Can you trust predictive uncertainty under real dataset shifts in digital pathology?

Thagaard, Jeppe; Hauberg, Søren; van der Vegt, Bert ; Ebstrup, Thomas; Hansen, Johan D. ; Dahl, Anders Bjorholm

Published in:
Proceedings of 23rd International Conference on Medical Image Computing and Computer Assisted Intervention

Link to article, DOI:
10.1007/978-3-030-59710-8_80

Publication date:
2020

Document Version
Peer reviewed version

Can you trust predictive uncertainty under real dataset shifts in digital pathology?

Jeppe Thagaard[1,2], Søren Hauberg[1], Bert van der Vegt[3], Thomas Ebstrup[2], Johan D. Hansen[2] & Anders B. Dahl[1]

[1] Technical Universtity of Denmark, Lyngby, Denmark
[2] Visiopharm A/S, Hørsholm, Denmark
jept@dtu.dk, jth@visiopharm.com
[3] University Medical Center Groningen, Groningen, The Netherlands

Abstract. Deep learning-based algorithms have shown great promise for assisting pathologists in detecting lymph node metastases when evaluated based on their predictive accuracy. However, for clinical adoption, we need to know what happens when the test set dramatically changes from the training distribution. In such settings, we should estimate the uncertainty of the predictions, so we know when to trust the model (and when not to). Here, we i) investigate current popular methods for improving the calibration of predictive uncertainty, and ii) compare the performance and calibration of the methods under clinically relevant in-distribution dataset shifts. Furthermore, we iii) evaluate their performance on the task of out-of-distribution detection of a different histological cancer type not seen during training. Of the investigated methods, we show that deep ensembles are more robust in respect of both performance and calibration for in-distribution dataset shifts and allows us to better detect incorrect predictions. Our results also demonstrate that current methods for uncertainty quantification are not necessarily able to detect all dataset shifts, and we emphasize the importance of monitoring and controlling the input distribution when deploying deep learning for digital pathology.

Keywords: Deep learning · Digital pathology · Predictive uncertainty

1 Introduction

Motivated by the predictive performance of deep learning (DL) in research [3, 21] and grand challenges [2], clinical-grade DL-tools for assisting pathologists in detection of lymph node metastases are now being developed. In clinical settings where algorithms can potentially affect medical decisions, it is crucial to know how well-calibrated the underlying model is, such that the model gives a reliable estimate of the quality of the predictions. However, there exists only limited research [4, 20, 22] on how different distributional shifts in pathology affect the accuracy of DL-based algorithms, and these do not consider predictive uncertainty. Dataset shifts are especially relevant in pathology as pre-analytical steps

can introduce large variability, and the spectrum of the target indication of an algorithm can also be broad. This makes it difficult to include the whole spectrum within the training set. Rare incidental findings, which are clinically relevant, may also be missed by an algorithm because they are outside the distribution of the training set.

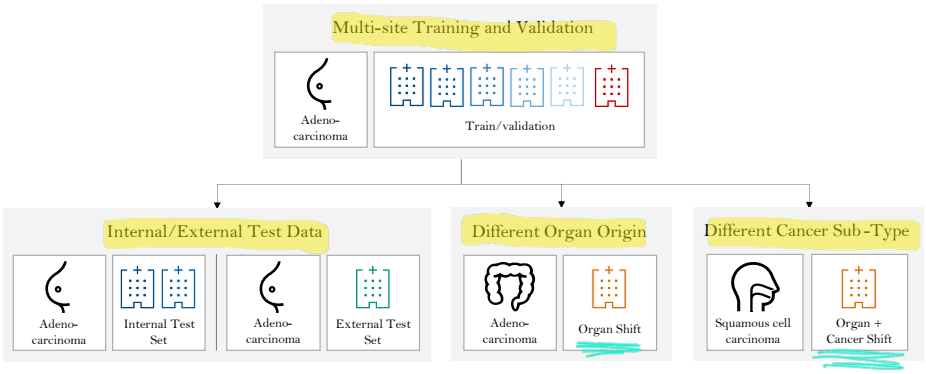

Fig. 1: Overview of experimental setup. Slides from 6 different sites are used as development data ($\mathcal{D}_{train}$ and $\mathcal{D}_{val}$), where blue (5 sites) represents CAM16-train and CAM17-train and red (one site) is DATASET2. CAM16-test defines the internal test set ($\mathcal{D}_{test,int.}$) as the 2 sites are also used as development data. DATASET3 (green) is denoted as the external test set ($\mathcal{D}_{test,ext.}$) as this site is not included in the development data. Slides from DATASET4 and DATASET5 (orange) with colon adenocarcinoma ($\mathcal{D}_{colon}$) and head and neck squamous cell carcinoma ($\mathcal{D}_{SCC}$) are used to test on different organ origin and different cancer sub-type than the original target task of detecting adenocarcinoma from breast cancer.

Our contribution is a thorough investigation of several state-of-the-art methods' ability to quantify uncertainty while keeping high accuracy. We focus on the problem of detecting cancerous tissue in digital pathology, specifically for the task of detecting lymph node metastases. This has not been covered in previous investigations such as [9, 17], because the appearance and variation resulting from distributional shifts of histopathology images is very different from that of natural images. Therefore, we i) extend our evaluation to a unique real-world pathology setting with a multi-hospital single indication training set and perform an extensive evaluation on both internal and external test sets and clinically plausible distributional shifts. We ii) compare the methods in terms of performance and calibration in addition to iii) how accurate their predictive uncertainty can detect both incorrect predictions and out-of-distribution (OOD) inputs.

1.1 Related work

Multiple popular methods have been proposed for quantifying predictive uncertainty for better calibration and robustness under distributional shifts and OOD inputs in deep neural networks (DNNs). Deep ensemble [13] is arguably the simplest method where multiple networks are trained individually and their predictions are averaged during inference. Monte Carlo Dropout (MC-Dropout) [6] is an approximate Bayesian method that uses dropout [19] during multiple forward passes during inference. Temperature scaling [7] is different as it serves as a post-processing method that learns a scaling parameter on a validation set but its performance has shown to be limited under distributional shifts [17]. Mixup [25] combines random pairs of images and their labels during training, originally aimed at increased performance but it has recently shown to improve the calibration of DNNs [23]. All methods have their advantages and limitations with regard to their complexity during training or inference.

Table 1: Details on data. * and ** denote adenocarcinoma and SCC, respectively. † [14], ‡ [3]

Dataset	Purpose	No. of slides	Site
CAM16-train	Development ($\mathcal{D}_{train}$,$\mathcal{D}_{val}$)	270 (160 normal, 110 tumor*)	2 hospitals†
CAM16-test	Evaluation ($\mathcal{D}_{test,int.}$)	129 (80 normal, 49 tumor*)	2 hospitals†
CAM17-train	Development ($\mathcal{D}_{train}$,$\mathcal{D}_{val}$)	46 (0 normal, 46 tumor*)	5 hospitals‡
DATASET2	Development ($\mathcal{D}_{train}$,$\mathcal{D}_{val}$)	56 (41 normal, 15 tumor*)	Hospital-A
DATASET3	Evaluation ($\mathcal{D}_{test,ext.}$)	135 (67 normal, 68 tumor*)	Hospital-B
DATASET4	Evaluation ($\mathcal{D}_{colon}$)	81 (43 normal, 38 tumor*)	Hospital-C
DATASET5	Evaluation ($\mathcal{D}_{SCC}$)	60 (40 normal, 20 tumor**)	Hospital-C

2 Methods

2.1 Experimental setup

To study a relevant application in pathology, we define the primary target task as detection of adenocarcinoma in hematoxylin and eosin (H&E) lymph node sections from breast cancer. To enable the development, we obtain datasets from public [2, 3, 14] and non-public sources (see details in Table 1) and evaluate both predictive accuracy and uncertainty using relevant metrics (see below).

In-distribution shift To evaluate whether we can trust the predictions on images not derived from the hospitals used in the development, we use DATASET3 as an external test set ($\mathcal{D}_{test,ext.}$) and CAM16$_{test}$ internal test set ($\mathcal{D}_{test,int.}$). The methods are evaluated based on their ability to generalize in terms of predictive accuracy and uncertainty.

As the same cancer sub-type can originate from different organs and metastasize to lymph nodes regardless of origin, we investigate the methods' ability to generalize to other organs than included in the training set. To enable this, we collect lymph node sections with adenocarcinoma from colon cancer ($\mathcal{D}_{colon}$).

Misclassification detection The ability to indicate incorrect classifications is attractive from a clinical automation perspective, so pathologists can better interfere and assess results when needed, especially when the input distribution change from the intended indication. It is easy to formulate as a binary classification problem using only the uncertainty as the prediction score, hence it is a popular downstream task to evaluate predictive uncertainty [10]. We hypothesize that current methods are better at detecting incorrect predictions when the dataset is more similar to the training distribution. To test the hypothesis, we use $\mathcal{D}_{test,int.}$, $\mathcal{D}_{test,ext.}$ and $\mathcal{D}_{colon}$ to assess the performance of the binary classification (correct vs. incorrect) on each dataset.

Out-out-distribution shift When pathologists assess lymph node sections for metastases, they are also aware of other clinically relevant abnormalities than the primary task. To mimic this setting, we collect slides that contain another histology sub-type (squamous cell carcinoma (SCC)) from head and neck cancer ($\mathcal{D}_{SCC}$), which includes both well- and un-differentiated SCCs. Since SCCs, especially well-differentiated cases, are morphological different than adenocarcinoma, we consider $\mathcal{D}_{SCC}$ a realistic out-of-distribution dataset because it contains unseen abnormalities from the same domain as the training set.

Here, our evaluation is two-fold: generalization to another cancer sub-type and the ability to detect novel classes using its predictive uncertainty. To achieve the latter, we denote all tumor regions from $\mathcal{D}_{SCC}$ as $\mathcal{D}_{out}$ and the in-distribution $\mathcal{D}_{test,ext.}$ as $\mathcal{D}_{in}$. We then compare each method to discriminate between $\mathcal{D}_{out}$ and $\mathcal{D}_{in}$.

Since poorly differentiated SCC can look morphologically similar to adenocarcinoma, we also take a subset of $\mathcal{D}_{SCC}$ diagnosed as well-differentiated SCC ($N = 5$) and treat only samples from these as OOD inputs in a final experiment.

Reference standard Similar to the Camelyon dataset, all ground truth annotations on the non-public datasets were carefully prepared under the supervision of expert pathologists with additional slides stained with cytokeratin immunohistochemistry (IHC). All work related to the non-public datasets was approved by their institutional review board.

2.2 Evaluation metrics

We employ *Accuracy*, *Area Under the Receiver Operating Characteristics curve* (AUROC) and *Precision-Recall curve* (AUPR) to report classification performance (normal vs. tumor). As suggested by Guo et al. [7], we use the *Expected*

Calibration Error ECE [16] to measure the calibration for each model. First, we compute the confidence of each of N observation denoted $p(\hat{y}_n)$, and bin these into H bins. We then calculate the ECE by comparing the content of each bin to its average accuracy. Let B_h be the set of indices for bin h. We calculate the bin accuracy

$$\text{acc}(B_h) = |B_h|^{-1} \sum_{n \in B_h} \delta(\hat{y}_n - y_n^*) \tag{1}$$

and the bin confidence

$$\text{conf}(B_h) = |B_h|^{-1} \sum_{n \in B_h} p_n(\hat{y}) \ . \tag{2}$$

Then we get

$$\text{ECE} = \frac{1}{N} \sum_{h=1}^{H} |B_h| \cdot |\text{acc}(B_h) - \text{conf}(B_h)| \tag{3}$$

$$= \frac{1}{N} \sum_{h=1}^{H} \left| \sum_{n \in B_h} p_n(y) - \sum_{n \in B_h} \delta(\hat{y}_n - y_n^*) \right| \tag{4}$$

where $\delta(x) = 1$ if $x = 0$ or $\delta(x) = 0$ if $x \neq 0$, and y_n^* is the true label.

For misclassification and OOD detection, we use also AUROC and AUPR but on the classification performance of correct vs. incorrect and in- vs. out-of-distribution, respectively. We use *False Positive Rate at 95% True Positive Rate* (FPR95) to compare method at a certain operating point. As noted by [1], these metrics are more reliable to compare for OOD detection as the task remains the same regardless of method.

2.3 Overview of methods

We focus on methods that model $p(y|x)$ as these are the most popular in medical image analysis [3, 15] and are known to scale well [12, 13]. As a baseline, we use the softmax of a standard DNN to obtain posterior probabilities. For all methods, we obtain the prediction as $\hat{y} = \arg\max_y p(y|x, \theta)$ and the confidence as the maximum softmax probability $p(\hat{y}) = \max_y p(y|x, \theta)$.

MC-Dropout We train using dropout [19] with rate p and apply L forward passes during inference with dropout enabled as described in Gal et al. [6].

Deep ensemble We train M standard DNNs independently of each other following [13] and combine the predictions as

$$p(y = k|x, \theta) = \frac{1}{M} \sum_{m=1}^{M} p_m(y = k|x, \theta_m) \tag{5}$$

Mixup Recently proposed as a simple method by [25] for training better DNNs where two random input samples (x_i, x_j) and their corresponding labels (y_i, y_j) are combined using:

$$\begin{aligned} \tilde{x} &= \lambda x_i + (1-\lambda)x_j \\ \tilde{y} &= \lambda y_i + (1-\lambda)y_j \end{aligned} \tag{6}$$

where $\lambda \in [0, 1]$ determines the mixing ratio of the linear interpolation. λ is drawn from a symmetric Beta distribution Beta(α, α), where α controls the strength of the input interpolation and the label smoothing. We train a DNN with mixup using standard cross-entropy calculated on the soft-labels instead of the hard labels. We refer to [25] for the full details on mixup.

Table 2: Evaluation of predictive performance. *$\alpha = 0.3$

	$\mathcal{D}_{test,int.}$			$\mathcal{D}_{test,ext.}$			$\mathcal{D}_{colon}$		
	Acc.	AUROC	AUPR	Acc.	AUROC	AUPR	Acc.	AUROC	AUPR
Baseline	90.5	96.5	95.1	**94.3**	97.9	94.3	**79.0**	90.7	92.8
Ensemble	90.1	**97.3**	**95.9**	**94.3**	**98.1**	**96.8**	78.1	**92.3**	**94.2**
MC-Dropout	**91.0**	97.0	95.7	93.8	97.7	96.2	78.0	90.9	93.4
Mixup*	86.5	95.6	94.2	93.4	97.1	94.6	75.8	91.0	92.6

2.4 Implementation and training details

We perform a train/validation split on the development dataset and use these to train and select hyper-parameters for all methods. All datasets are sampled in patches (512×512 pixels) at 20× magnification with 50% (strided) and 150% (overlapping) sampling fraction for normal and tumor, respectively. We employ a ResNet-50 [8] architecture as the backbone for all methods because there are negligible changes between different image classifiers [9]. We use $M = 5$ to create the ensemble as reported by [17] to be sufficient. For MC-dropout, initial experiments of different implementation variations showed no performance differences. Hence, we add a dropout before the logit layer similar to [12] with $p = 0.5$ and use $L = 50$. All models are trained for 15 epochs with ADAM [11] ($\beta = (0.9, 0.999)$) with weight decay (0.0005) using a mini-batch size of 16. We use an initial learning rate of 0.01 and drop it with factor 10 every 5th epoch for all methods except mixup which required a lower initial learning rate of 0.001 to converge. For mixup, we experimented with $\alpha \in [0.1, 0.3, 0.5, 1.0]$ and we report results with $\alpha = 0.3$ as this performed best on $\mathcal{D}_{val}$. In all experiments, we apply data augmentation similar to [15] and use Pytorch [18] and Pytorch-Lightning [5].

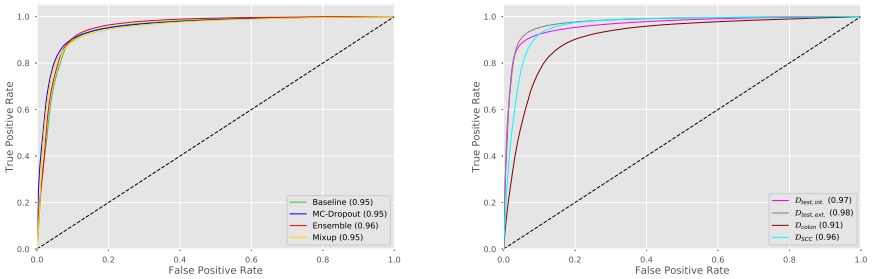

Fig. 2: ROC-curves for predictive performance. Left shows each methods with ROC curves averaged across all datasets. Right shows each dataset with ROC curves averaged across all methods.

3 Results

3.1 Evaluating predictive performance under dataset shifts

First, we evaluate the predictive performance on the primary task of detecting adenocarcinoma in lymph node sections. We summarize the results in Table 2, and the ROC-curves for all methods and dataset shifts are shown in Figure 2. The results show that all methods can archive high predictive performance on both the internal and external test sets. All methods perform significantly worse when evaluated on the colon dataset $\mathcal{D}_{colon}$ with mixup performing worst. Interestingly, all methods have higher AUROC on $\mathcal{D}_{SCC}$ (see Table 4) compared to $\mathcal{D}_{colon}$ even though the cancer sub-type is histological different, especially in the well-differentiated cases. In general, deep ensemble slightly outperforms all other methods on threshold independent metrics like AUROC and AUPR.

3.2 Evaluating predictive uncertainty under dataset shifts

We present results of calibration and detection of incorrect classified examples together in Table 3. In terms of ECE, deep ensemble and mixup improve calibration compared to the baseline method, whereas MC-dropout performs worse for the external and colon dataset. When using each method's predictive uncertainty to detect misclassifications on the test set, deep ensemble and MC-dropout have higher AUROC and AUPR on all three datasets than baseline and mixup. However, the quality of the predictive uncertainty for decreases slightly when dataset shift increases.

3.3 Evaluating on different cancer sub-type

The left part of Table 4 shows the performance on $\mathcal{D}_{SCC}$, while the right side summarizes the result of the OOD experiment. All methods show strong predic-

Table 3: Evaluation of calibration and misclassification detection. *$\alpha = 0.3$

	$\mathcal{D}_{test,int.}$			$\mathcal{D}_{test,ext.}$			$\mathcal{D}_{colon}$		
	ECE	AUROC	AUPR	ECE	AUROC	AUPR	ECE	AUROC	AUPR
Baseline	4.9	82.6	35.7	2.1	77.7	28.6	11.8	76.7	42.0
Ensemble	**2.1**	83.9	35.6	**0.6**	**82.3**	**30.2**	**7.5**	**78.6**	**44.5**
MC-Dropout	4.6	**84.0**	35.3	2.6	79.8	29.7	13.3	77.2	43.5
Mixup*	4.2	79.1	**36.5**	0.9	80.9	29.3	9.7	71.5	41.4

tive accuracy, but fail to recognize SCC as an unseen class. Here, both ensemble and mixup outperform the baseline and MC-dropout methods.

Table 4: Evaluation of performance and OOD detection on $\mathcal{D}_{SCC}$. *$\alpha = 0.3$

	Performance			OOD			ODD (only well-diff.)		
	Acc.	AUROC	AUPR	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95
Baseline	89.3	95.4	88.4	64.1	37.3	97.6	70.6	5.2	90.9
Ensemble	**89.7**	**96.3**	**91.8**	73.2	46.2	92.6	81.6	7.4	71.1
MC-Dropout	89.0	95.9	91.5	59.8	35.6	99.3	67.5	4.7	84.8
Mixup*	87.5	95.8	89.2	**86.3**	**53.6**	**47.5**	**86.5**	**8.1**	**44.6**

4 Discussion and Conclusion

We have evaluated current popular methods for predictive uncertainty on clinically relevant dataset shifts for the detection of lymph node metastases in pathology slides. All methods can generalize predictive accuracy from the internal test set to the external dataset while maintaining the quality of the predictive uncertainty. When applied to another organ, all investigated methods show both decreased performance and increased overconfidence. We have shown similar behavior when evaluated on the different cancer sub-type even-though the performance decrease was smaller than under organ shift.

As site-specific variations such as sectioning, staining and scanning variability are present in the experimental internal and external setup, we have shown that current methods are able to generalize to these sources of variability. We leave it to future work to quantify how site-specific pre-analytical variations affect the current methods as it requires a more controlled data acquisition scheme.

Our experiments show minimal benefits of MC-Dropout compared to the baseline method, and it can hurt the calibration performance on all dataset shifts. We contribute this to MC-Dropout being a too weak ensemble to achieve the same effect as a true ensemble. In general, deep ensemble increases predictive performance but also shows robustness in calibration under distributional shifts. It also displays decent capability in detecting incorrect predictions, but none of the methods are sufficient on this task. Based on the results and its simplicity,

deep ensemble is an attractive method for predictive uncertainty but it comes with a computational overhead during both training and inference. Here, mixup might seem to be a cheaper alternative as our results show better calibration than baseline and MC-Dropout with a slight decrease in performance. We leave it to future work to investigate effects of different implementation of MC-Dropout and mixup extensions such as [24].

The ODD experiments indicate that adenocarcinoma and SCC, especially moderate and undifferentiated, are too similar in their morphological patterns to be treated as OOD. However, when we only assume well-differentiated SCC as an unseen class, ensemble and mixup are better to indicate the dataset shift without being sufficient for ODD detection.

Based on our results, we recommend that deep learning-based algorithms are ready for clinical implementation with reliable uncertainty estimates if used within the indication and organ included in the training set, but one should not expect current methods to alarm novel abnormalities.

Acknowledgement The work was mainly supported by Innovation Fund Denmark (8053-00008B). Furthermore, it was partly supported by a research grant (15334) from VILLUM FONDEN, by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 757360) and by The Center for Quantification of Imaging Data from MAX IV (QIM) funded by The Capital Region of Denmark.

References

1. Ashukha, A., Lyzhov, A., Molchanov, D., Vetrov, D.: Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. arXiv preprint `arXiv:2002.06470` (2020)
2. Bandi, P., et al.: From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. IEEE Transactions on Medical Imaging **38**(2), 550–560 (2019). https://doi.org/10.1109/TMI.2018.2867350
3. Bejnordi, B.E., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA - Journal of the American Medical Association **318**(22), 2199–2210 (2017). https://doi.org/10.1001/jama.2017.14585
4. Ciompi, F., Geessink, O., Bejnordi, B.E., De Souza, G.S., Baidoshvili, A., Litjens, G., Van Ginneken, B., Nagtegaal, I., Van Der Laak, J.: The importance of stain normalization in colorectal tissue classification with convolutional networks. In: ISBI, pp. 160–163 (2017)
5. Falcon, W.: Pytorch lightning. GitHub. Note: `https://github.com/PyTorchLightning/pytorch-lightning` (2019)
6. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: ICML, pp. 1050–1059 (2016)
7. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML, pp. 1321–1330 (2017)

8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
9. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: ICLR (2019)
10. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: ICLR (2017)
11. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: ICLR, pp. 1–15 (2014)
12. Kirsch, A., van Amersfoort, J., Gal, Y.: Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In: NeurIPS, pp. 7026–7037 (2019)
13. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: NeurIPS, pp. 6402–6413 (2017)
14. Litjens, G., et al.: 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. GigaScience **7**(6) (2018). https://doi.org/10.1093/gigascience/giy065
15. Liu, Y., Kohlberger, T., Norouzi, M., Dahl, G., Smith, J., Mohtashamian, A., Olson, N., Peng, L., Hipp, J., Stumpe, M.: Artificial intelligence based breast cancer nodal metastasis detection: Insights into the black box for pathologists. Archives of Pathology & Laboratory Medicine **143**(7), 859–868 (2018)
16. Naeini, M.P., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using bayesian binning. In: AAAI (2015)
17. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In: NeurIPS, pp. 13991–14002 (2019)
18. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS, pp. 8024–8035 (2019)
19. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research **15**(1), 1929–1958 (2014)
20. Stacke, K., Eilertsen, G., Unger, J., Lundström, C.: A closer look at domain shift for deep learning in histopathology. arXiv preprint `arXiv:1909.11575` (2019)
21. Steiner, D.F., MacDonald, R., Liu, Y., Truszkowski, P., Hipp, J.D., Gammage, C., Thng, F., Peng, L., Stumpe, M.C.: Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. The American journal of surgical pathology **42**(12), 1636 (2018)
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 (2019)
23. Thulasidasan, S., Chennupati, G., Bilmes, J.A., Bhattacharya, T., Michalak, S.: On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In: NeurIPS, pp. 13888–13899 (2019)
24. Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., Bengio, Y.: Manifold mixup: Better representations by interpolating hidden states. In: ICLR, pp. 6438-6447 (2019)
25. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: ICLR (2018)