Test Time Transform Prediction for Open Set Histopathological Image Recognition

Adrian Galdran[1,3,✉], Katherine J. Hewitt[2], Narmin L. Ghaffari[2], Jakob N. Kather[2], Gustavo Carneiro[3], and Miguel A. González Ballester[1,4]

[1] BCN Medtech, Dept. of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain, {adrian.galdran,ma.gonzalez}@upf.edu
[2] Department of Medicine III, University Hospital RWTH Aachen, Germany, {khewitt,nghaffarilal,jkather}@ukaachen.de
[3] University of Adelaide, Adelaide, Australia, gustavo.carneiro@adelaide.edu
[4] Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

Abstract. Tissue typology annotation in Whole Slide histological images is a complex and tedious, yet necessary task for the development of computational pathology models. We propose to address this problem by applying Open Set Recognition techniques to the task of jointly classifying tissue that belongs to a set of annotated classes, *e.g.* clinically relevant tissue categories, while rejecting in test time Open Set samples, *i.e.* images that belong to categories not present in the training set. To this end, we introduce a new approach for Open Set histopathological image recognition based on training a model to accurately identify image categories and simultaneously predict which data augmentation transform has been applied. In test time, we measure model confidence in predicting this transform, which we expect to be lower for images in the Open Set. We carry out comprehensive experiments in the context of colorectal cancer assessment from histological images, which provide evidence on the strengths of our approach to automatically identify samples from unknown categories. Code is released at https://github.com/agaldran/t3po .

Keywords: Histopathological Image Analysis · Open Set Recognition

1 Introduction and Related Work

Computational pathology has become fertile ground for deep learning techniques, due to several factors like the availability of large scale annotated data coupled with the increase in computational power, or the extremely time-consuming and tedious nature of visual histology examination [14,6,18]. In this context, the advanced pattern recognition capabilities of modern neural networks represents a great match for the challenges posed by digital histopathology.

However, for each new dataset that a practitioner needs to analyze, there is a requirement to annotate large whole slide images, which contain many different tissues, some of them relevant for the task at hand, whereas some others not.

In this situation, this manual annotation processing can be focused only on the labeling of the relevant tissues. Hence, an algorithm that could automatically disregard data samples outside the set of categories initially labeled by the user would be greatly useful. Another plausible scenario arises if the practitioner has labeled all of the tissue typologies that might be of interest to them, but regions of anomalous appearance show up at a later stage. These could belong to a category of clinical interest, such as rare disease signs, or simply be acquisition artifacts, but manual review of these findings could be advisable to prevent potential misdiagnosis. An obvious solution to these problems would be to flag samples in test time for which the computational model is unconfident on its prediction, assuming this could point to atypical data. Unfortunately, deep neural networks are known to be incapable of associating anomalous inputs to meaningful low confidence values [8], and there is the need for specific solutions [15].

A suitable framework to solve the above problems is based on Open Set Recognition (OSR) techniques. These are a class of learning algorithms designed to handle the presence in test time of data out of the categories on which a model was trained. This is closely related to Out-of-Distribution (OoD) detection; for the sake of clarity, we stress that here we follow the definitions given in [24], and consider OoD detection as the problem of identifying in test time samples that do not belong to the data distribution where the model was trained, without the simultaneous goal of also performing classification on data belonging to known categories. For example, a popular approach to OoD detection involves training a model to solve some pretext task for which we know the solution beforehand, *e.g.* predicting the way in which an image has been geometrically transformed [7]. The rationale is that after training, for in-distribution data the model will be able to accurately predict the applied transformation, whereas for OoD data it will most likely fail to recognize it. Other common OoD detection methods include exposing the model to outliers during training [10], observing the maximum softmax probability [9], or adding extra branches to the model to account for predictive confidence [5]. These and most other techniques have been proposed in the context of natural images, and it has been shown that they may not translate satisfactorily for OoD detection in medical imaging [1,26].

OSR and OoD detection are also related to Domain Shift/Adaptation (DS/A), the task of training a model to accurately classify data collected in a particular domain, and having the same model generalize to data with the same categories but gathered from a different domain, *e.g.* a second hospital with a different tissue preparation protocol or acquisition device [11]. In histopathological image analysis, OoD detection and DS/A have been more studied in recent years than OSR. For instance, in [22] the effect of color augmentation techniques on domain generalization in image classification on slides acquired in 9 different pathology laboratories was analyzed, and in [25] unsupervised style transfer techniques from non-medical data were applied to enhance robustness to domain shift. Stacke *et al.* also studied domain shift in histological imaging in [20,21], defining a measure in the space of learned image representations to quantify it and using it to detect data for which a model may struggle to generalize. Ensembling techniques

are also popular for uncertainty quantification in histological data, and can be put to use for identifying unreliable predictions, which can then be associated to OoD data [23]. This was proposed for instance in [16], where multi-head CNNs were shown to be superior to Monte Carlo dropout and deep ensembles for the task of flagging breast histologies containing lymph node tissue showing signs of diffuse large B-cell lymphoma, an anomaly that was not present in the training set. Self-supervision based on contrastive learning and multi-view consistency has also recently been leveraged for learning robust representations that may enable DA, namely in [4]. In [2], the authors used a similar approach to learn representations that could be useful for performing OoD detection under DS.

In this paper we introduce a novel method for OSR on histological images based on recycling information obtained during training regarding the kind of data augmentation operations that are applied online to the training data. We conjecture that for images belonging to known categories, a model trained to predict those operations will be more confident in test time, whereas for OoD data the model will be uncertain when solving this pretext task. We validate our hypothesis on two popular datasets related to colorectal cancer detection, where our experiments show that the proposed approach can accurately classify images from categories used for training and simultaneously reject clinically uninformative regions in an image without the need to manually label them.

2 Methodology

In this section, we introduce basic definitions related to the OSR setting and explain our data augmentation pipeline, which allows us to define transform prediction in a well-posed way. We then define our OSR method that jointly classifies in-distribution data and measures confidence in predicting if a data transform operation has been applied in order to declare a sample as OoD.

2.1 Open Set Recognition - Max over Softmax as a strong baseline

In an OSR scenario, we start from a labeled training set $\mathcal{C}_{\text{train}}$ with examples belonging to N known categories $\mathcal{K} = \{k_1, ..., k_N\}$, which compose the known, or *Closed Set*. However, in test time the classifier encounters samples from an *Open Set* $\mathcal{O}_{\text{test}}$ with M unknown categories $\mathcal{U} = \{u_1, ..., u_M\}$ not seen during training, *i.e.* $\mathcal{D}_{\text{test}} = \mathcal{C}_{\text{test}} \cup \mathcal{O}_{\text{test}}$. The goal of an open set classifier is to generate a reliable prediction on $\mathcal{C}_{\text{test}}$ while also rejecting samples from $\mathcal{O}_{\text{test}}$.

There exist many approaches to OSR [9]. However, it has been recently demonstrated in [24] that the simplest of all OSR methods, when optimized so as to maximize closed set accuracy with modern model architectures U_θ and training techniques, attains state-of-the-art OSR results. This baseline method consists of minimizing the cross-entropy loss between one-hot labels y and softmax probabilities $p_\theta(y|x)$ for $x \in \mathcal{C}_{\text{train}}$, and then define an OSR score as the maximum softmax probability $S(y \in \mathcal{C}_{\text{test}}|x) = \max_{y \in \mathcal{C}} p_\theta(y|x)$, assuming that U_θ will distribute probabilities with high entropy for unknown classes, resulting in a low $S(y \in \mathcal{O}_{\text{test}}|x)$ value.

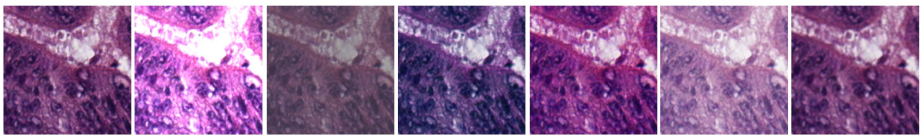

Fig. 1: Transform space $\mathcal{T}_{\text{app}}$, shown left to right: $\mathcal{T}_{\text{app}}$ ={`Identity`, `Brightness`, `Contrast`, `Saturation`, `Hue`, `Gamma`, `Sharpness`}. Our model learns to predict the applied transform during training. In test time, the model only receives un-transformed images, and we measure its confidence on transform prediction.

2.2 Decoupled Color-Appearance Data Augmentation

Data augmentation operations (image transforms in computer vision), are a conventional technique to increase generalization and reduce overfitting when training deep neural networks. Recently, learned data augmentation, which learns an optimal transformation policy from a validation set, has gained popularity, with increasingly complex techniques being proposed. However, this comes at a noticeable training overhead that has been recently shown to be indeed unnecessary [17]: the simpler scheme of randomly selecting, for each optimization step, *a single image transform* (instead of a composition of transforms) from a fixed transform space $\mathcal{T}$, with a variable strength, works remarkably well.

Inspired by [17], we define a data augmentation policy with a single transform at a time, allowing us to pose the auxiliary problem of predicting which transform has been applied to a training sample. Also, noting that geometric transforms are hardly predictable on histological data (as opposed to natural images, there is no meaningful notion of top/bottom, rotations, etc.), we decouple geometry from appearance, and define our transform space as $\mathcal{T} = \mathcal{T}_{\text{geom}} \cup \mathcal{T}_{\text{app}}$, where $\mathcal{T}_{\text{geom}}$ contains geometric transformations - rotations, shears, and translations - whereas $\mathcal{T}_{\text{app}}$ contains only color transformations. These transforms are illustrated and listed in Fig. 1; definitions can be found in the standard Python Image Library `https://github.com/python-pillow/Pillow`.

2.3 Test-Time Transform Prediction and Open Set Recognition

We formulate the training of our model as a joint optimization of two tasks. During training, we sample an image x from $\mathcal{C}_{\text{train}}$, apply a random geometric transform $\tau_g \in \mathcal{T}_{\text{geom}}$, then an appearance transform $\tau_a \in \mathcal{T}_{\text{app}}$, and pass it through a CNN U_θ, which produces an internal representation x_θ. This is then sent to the main branch f_α, a fully connected layer followed by a softmax operation, which generates a probability of x belonging to a known category from $\mathcal{K}$, but also to an auxiliary branch g_β that predicts the actual appearance transform τ_a that was applied. Among these there is the `Identity` operation, meaning that the model needs to learn what an image $x \in \mathcal{C}$ looks like.

Finally, in test time, an image x is processed by U_θ without applying any transform, resulting in a classification score $f_\alpha(U_\theta(x))$, and we define as our OSR

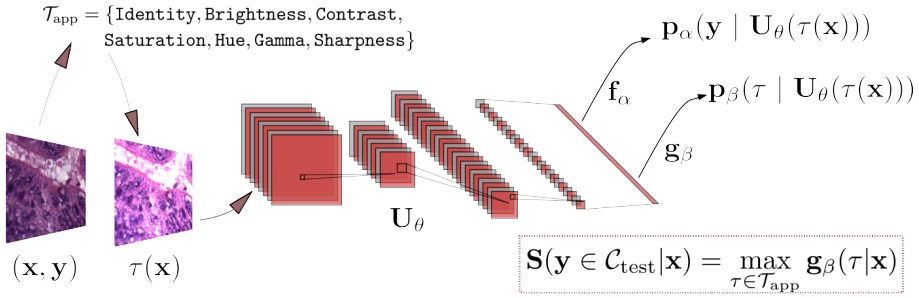

Fig. 2: Visual scheme of our OSR Test-Time Transform Prediction technique. A shared representation $U_\theta(\tau(x))$ is sent to two linear layers, f_α performs Closed Set classification and g_β predicts the applied transform τ. In test time, our OSR score is the confidence of the transform prediction branch g_β on its prediction.

score the maximum of softmax probabilities on the transform prediction task:

$$S(y \in \mathcal{C}_{\text{test}}|x) = \max_{\tau_a \in \mathcal{T}_{\text{app}}} g_\beta(\tau_a|x). \tag{1}$$

In essence, we expect the model to be more confident when predicting the transform on $\mathcal{C}_{\text{test}}$ than on $\mathcal{O}_{\text{test}}$. Let us note that we could also generate and aggregate predictions on transformed test images (Test-Time Augmentation), although this would induce an inference overhead that we prefer to avoid in this work. An illustration of the proposed OSR approach is shown in Fig. 2.

3 Experimental Analysis

In this section we introduce our experimental setup: datasets, proposed OSR tasks, and detailed performance evaluation with a discussion on the numerical differences between compared methods, as well as limitations of our technique.

3.1 Datasets and Open Set Splits

We evaluate our technique on a clinically meaningful task, namely colorectal cancer (CRC) assessment. In this context, tumor tissue composition is heterogeneous, non-stationary, and its study is key to disease prognosis [13]. A common technique for CRC monitoring is quantification of tissue configuration by histological evaluation of Hematoxylin and Eosin (H&E) stained tissue sections.

We consider two publicly available datasets[5] that enable CRC tissue characterization, referred to as Kather-5k [13] and Kather-100k [12]. Examples of images from each tissue type in these datasets are shown in Fig. 3. Specifically:

[5] Kather-5k: `http://doi.org/10.5281/zenodo.53169`
Kather-100k: `http://doi.org/10.5281/zenodo.1214456`

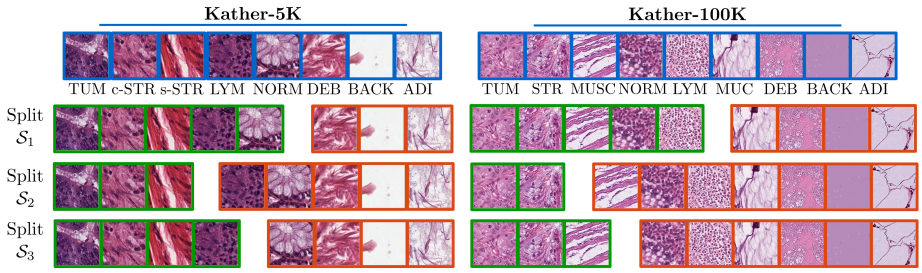

Fig. 3: Closed (green) and Open (orange) Set partitions defined on the two considered datasets, please see the text for acronym definitions and motivation.

- Kather-5k contains 5,000 image patches extracted from 10 tumoral tissue slides with 8 classes: *tumor* epithelium (TUM), *simple stroma* (s-STR), homogeneous tissue, with tumoral and extra-tumoral stroma, but also muscle, *complex stroma* (c-STR), which may contain some immune cells, *immune cell conglomerates* (LYM), *debris* (DEB), which includes necrosis or mucus, *normal* colon mucosa (NORM), *adipose* tissue (ADI), and *background* tissue (BACK). Data is balanced, with 150×150 pixel size and $74\mu m$/px resolution.
- Kather-100k is larger, with 100,000 image patches extracted from 86 CRC tissue slides, originally used for overall CRC survival prediction. It has 9 different tissue types: *tumour* epithelium (TUM), cancer-related *stroma* (t-STR), smooth *muscle* (MUS), *immune cell conglomerates* (LYM), *debris/necrosis* (DEB), *mucus* (MUC), *normal* colon mucosa (NORM), *adipose* tissue (ADI), and *background* (BACK). Note some subtle differences with Kather-5k: the category debris is split into debris/necrosis and mucus; also, stroma is not divided into simple an complex: only tumoral stroma is considered, whereas muscle is a new category. Data is approximately balanced and color-normalized, with images of size 224×224 and 122μ/px resolution.

Next, following expert pathologist's advice, we define several Closed/Open set splits in each dataset, illustrated in Fig. 3. We first design a split $\mathcal{S}_1$ mimicking the hypothetical situation in which a practitioner decides to label only clinically informative tissue regions, and leaves uninformative regions unlabeled, expecting the OSR model to automatically identify it as part of the Open Set $\mathcal{O}$, while still achieving high accuracy in the Closed Set $\mathcal{C}$. Note that this is not a trivial task, since necrotic tissue, part of the debris category, can be infiltrated by inflammatory cells, and therefore the lymphocytes class acts as a confounder in the closed set. To supplement our experimental analysis and understand the weaknesses of OSR systems in this application, we also define two other splits $\mathcal{S}_2$ and $\mathcal{S}_3$. In $\mathcal{S}_2$ we aim at analyzing if an OSR classifier can classify tumoral regions while rejecting healthy tissue as well as uninformative samples in test time, so we include tumor and stroma patches in the closed set. Note that for both datasets there are now some confounders in the Open Set. Namely, in the Kather-5k dataset complex stroma images may include some immune cells, but the immune-cell conglomerate category is in the Open Set of $\mathcal{S}_2$. On the other hand, in the Kather-100k dataset stroma images do not include immune cells,

Table 1: Performance averaged over 10 training runs of our approach and other OSR techniques on several Open/Closed splits of the Kather-5k dataset. Best performance is underlined, results within its confidence interval are bold.

	Split 0		**Split 1**		**Split 2**	
	ACC	AUC	ACC	AUC	ACC	AUC
CE+	**93.03**	91.66	**94.27**	82.51	92.88	90.02
ARPL	**92.84**	88.96	92.51	80.28	**93.39**	82.39
MC-Dropout	**93.16**	91.52	**94.02**	82.19	92.80	85.45
T3PO (Ours)	**92.54**	**93.55**	**94.27**	**84.73**	91.80	**91.24**

but the stroma and the muscle categories share a fibrous aspect, and muscle images belong to the Open Set. For comparison purposes, in the last split $\mathcal{S}_3$ we move the lymphocites class to the Closed Set, which should result in an easier OSR task at the expense of a more challenging Closed Set classification task, since now $\mathcal{C}$ contains two similar classes.

3.2 Implementation Details and Performance Evaluation

We compare Test-Time Transform Prediction (T3PO) with the state-of-the-art ARPL technique [3], and CE+, the strong baseline proposed recently in [24], which consists of maximizing the Closed Set accuracy of the classifier and, instead of taking the maximum over the softmax probabilities as the OSR score, use the maximum over the logits, *i.e.* pre-softmax activations of the network. We also adopt the MC-Dropout baseline (applying dropout multiple times ($n = 32$) in test time and collecting the entropy of the resulting set of softmax probabilities as the OSR score), popular in medical image analysis problems [16].

Since previous work has shown that relatively small architectures are capable of achieving high accuracy on the two considered datasets, for the sake of quick experimentation we always train a MobileNet V2 network as our backbone [19], starting from ImageNet weights. Following [13], we split the data into 70% for training, 15% for validation and early-stopping, and 15% for testing. In all cases we train with a cyclical learning rate starting at $l = 0.01$ and a batch-size of 128, for 200 epochs in the Kather-5k dataset. Due to the larger amount of training samples, we only train for 20 epochs in the Kather-100k dataset, which is enough for all models to converge. We use the Adam optimizer, monitor the Closed Set accuracy during training, and keep the highest-performing checkpoint. After training, we collect model accuracy on the Closed test set, and OSR scores in the Closed and Open test sets. We perform ten training runs per split and report mean Closed Set accuracy and Closed/Open AUC.

Table 2: Performance averaged over 10 training runs of our approach and other OSR techniques on several Open/Closed splits of the Kather-100k dataset. Best performance is underlined, results within its confidence interval are bold.

	Split 0		**Split 1**		**Split 2**	
	ACC	AUC	ACC	AUC	ACC	AUC
CE+	**99.54**	96.50	**<u>99.69</u>**	**84.59**	**<u>99.62</u>**	82.96
ARPL	98.88	91.76	**99.33**	78.00	98.98	79.96
MC-Dropout	**<u>99.57</u>**	96.23	**99.64**	**<u>84.93</u>**	**99.58**	84.52
T3PO (Ours)	**99.46**	**<u>96.57</u>**	**99.66**	83.32	**99.56**	**<u>92.42</u>**

3.3 Results and Discussion

Tables 1 and 2 show the performance of the considered OSR techniques on the Kather-5k and Kather-100k datasets respectively. The first split, which in both cases sets out the task of classifying clinically relevant tissue categories, is successfully solved to a high accuracy by all approaches, with no statistically significant difference between our proposed T3PO and the top performer MC-Dropout. If we analyze the ability of each method to reject uninteresting data in test time, however, we see that T3PO outperforms the other techniques, by a relatively wide margin in the Kather-5k dataset, in terms of Closed/Open Set AUC, indicating that our method can better identify Open Set data in this case.

The second and third split in the Kather-5k dataset illustrate a limitation of OSR approaches. In the second split, the Open Set contains images from the immune cell category, and immune cells are also present on some images from the complex stroma class, which belongs to the Closed Set. This results in a generally lower AUC for all methods, although T3PO continues to outperform other techniques. In addition, when we move the immune-cell category from the Open to the Closed Set, we see a noticeable increase in AUC for all methods (and a decrease in accuracy, since two visually similar categories are now in the Closed Set), with T3PO still significantly attaining top performance in Open Set recognition. It should be noted that this is achieved at the cost of a modest, but statistically significant decrease in Closed Set accuracy for the third split.

Lastly, the second and third split in the Kather-100k dataset also show a similar phenomenon. In this case the muscle class belonging to the Open Set in the second split drives the Closed/Open AUC down for all methods, since it is confounded with the stroma category from the Closed set, and we see that T3PO is among the worst techniques now. However, when we move the muscle class into the Closed Set, T3PO increases the AUC by more than 9 points, outperforming all other methods, and losing very little accuracy.

3.4 Conclusion and Future Work

We have illustrated how a clinically meaningful task, disregard irrelevant image regions from histological slides without explicitly training a model to discriminate them, can be addressed with OSR techniques. We have also introduced T3PO, a new OSR method that outperforms several recent approaches in most cases. We have also discussed its limitations, namely T3PO consists of the identification of global image transformations in test time, thereby relying on low-level image characteristics like color and aspect, but not taking full advantage of other semantic cues, which may result in sub-optimal performance. We leave the integration of the knowledge of image content into our approach for future work.

Acknowledgments

This work was partially supported by a Marie Skłodowska-Curie Global Fellowship (No 892297) and by Australian Research Council grants (DP180103232 and FT190100525).

References

1. Berger, C., Paschali, M., Glocker, B., Kamnitsas, K.: Confidence-Based Out-of-Distribution Detection: A Comparative Study and Analysis. In: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis. pp. 122–132. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-87735-4-12
2. Bozorgtabar, B., Vray, G., Mahapatra, D., Thiran, J.P.: SOoD: Self-Supervised Out-of-Distribution Detection Under Domain Shift for Multi-Class Colorectal Cancer Tissue Types. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 3317–3326 (Oct 2021). https://doi.org/10.1109/ICCVW54120.2021.00371, iSSN: 2473-9944
3. Chen, G., Peng, P., Wang, X., Tian, Y.: Adversarial Reciprocal Points Learning for Open Set Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2021). https://doi.org/10.1109/TPAMI.2021.3106743, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence
4. Ciga, O., Xu, T., Martel, A.L.: Self supervised contrastive learning for digital histopathology. Machine Learning with Applications **7**, 100198 (Mar 2022). https://doi.org/10.1016/j.mlwa.2021.100198
5. DeVries, T., Taylor, G.W.: Learning Confidence for Out-of-Distribution Detection in Neural Networks. arXiv:1802.04865 [cs, stat] (Feb 2018), `http://arxiv.org/abs/1802.04865`, arXiv: 1802.04865
6. Echle, A., Rindtorff, N.T., Brinker, T.J., Luedde, T., Pearson, A.T., Kather, J.N.: Deep learning in cancer pathology: a new generation of clinical biomarkers. British Journal of Cancer **124**(4), 686–696 (Feb 2021). https://doi.org/10.1038/s41416-020-01122-x, number: 4 Publisher: Nature Publishing Group
7. Golan, I., El-Yaniv, R.: Deep Anomaly Detection Using Geometric Transformations. In: Advances in Neural Information Processing Systems. vol. 31 (2018)

8. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On Calibration of Modern Neural Networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1321–1330. PMLR (Aug 2017)
9. Hendrycks, D., Gimpel, K.: A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In: International Conference on Learning Representations. OpenReview.net (2017)
10. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep Anomaly Detection with Outlier Exposure. In: International Conference on Learning Representations (2019)
11. Howard, F.M., Dolezal, J., Kochanny, S., Schulte, J., Chen, H., Heij, L., Huo, D., Nanda, R., Olopade, O.I., Kather, J.N., Cipriani, N., Grossman, R.L., Pearson, A.T.: The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nature Communications **12**(1), 4423 (Jul 2021). https://doi.org/10.1038/s41467-021-24698-1
12. Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., Jansen, L., Reyes-Aldasoro, C.C., Zörnig, I., Jäger, D., Brenner, H., Chang-Claude, J., Hoffmeister, M., Halama, N.: Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLOS Medicine **16**(1), e1002730 (Jan 2019). https://doi.org/10.1371/journal.pmed.1002730
13. Kather, J.N., Weis, C.A., Bianconi, F., Melchers, S.M., Schad, L.R., Gaiser, T., Marx, A., Zöllner, F.G.: Multi-class texture analysis in colorectal cancer histology. Scientific Reports **6**(1), 27988 (Jun 2016). https://doi.org/10.1038/srep27988
14. van der Laak, J., Litjens, G., Ciompi, F.: Deep learning in histopathology: the path to the clinic. Nature Medicine **27**(5), 775–784 (May 2021). https://doi.org/10.1038/s41591-021-01343-4, number: 5 Publisher: Nature Publishing Group
15. Lee, K., Lee, H., Lee, K., Shin, J.: Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples. In: International Conference on Learning Representations (2018)
16. Linmans, J., Laak, J.v.d., Litjens, G.: Efficient Out-of-Distribution Detection in Digital Pathology Using Multi-Head Convolutional Neural Networks. In: Medical Imaging with Deep Learning (2020)
17. Müller, S.G., Hutter, F.: TrivialAugment: Tuning-Free Yet State-of-the-Art Data Augmentation. pp. 774–782 (2021)
18. Picon, A., Terradillos, E., Sánchez-Peralta, L.F., Mattana, S., Cicchi, R., Blover, B.J., Arbide, N., Velasco, J., Etzezarraga, M.C., Pavone, F.S., Garrote, E., Saratxaga, C.L.: Novel Pixelwise Co-Registered Hematoxylin-Eosin and Multiphoton Microscopy Image Dataset for Human Colon Lesion Diagnosis. Journal of Pathology Informatics **13**, 100012 (Jan 2022). https://doi.org/10.1016/j.jpi.2022.100012
19. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: Inverted Residuals and Linear Bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4510–4520 (Jun 2018)
20. Stacke, K., Eilertsen, G., Unger, J., Lundström, C.: A closer look at domain shift for deep learning in histopathology. In: MICCAI COMPAY Workshop (Jul 2019)
21. Stacke, K., Eilertsen, G., Unger, J., Lundström, C.: Measuring Domain Shift for Deep Learning in Histopathology. IEEE Journal of Biomedical and Health Informatics **25**(2), 325–336 (Feb 2021). https://doi.org/10.1109/JBHI.2020.3032060, conference Name: IEEE Journal of Biomedical and Health Informatics

22. Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.M., Ciompi, F., van der Laak, J.: Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Medical Image Analysis **58**, 101544 (Dec 2019). https://doi.org/10.1016/j.media.2019.101544
23. Thagaard, J., Hauberg, S., van der Vegt, B., Ebstrup, T., Hansen, J.D., Dahl, A.B.: Can You Trust Predictive Uncertainty Under Real Dataset Shifts in Digital Pathology? In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. pp. 824–833. Lecture Notes in Computer Science, Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8-80
24. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Open-Set Recognition: A Good Closed-Set Classifier is All You Need (Sep 2021)
25. Yamashita, R., Long, J., Banda, S., Shen, J., Rubin, D.L.: Learning Domain-Agnostic Visual Representation for Computational Pathology Using Medically-Irrelevant Style Transfer Augmentation. IEEE Transactions on Medical Imaging **40**(12), 3945–3954 (Dec 2021). https://doi.org/10.1109/TMI.2021.3101985, conference Name: IEEE Transactions on Medical Imaging
26. Zhang, O., Delbrouck, J.B., Rubin, D.L.: Out of Distribution Detection for Medical Images. In: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis. pp. 102–111. Lecture Notes in Computer Science, Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-87735-4-10