resulted in decreased heterogeneity in the model predictions (Figure 20); that is, a DRO model abstained on tiles that an ERM model predicted wrongly (since they might have presented features corresponding to classes different from the ground truth) and thus learnt more robust features. Further, DRO models predicted the same class in a greater majority of the patches that they did not abstain on. The ERM methods wrongly predicted non-tumorous regions of a WSI as tumorous, even at higher confidences. On the other hand the abstention method abstained on tiles where the features of the tile do not align with the distribution of features pertinent to the label given to the WSI. Also, DRO models trained at high confidence thresholds abstained from making predictions on regions of the WSI covered by slide-preparation artifacts, such as air bubbles (Figure 21). By its abstention from artifacts, the abstention method is less

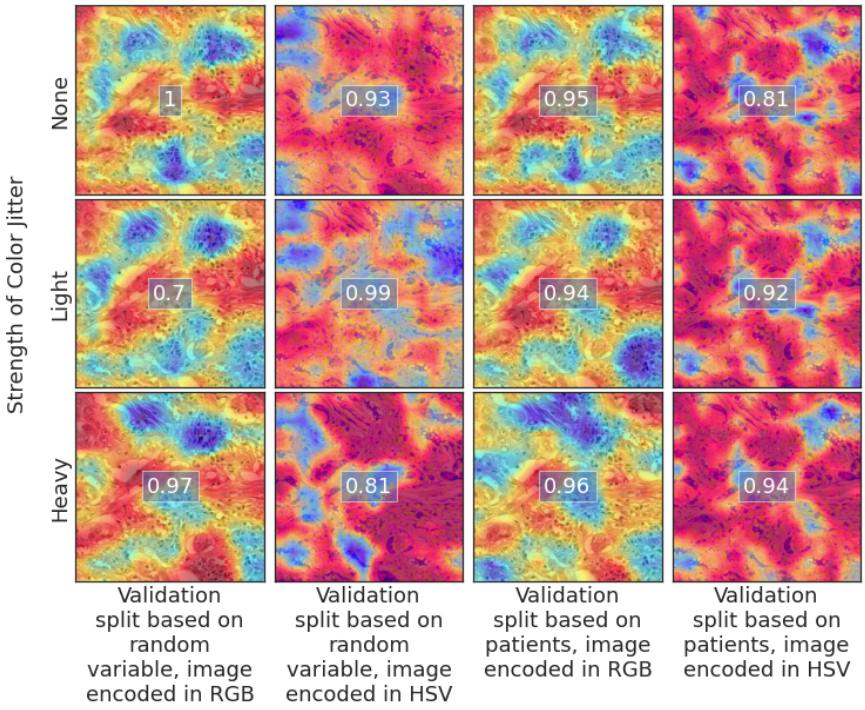


Figure 3: Using Grad-CAM to highlight features of an image relevant to source-site prediction shows that image augmentation techniques do not affect the areas important to recognizing source site of the scan. Inset shows average overlap of mask with top-left mask (as averaged across all images of the test set).

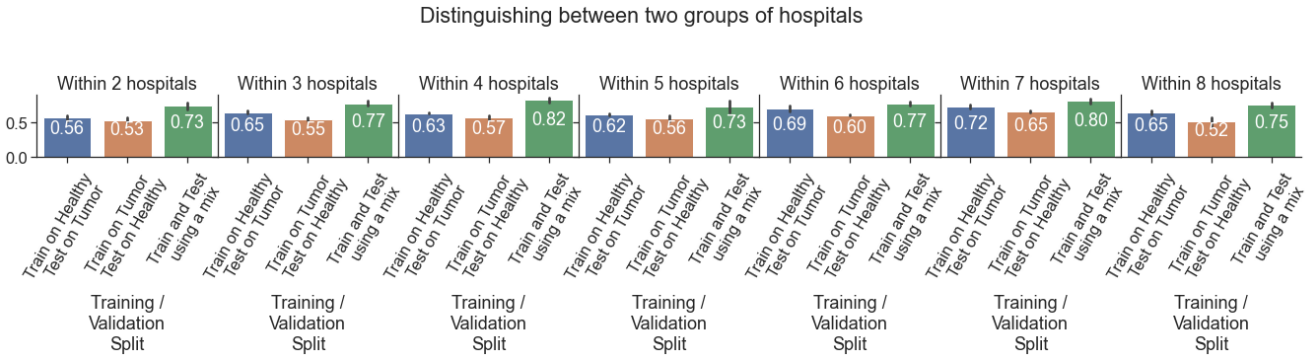


Figure 4: Models trained to distinguish between two groups of source sites, G_1 and G_2, produce different results depending on whether the model was trained using images with only tumor tissue, only healthy tissue, or a mix of both

likely to learn spurious correlates.

3.6. Using group-DRO to improve generalization in grade prediction in TCGA cc-RCC

Regarding cc-RCC analyses, we observed an improvement by 18.5% F1 after first removing tiles that do not contain tumor and up to 19.4% F1 when including non-tumor tiles (Table 2) in the task of identifying whether a tile comes from a slide of grade 2 or 4 tumor.

3.7. Predicting Gleason score in TCGA-PRAD

Next, we compared the performance of a group-DRO method to a model trained with ERM on predicting Gleason score in PRAD. The model trained with group-DRO performs up to 5.9% better than a model trained with ERM without first removing

Task	ERM	Abstention threshold			
		0.6	0.7	0.8	0.9
LUAD-1	91.1	93.1	96.6	**98.4**	97.2
LUAD-2	88.3	88.6	89.1	93.3	**95.2**
LUAD-3	78.6	77.5	78.9	78.6	**82.0**
LUAD-4	81.1	78.9	80.4	84.0	**91.0**
LUAD-5	72.2	73.4	71.9	76.3	**79.7**
PRAD	68.8	67.1	62.8	**74.7**	65.2
PRAD-TF	74.7	50.3	78.3	**84.7**	84.4
cc-RCC	64.4	68.6	71.2	**83.8**	76.1
cc-RCC-TF	68.1	69.7	72.3	**86.6**	73.2

Table 2: Comparing the test F1 of a normal CNN trained with ERM with a group-DRO method trained with abstention and a group-DRO method trained with abstention and temperature scaling. Our proposed model outperformed a conventional CNN in F1.

non-tumor tiles (Table 2). After filtering out the patches that did not contain tumor, the group-DRO method performed up to 13% better (Table 2). Thus, we propose using group-DRO models to learn outcomes based on weak labels, since group-DRO models are more robust to spurious correlates in the image that do not align with the WSI label.

4. Discussion

In this study, we showed that stain profile can be used to identify the source site of a histopathology scan and contribute to significant heterogeneity in model performance. This artifact might lead a model to overfit spuriously correlated features of the slide while training on a label with weak morphological evidence.

In our analyses, we took five slides from each hospital and one hundred tiles from each slide. The differences between source sites could reflect differences specific to those tiles that were selected. However, the models' ability to correctly identify the source site of a tile among ten sources despite using image re-coloring techniques and stain normalization implies that there are features of an image that provide sufficient visual evidence for a model to identify the source site of an image. It is possible that these features could be biological, (e.g., differences in grade, tumor-infiltrating lymphocyte infiltration, metastatic potential, or other features that are enriched in the source site's data), so consideration of such batch effects are key for successful analysis of these data types.

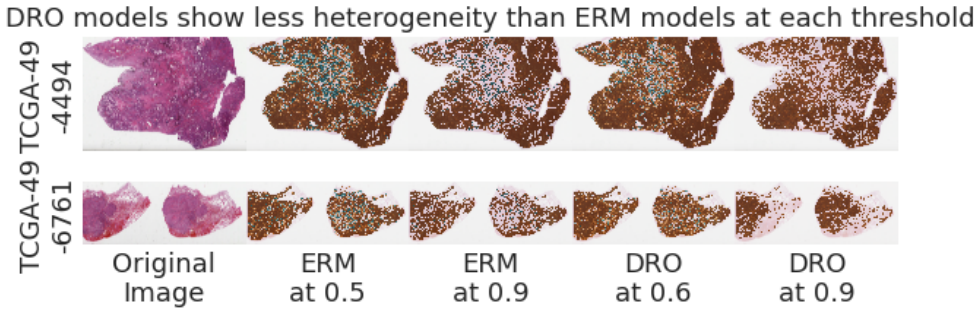
We found that models achieved different performances in the task of identifying LUAD when trained on data from one hospital and tested on those of another site. We attributed this to a difference in the distributions of spurious variables between the training and validation datasets. We hypothesize that if a model tested well on data from a hospital while using data from other hospitals to train, the testing data are a narrow distribution of spurious and core variables that fall within the training data manifold.

The prevalent assumption is that source hospitals of histopathology tissue differ only in their stain profiles and staining techniques. We did not expect a model trained to distinguish between two groups of source hospitals to learn some biological features. However, the features learned by the model did not hold when evaluated on a different tissue type, implying that the model is classifying based on features specific to the tissue type, different from the core variable of stain template used by the hospital.

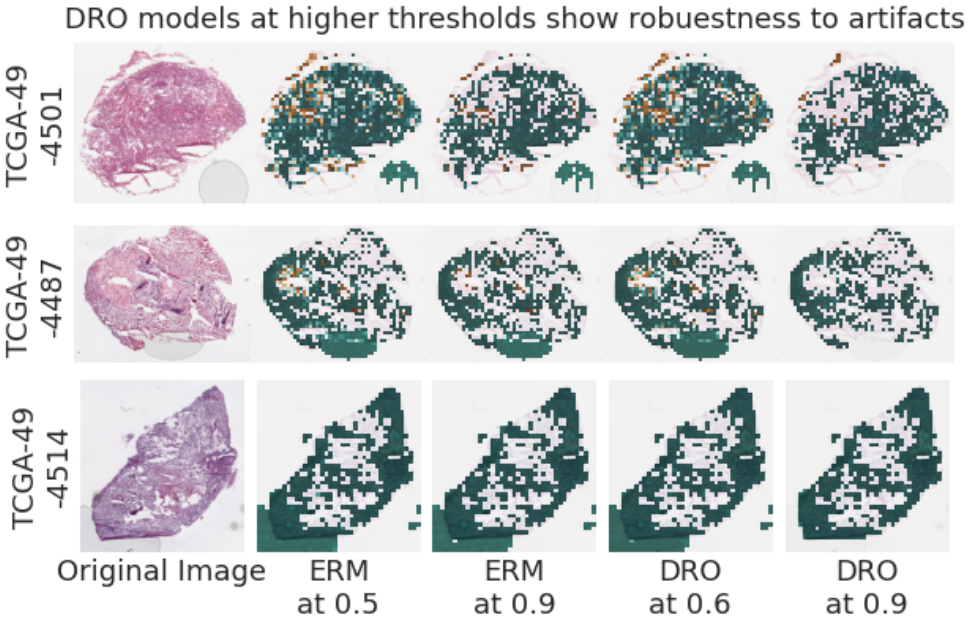
Ultimately, we found that DRO methods that aim to either optimize the model's performance on a previously defined subgroup or a learned subgroup, defined in our case by the training samples that the model performs well on, were able to provide better performances on an external validation set.

5. Conclusion

Here, we evaluated the impact of batch effects and developed approaches to mitigate these fundamental challenges to digital pathology. We assessed how source sites can be learned by models, evaluated existing approaches to address known sources of batch effects, and highlighted batch effect features that, although unseen, can still impact downstream analyses. We also evaluated the role of the interpretability tool, Grad-CAM, and proposed a neural network that is robust to the distributional shifts between training and held-out test sets. Prospectively, careful consideration of seen and unseen batch



(a) Reduced heterogeneity in a model trained using DRO compared to ERM in tumor vs. normal identification in LUAD. Brown indicates patches that were predicted as tumor, blue indicates patches that were predicted as normal tissue. DRO models with higher confidence thresholds abstain on tiles that an ERM model predicts as normal tissue, thus avoiding learning contradictory features (first row). Second row: ERM methods call non-tumor region on the right hand side of the tissue as tumor, even at high confidence thresholds. DRO methods abstain on tiles where the tissue does not bear tumor.



(b) ERM methods predict bubble artifacts as healthy surrounding tissue. DRO methods at higher confidence thresholds abstain from making predictions on artifacts.

Figure 5

effects in CV digital pathology analysis will guide successful biological investigations with potential clinical impact.

6. Acknowledgements

SNH thanks Sneha Jha, Haitham Elmarakeby, Brendan Reardon, Parimarjan Negi and for their helpful comments.

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] X. Zhu, J. Yao, F. Zhu, and J. Huang, "WSISA: Making Survival Prediction from Whole Slide Histopathological Images," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, Jul. 2017, pp. 6855–6863. [Online]. Available: http://ieeexplore.ieee.org/document/8100208/

[8] W. Lu, S. Graham, M. Bilal, N. Rajpoot, and F. Minhas, "Capturing Cellular Topology in Multi-Gigapixel Pathology Images," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, WA, USA: IEEE, Jun. 2020, pp. 1049–1058. [Online]. Available: https://ieeexplore.ieee.org/document/9150693/

[9] 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

[10] M. Schmitt, R. C. Maron, A. Hekler, A. Stenzinger, A. Hauschild, M. Weichenthal, M. Tiemann, D. Krahl, H. Kutzner, J. S. Utikal, S. Haferkamp, J. N. Kather, F. Klauschen, E. Krieghoff-Henning, S. Fröhling, C. von Kalle, and T. J. Brinker, "Hidden Variables in Deep Learning Digital Pathology and Their Potential to Cause Batch Effects: Prediction Model Study," *Journal of Medical Internet Research*, vol. 23, no. 2, p. e23436, Feb. 2021. [Online]. Available: https://www.jmir.org/2021/2/e23436

[11] F. M. Howard, J. Dolezal, S. Kochanny, J. Schulte, H. Chen, L. Heij, D. Huo, R. Nanda, O. I. Olopade, J. N. Kather, N. Cipriani, R. Grossman, and A. T. Pearson, "The Impact of Digital Histopathology Batch Effect on Deep Learning Model Accuracy and Bias," Bioinformatics, preprint, Dec. 2020. [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/2020.12.03.410845

[12] B. Schömig-Markiefka, A. Pryalukhin, W. Hulla, A. Bychkov, J. Fukuoka, A. Madabhushi, V. Achter, L. Nieroda, R. Büttner, A. Quaas, and Y. Tolkach, "Quality control stress test for deep learning-based diagnostic model in digital pathology," *Modern Pathology*, Jun. 2021. [Online]. Available: http://www.nature.com/articles/s41379-021-00859-x

[13] 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

[14] 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

[15] 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

[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] S. A. Fuhrman, L. C. Lasky, and C. Limas, "Prognostic significance of morphologic parameters in renal cell carcinoma:," *The American Journal of Surgical Pathology*, vol. 6, no. 7, pp. 655–664, Oct. 1982. [Online]. Available: http://journals.lww.com/00000478-198210000-00007

[19] D. Karimi, G. Nir, L. Fazli, P. C. Black, L. Goldenberg, and S. E. Salcudean, "Deep Learning-Based Gleason Grading of Prostate Cancer From Histopathology Images-Role of Multiscale Decision Aggregation and Data Augmentation," *IEEE journal of biomedical and health informatics*, vol. 24, no. 5, pp. 1413–1426, May 2020.

[20] J. Wang, R. J. Chen, M. Y. Lu, A. Baras, and F. Mahmood, "Weakly Supervised Prostate TMA Classification via Graph Convolutional Networks," *arXiv:1910.13328 [cs, eess, q-bio]*, Nov. 2019, arXiv: 1910.13328. [Online]. Available: http://arxiv.org/abs/1910.13328

[21] 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

[22] 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.

[23] T. Pytorch, "pytorch/pytorch," Sep. 2021, original-date: 2016-08-13T05:26:41Z. [Online]. Available: https://github.com/pytorch/pytorch/blob/88fff22023b201ee237ab0856d53a154cc1784bb/torch/nn/modules/linear.py

[24] 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

[25] 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/

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

[27] A. Janowczyk, R. Zuo, H. Gilmore, M. Feldman, and A. Madabhushi, "HistoQC: An Open-Source Quality Control Tool for Digital Pathology Slides," *JCO Clinical Cancer Informatics*, no. 3, pp. 1–7, Dec. 2019. [Online]. Available: https://ascopubs.org/doi/10.1200/CCI.18.00157

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Feb. 2020, arXiv: 1610.02391. [Online]. Available: http://arxiv.org/abs/1610.02391

[29] F. M. Howard, J. Dolezal, S. Kochanny, J. Schulte, H. Chen, L. Heij, D. Huo, R. Nanda, O. I. Olopade, J. N. Kather, N. Cipriani, R. L. Grossman, and A. T. Pearson, "The impact of site-specific digital histology signatures on deep learning model accuracy and bias," *Nature Communications*, vol. 12, no. 1, p. 4423, Dec. 2021. [Online]. Available: http://www.nature.com/articles/s41467-021-24698-1

[30] P. Byfield, "Peter554/StainTools: Patch release for DOI," Sep. 2019. [Online]. Available: https://zenodo.org/record/3403170

A. Supplementary methods

A.1. Data Extraction and Preprocessing

We obtained whole slide image (WSI) scans from TCGA. These WSIs underwent quality control (QC) using HistoQC [27] to filter out artefacts introduced during the slide preparation, digitization, and evaluation, such as tissue folding, whitespace, and pen marks, that might confound biological signals. The QC-checked images were processed into tiles by taking non-overlapping patches at 20x resolution of size 512 x 512 pixels.

A.2. Models and Loss Functions

To train a model to identify the source site of histopathology images using a multi-class prediction architecture with a cross-entropy loss function to predict the source site of a histopathology tile. Here, we use a pretrained ResNet-50 model [21] as our image-to-embedding encoder. We obtained the model from the Pytorch TorchVision library, and it was pre-trained on the ImageNet dataset [22].

We also trained a contrastive network to identify if two histopathology images come from the same source site. The loss function we used is described in equation 1, in which the embeddings of two images, x_1 and x_2, from a common encoder are trained to be similar if the two images come from the same source site and different otherwise.

$$\text{loss}(x, y) = \begin{cases} 1 - \cos(x_1, x_2), & \text{if } y = 1 \\ \max(0, \cos(x_1, x_2) - \text{margin}), & \text{if } y = -1 \end{cases} \tag{1}$$

The loss function described in equation 1 provides separation between the classes in the latent space. In order to augment interpretability to the latent space of source-sites and learn the specific correspondence between latent subspace and source-site, we added a cross-entropy loss to the contrastive embedding loss, thus making the model separate the embeddings in latent space and learn the correct mapping between the label and latent embedding distribution.

A.3. Preventing Overfitting

A.3.1 Early Stopping

We train our models to minimize error and stop training if the error does not improve on the validation set over five consecutive measurements [25]. The validation performance was measured four times per epoch. Thus, a lack of performance improvement for five consecutive measurements implies that the model's validation performance did not increase over one epoch. This number was not optimized for, since our goal was not to optimize for the outcome, but rather to compare the impact of training algorithm, image augmentation and validation split methods with the other hyperparameters aligned with common practice.

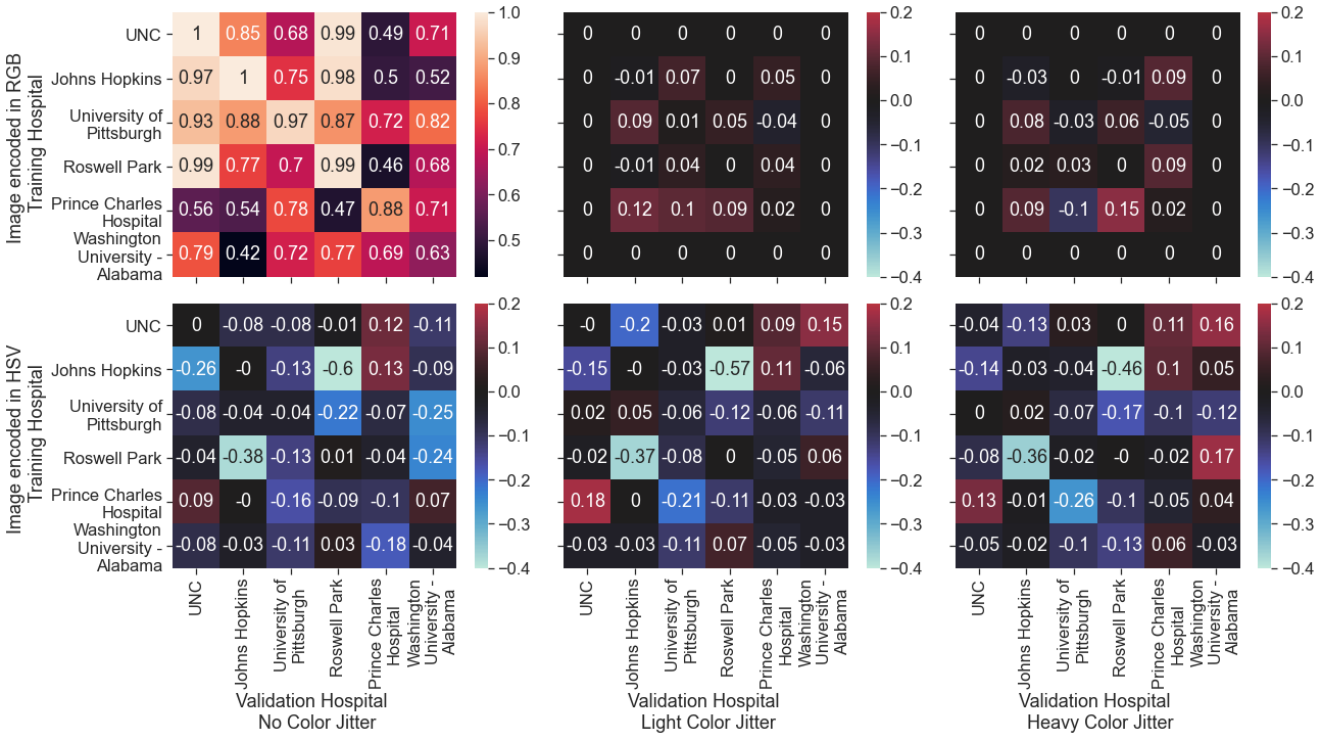
A.3.2 Color Augmentation and Normalization

We used image augmentation via jittering the RGB pixel values in the RGB space to prevent overfitting to the color distribution. In addition to using color augmentation, we also used stain normalization using Staintools [30]. We performed stain normalization in two ways: 1) Where the images in the validation set were normalized to the same template as the images in the training set and 2) Where the images in the validation set were normalized to a different template compared to the images in the training set. The first method was used to prevent the stain template of the image from creating a spurious correlate. The second method was used to test the model's reliance on morphological features that are still observable despite a distributional shift in the color profile.

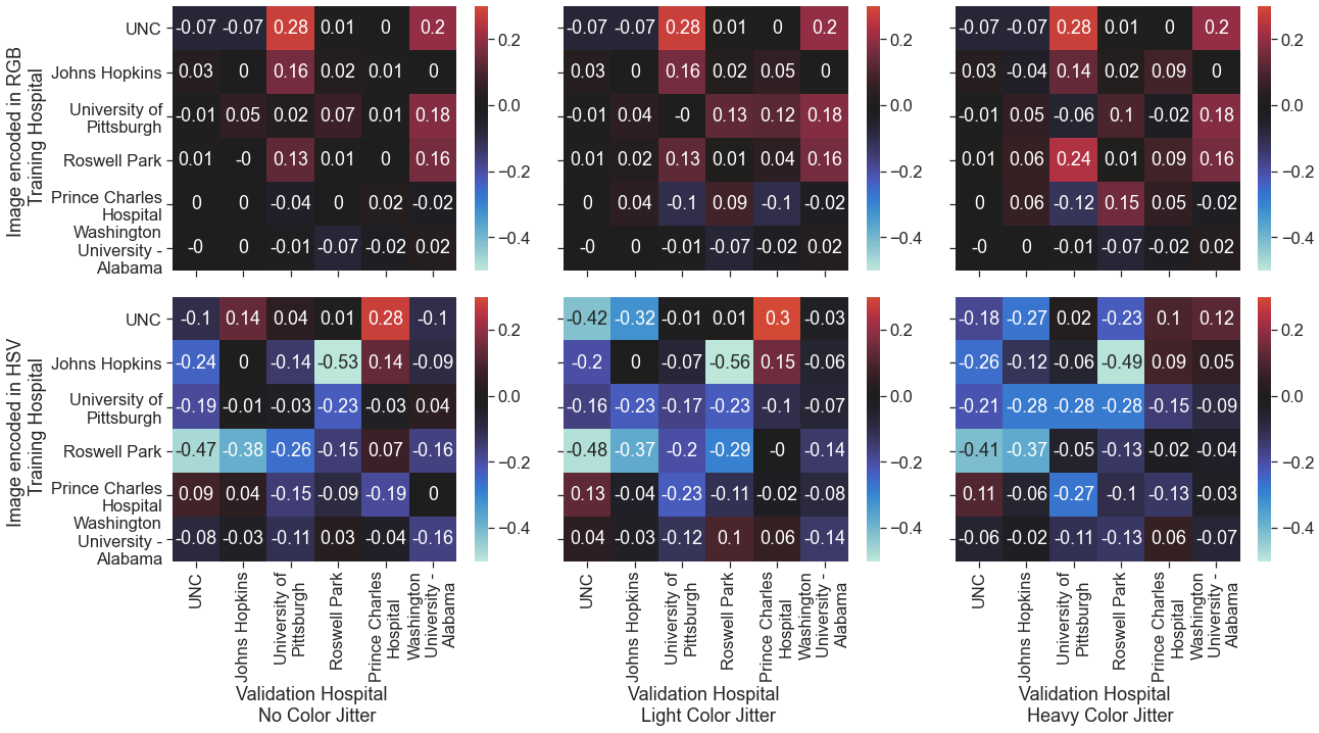
We used a random-crop size of 224 pixels within the 512 pixel patch during our training process as a data-augmentation technique.

A.4. Reporting F1

We reported the best validation F1 achieved by the model, unless stated otherwise. We continued to track the loss metric to evaluate further improvement by the model; however, an improvement in loss does not necessarily improve F1 and can lead to a worse F1 as well. Thus, we report the F1 at the training instant where the F1 is highest even if the model achieves a lower loss at a different time point.



(a)



(b)

Figure 6: Best Validation F1 achieved by a regular CNN model (6a) and a model trained with abstention (6b) trained on one hospital (y axis) and validated on another (x axis). All heatmaps except for top left of 6a are shown as change over top-left heatmap of 6a

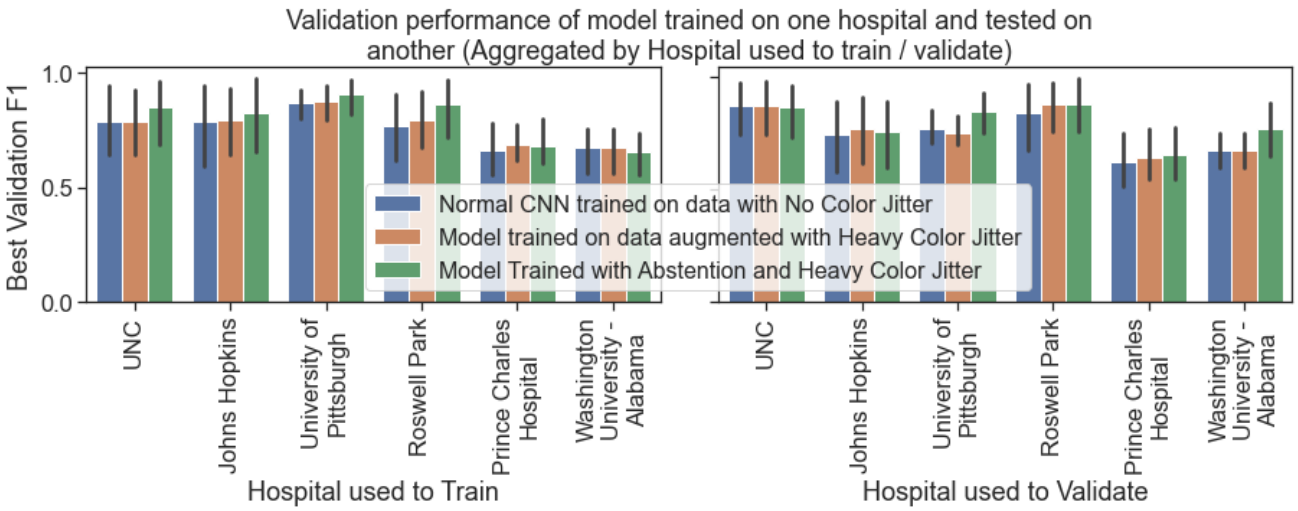


Figure 7: Data from Figure 1 aggregated across hospitals used to train and validate

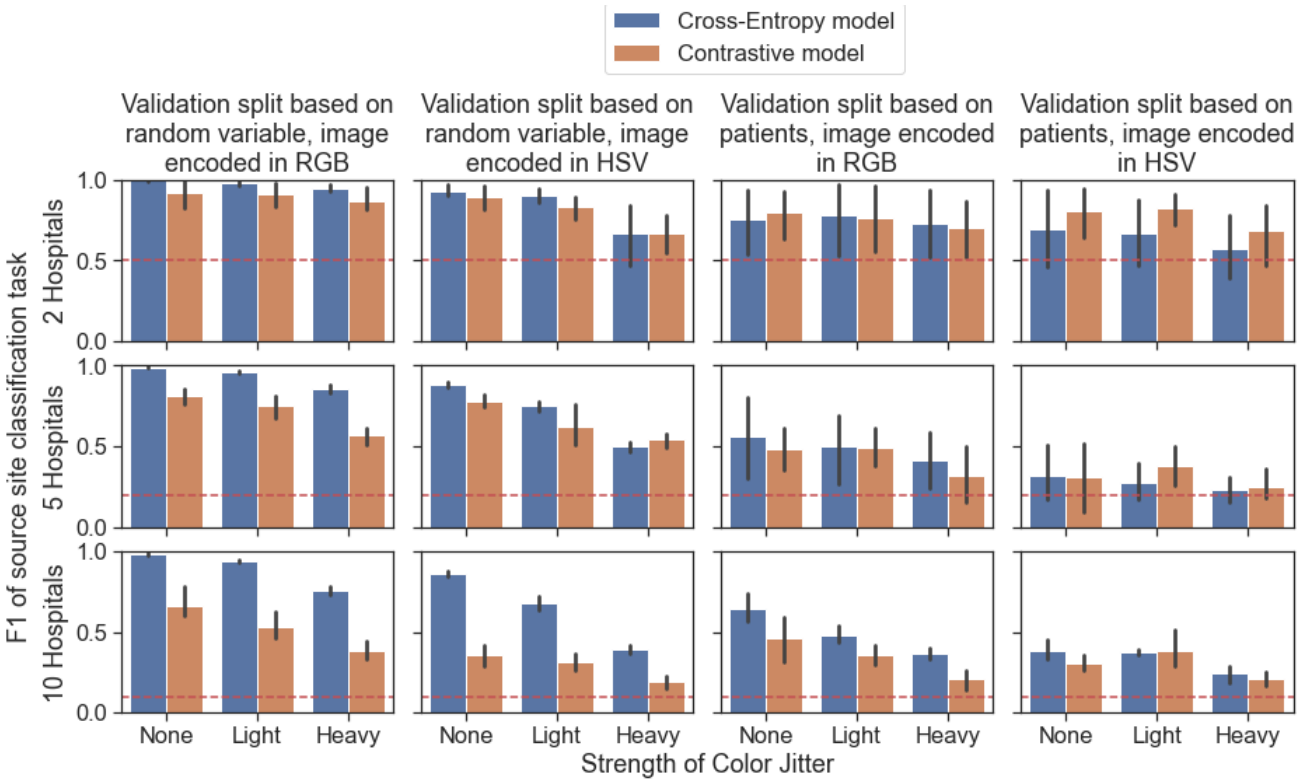


Figure 8: Validation F1 of a model used to predict the source site of a histopathology scan. We trained two models to predict the source site of a histopathology image (Figure 8). One model was trained using a Cross Entropy Loss (Blue) and the other was trained using a Contrastive Cosine Embedding Loss function (Orange).

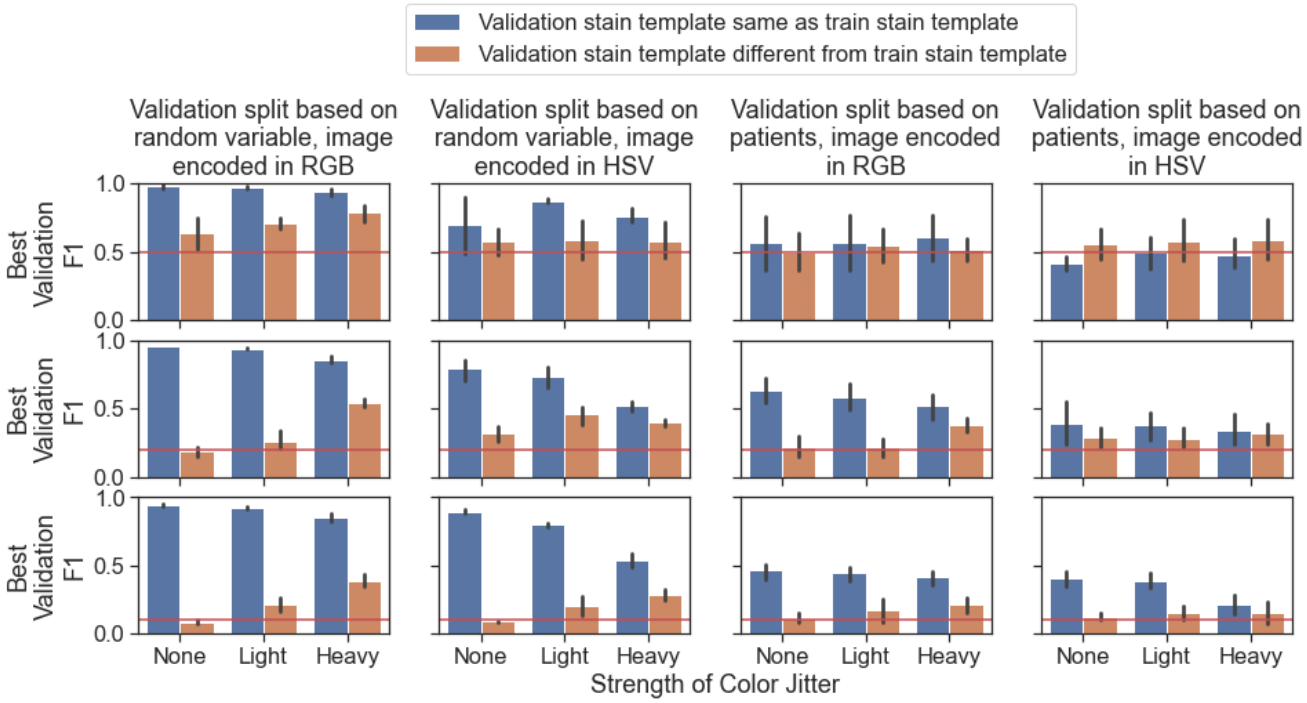


Figure 9: Validation F1 of a model used to predict the source site of a histopathology scan where the scans are stain normalized.

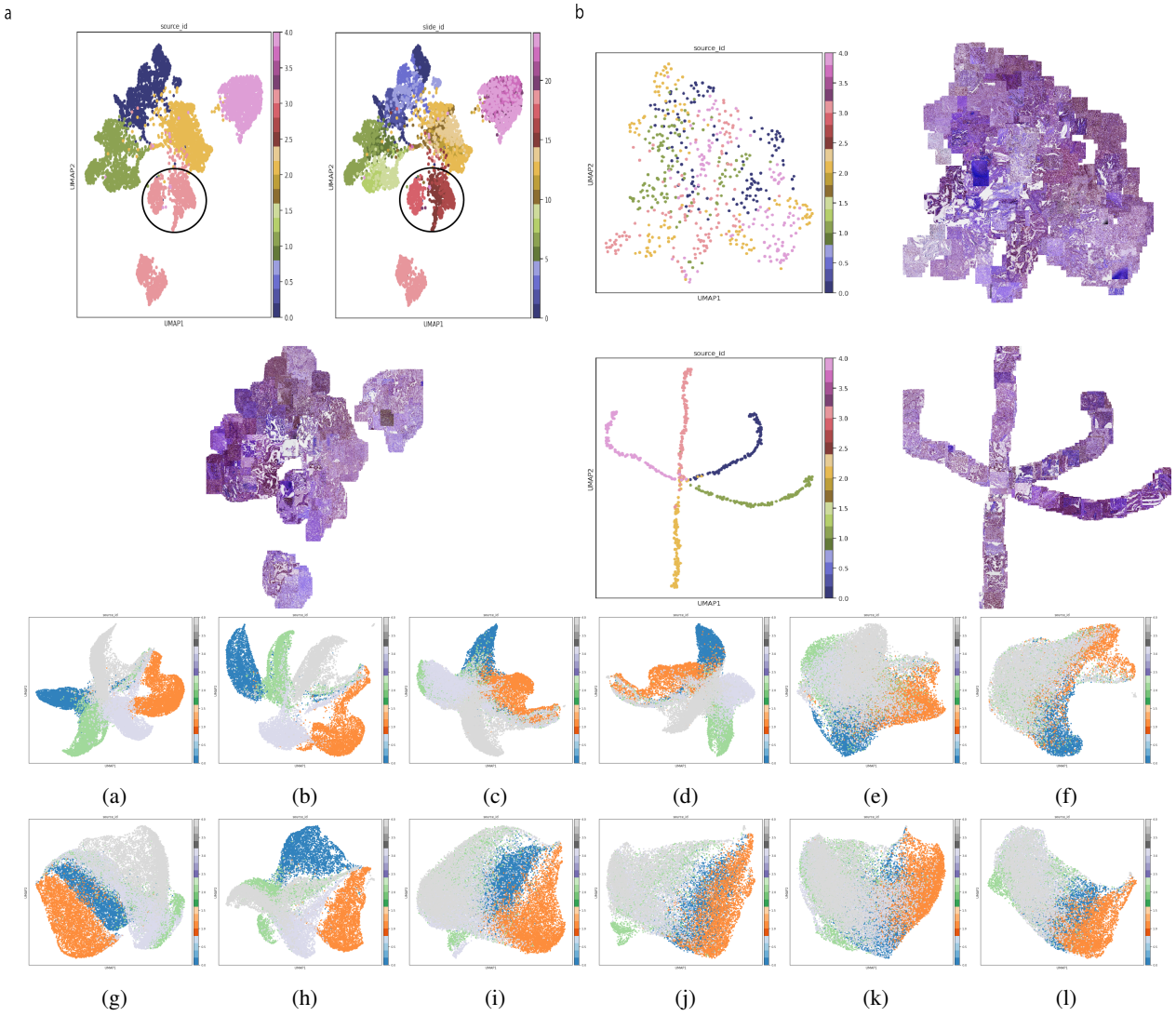


Figure 10: (Top Panel) a) (Left) UMAP of Embeddings from neural network trained using Cross-Entropy loss in identifying the source site of an image. Top-Left shows coloring by source site, Right, shows coloring by slide. Circle shows within a source site, slides cluster together without explicit training. Bottom shows UMAP with patchces of the images replacing points. (b) (Top Left, Right) UMAP from a CNN pre-trained on Res-Net with no fine-tuning. No clustering is evident. (Bottom Left, Right) UMAP of embeddings of a network trained using a Contrastive loss function. The latent space shows stronger clustering than that of a Cross-Entropy Function. (Bottom Panel) Effect of validation split, Color Jitter and Image encoding on a model trained with Contrastive Loss. a, b, g, h have No Color Jitter, c, d, i, j - Light, e, f, k, l have heavy Color Jitter. a, b, c, d, e, f are in RGB. g, h, i, j, k, l are in HSV. a, c, e, g, i, k don't hold out patients. b, d, f, h, j, l, hold out patients during validation

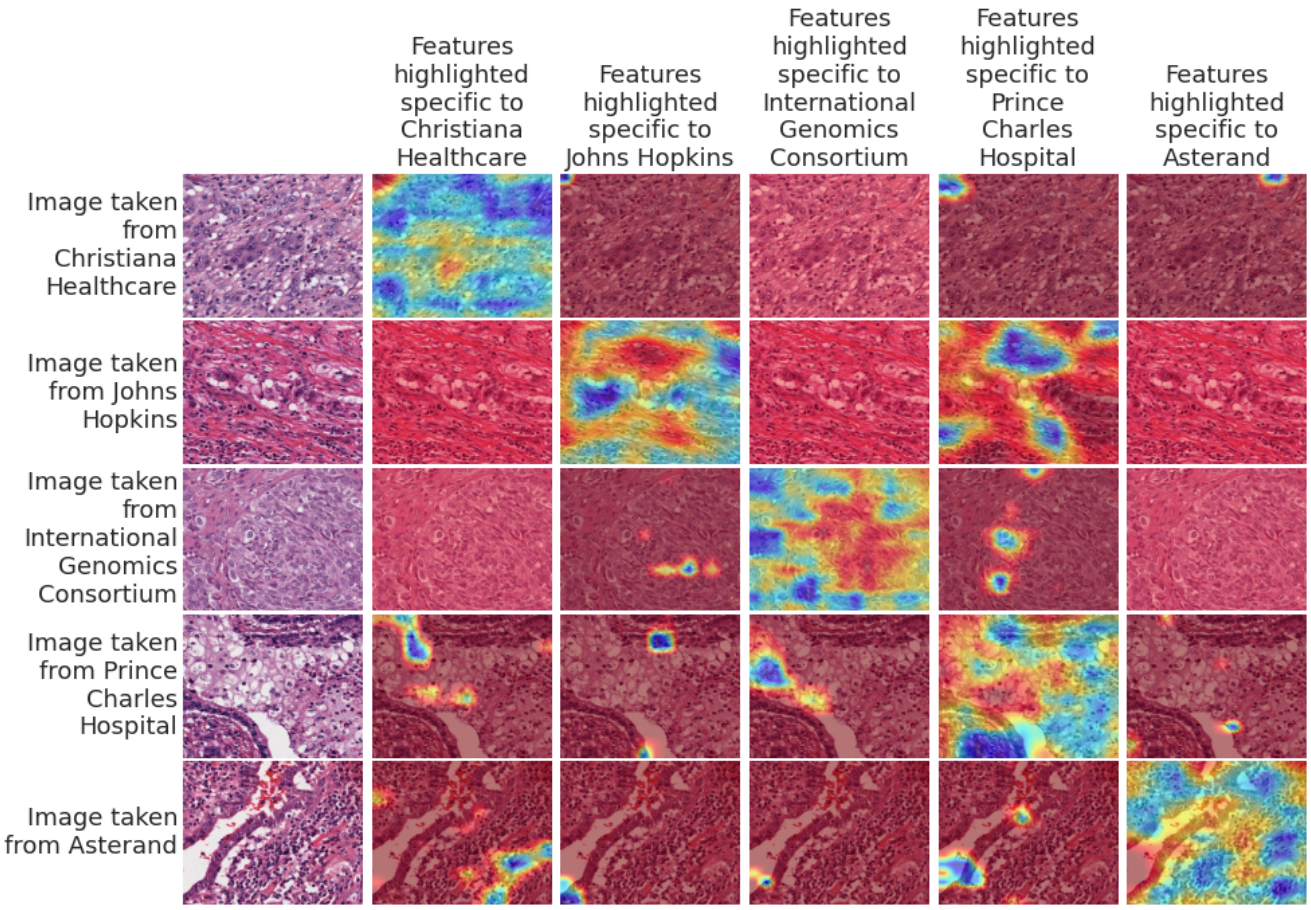


Figure 11: In order to interpret how models were able to identify the source site of the histopathology image, we used Grad-Cam to interpret the model's results. We present a grid of heatmaps such that the image in position (i, j) is the heatmap produced for an image from source i, with respect to what features of the image make the model think that it might be from source j. We note that in each row, the correct class produces the most Grad-Cam activity, aligned with our expectations.

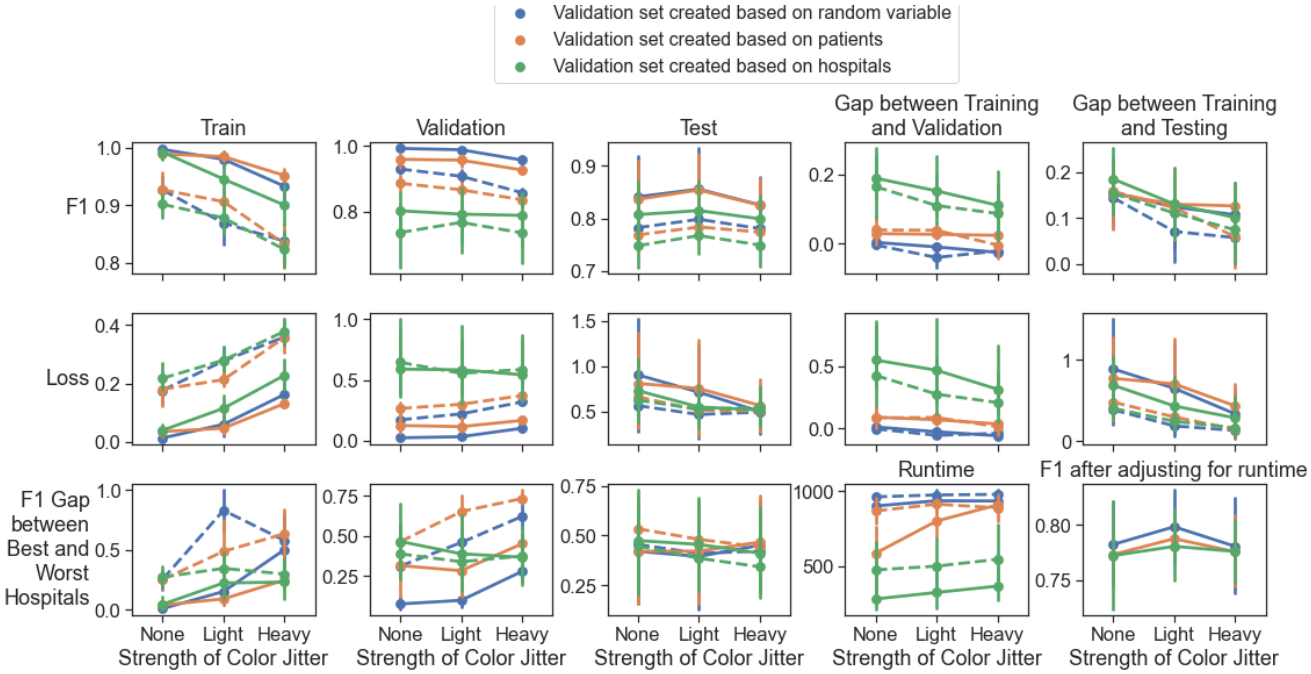


Figure 12: Measuring the impact of image augmentation techniques on the performance of a model to detect presence of tumor on a histopathology tile. Images encoded in RGB shown in solid, dotted lines show results for images encoded in HSV.

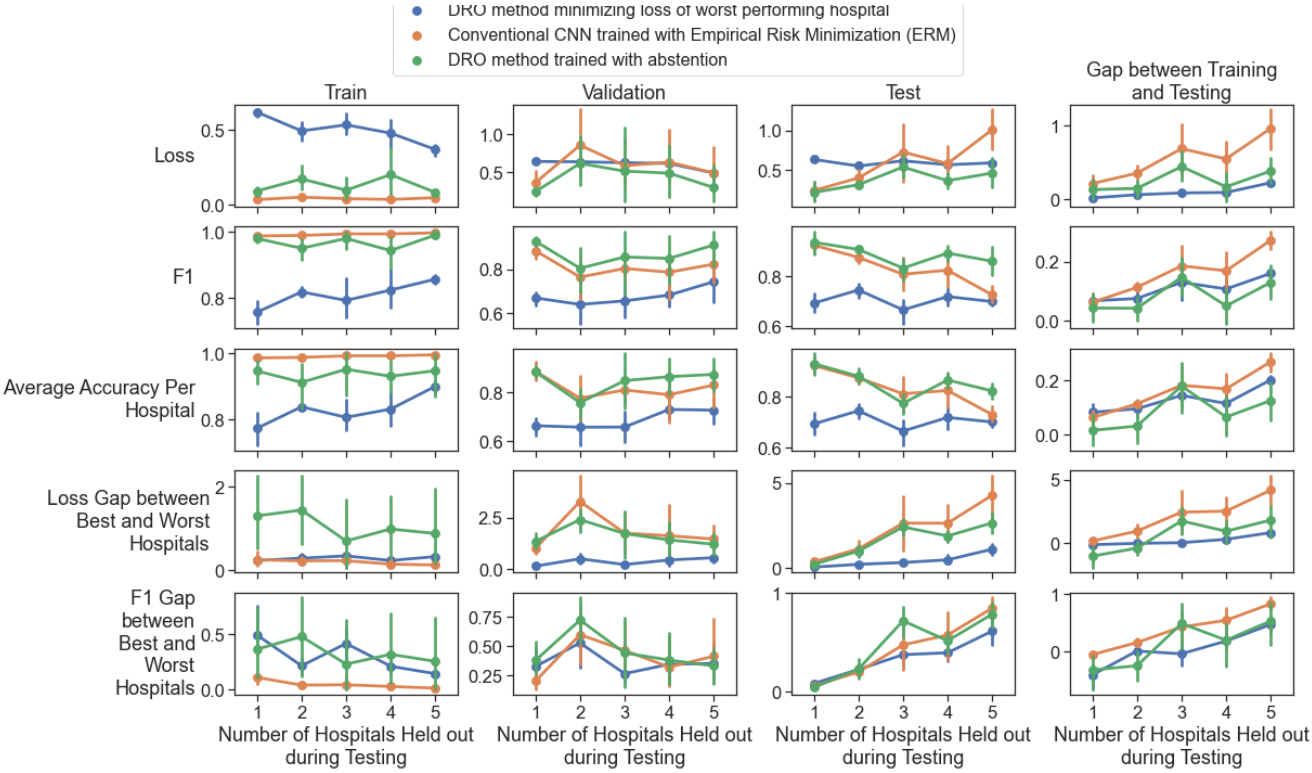


Figure 13: Comparing the performance of a Distributionally Robust Optimization (DRO) method minimizing worst case loss, a DRO method with abstention and a conventional CNN trained using Empirical Risk Minimization (ERM)

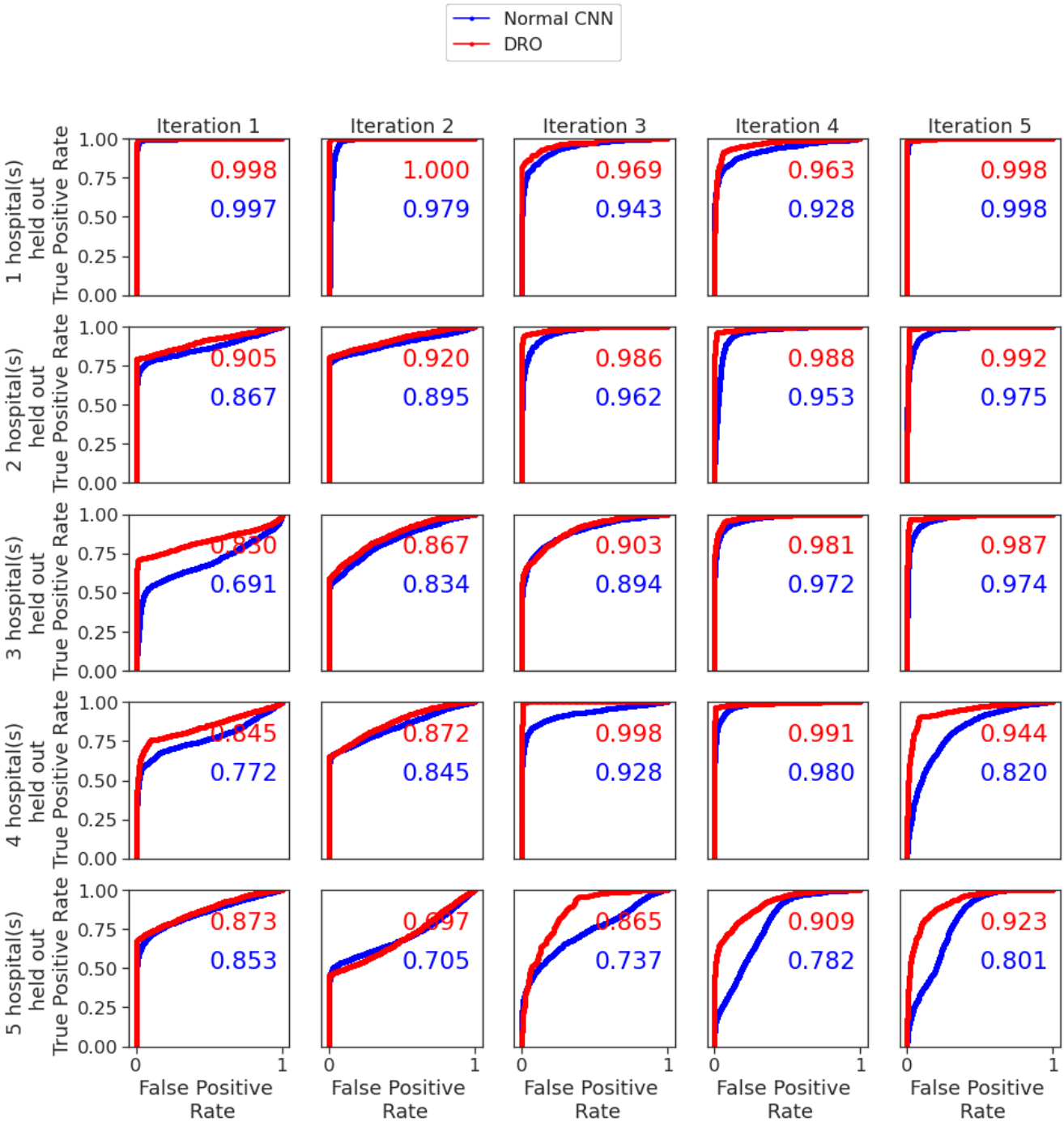


Figure 14: Receiver Operator Curves comparing the performance on a test set of an abstention method without temperature scaling to that of a conventional CNN. AUROC shown on figure

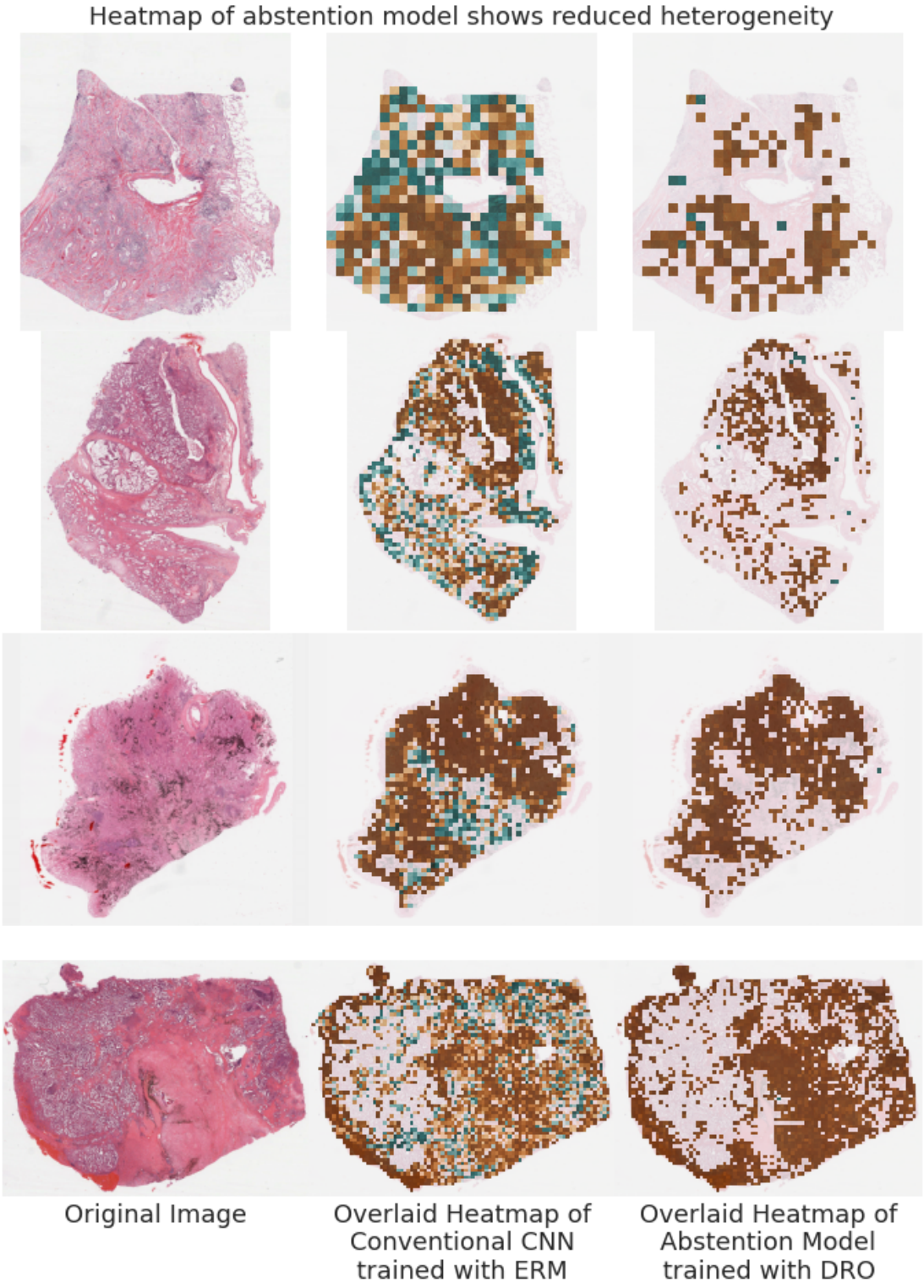


Figure 15: Reduced heterogeneity in model trained using Distributionally Robust Optimization (DRO). Brown indicates patches predicted as tumor in tissue, blue indicates patch was predicted as normal surrounding tissue

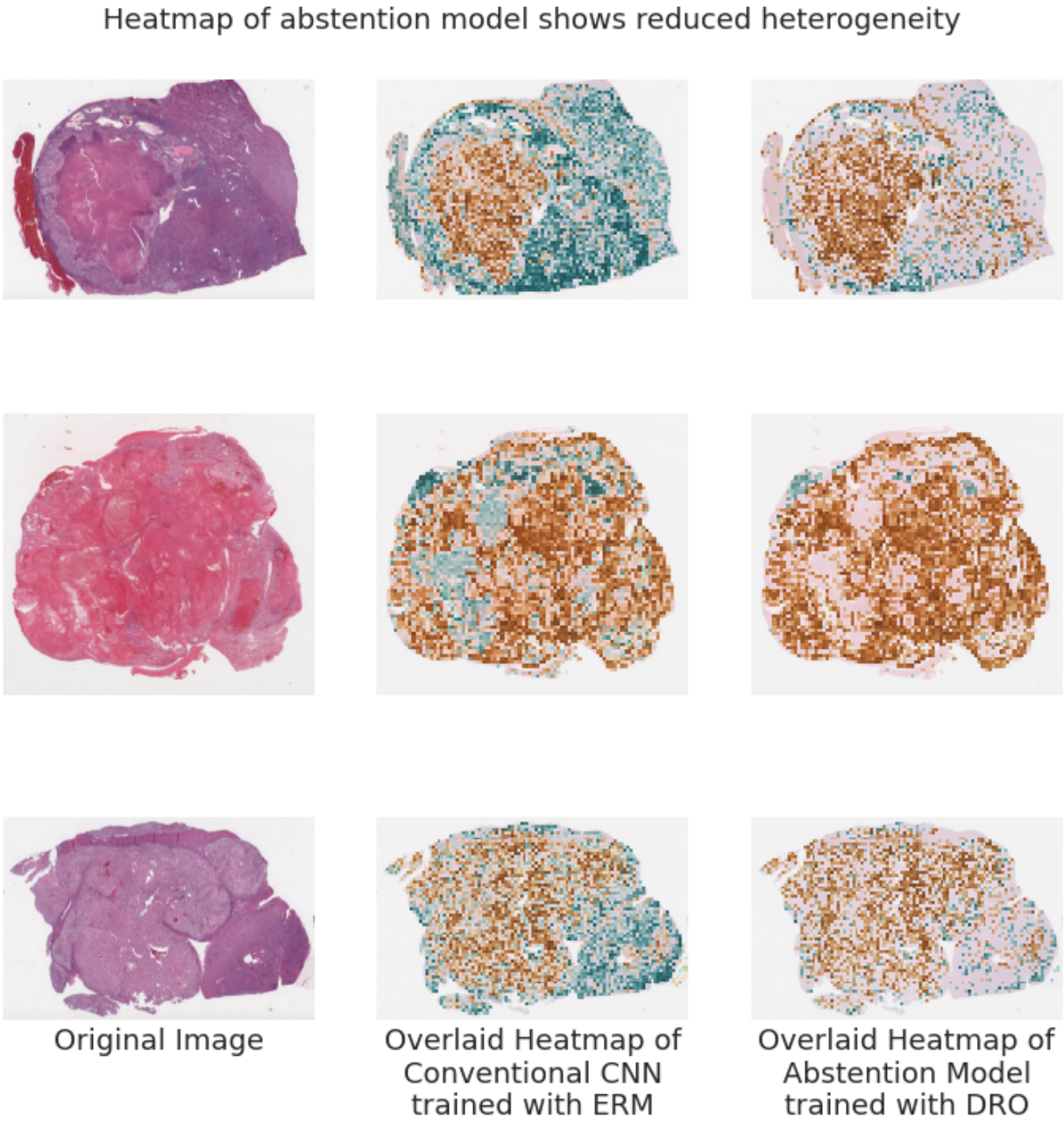


Figure 16: Reduced heterogeneity in model trained using Distributionally Robust Optimization (DRO). Brown indicates patches predicted as tiles predicted to be grade 2 tumor, blue indicates tiles predicted to be grade 4.

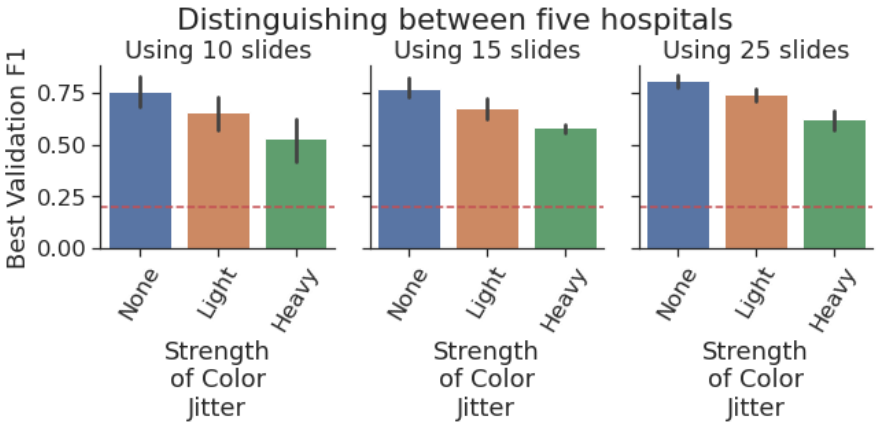


Figure 17: Distinguishing between 5 sources varying the number of slides

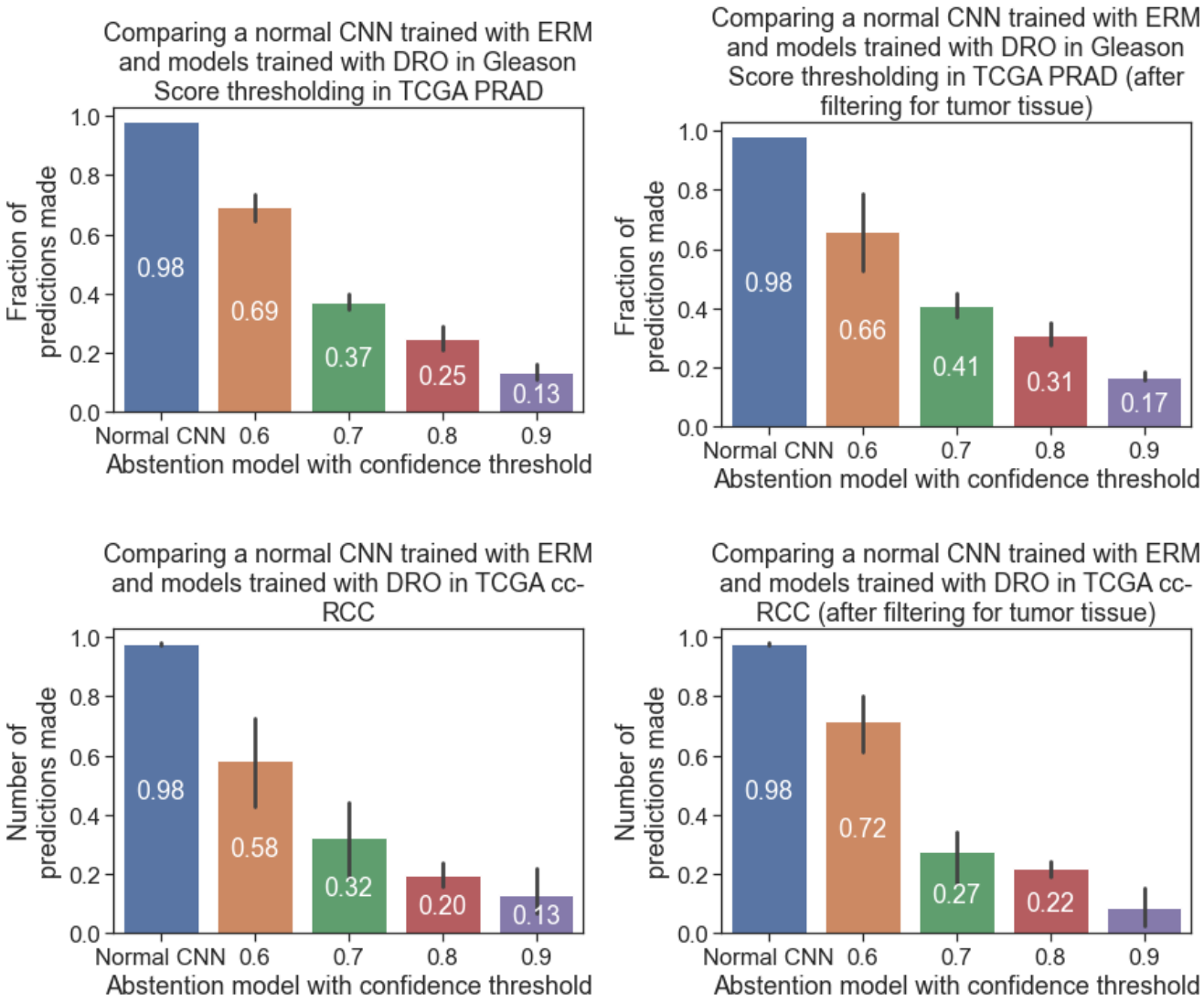


Figure 18: Showing fraction of tiles answered by abstention model in grade prediction tasks in TCGA-PRAD and ccRCC

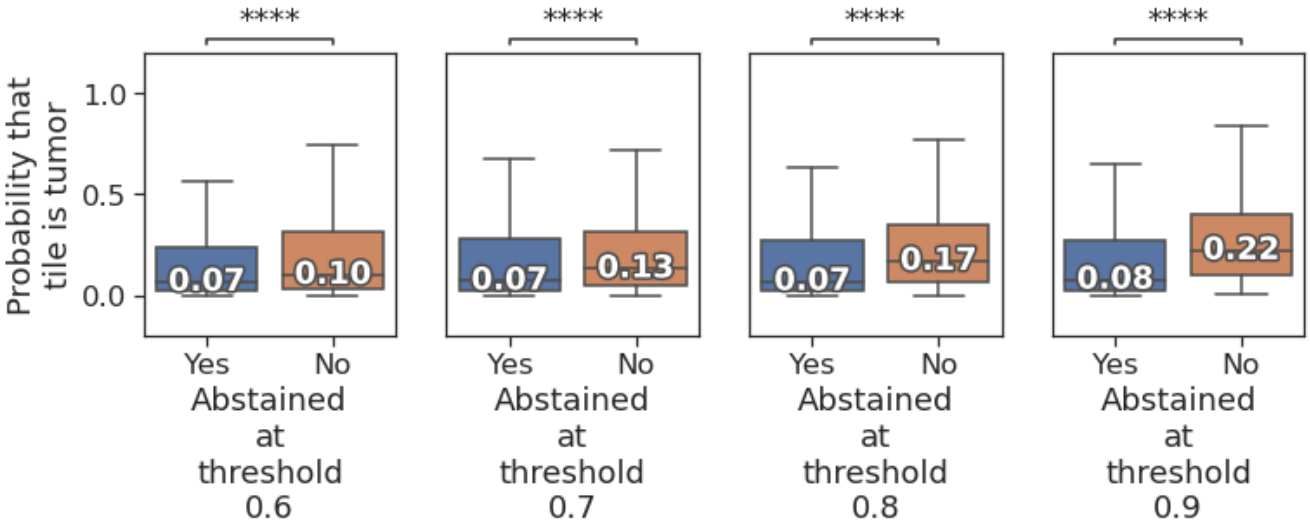


Figure 19: We explored the distribution of predicted softmax outputs among the tiles that the model was abstaining on in TCGA-PRAD. We found that the group-DRO model abstained more on tiles that did not contain tumor tissue (p values $< 10^{-4}$ using a two-sided Mann-Whitney test).

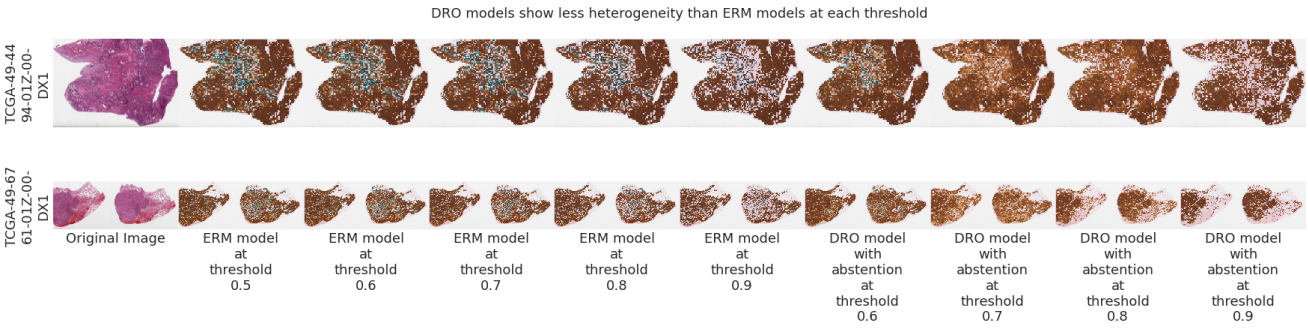


Figure 20: Reduced heterogeneity in a model trained using group-DRO for tumor versus normal identification in LUAD. Brown indicates patches that were predicted as tumor, blue indicates patches that were predicted as normal, surrounding tissue. Group-DRO models with higher confidence thresholds abstain on tiles where the features on the tile are OOD relative to the features pertinent to the WSI label (first row). Second row: group-DRO methods abstain on tiles on the right hand side of the tissue where the tissue does not bear tumor. ERM methods call non-tumor region as tumor, even at high confidence thresholds.

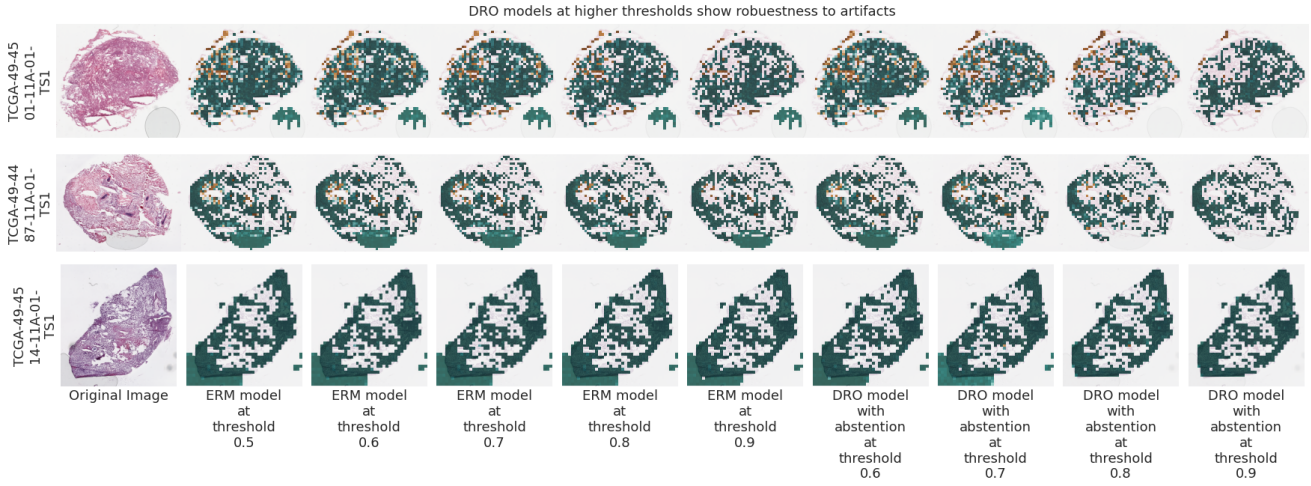


Figure 21: ERM methods predict bubble artifacts as healthy surrounding tissue. Group-DRO methods at higher confidence thresholds abstain from making predictions on artifacts.