BENCHMARKING COMMON UNCERTAINTY ESTIMATION METHODS WITH HISTOPATHOLOGICAL IMAGES UNDER DOMAIN SHIFT AND LABEL NOISE

PREPRINT

Hendrik Mehrtens[1], **Alexander Kurz**[1], **Tabea-Clara Bucher, Titus J. Brinker**[2]
Division of Digital Biomarkers for Oncology
German Cancer Research Center (DKFZ)
Heidelberg, Germany
{hendrikalexander.mehrtens},{alexander.kurz},{tabea.bucher}@dkfz-heidelberg.de
titus.brinker@nct-heidelberg.de

September 2022

ABSTRACT

In the past years, deep learning has seen an increase of usage in the domain of histopathological applications. However, while these approaches have shown great potential, in high-risk environments deep learning models need to be able to judge their own uncertainty and be able to reject inputs when there is a significant chance of misclassification. In this work, we conduct a rigorous evaluation of the most commonly used uncertainty and robustness methods for the classification of Whole-Slide-Images under domain shift using the H&E stained Camelyon17 breast cancer dataset. Although it is known that histopathological data can be subject to strong domain shift and label noise, to our knowledge this is the first work that compares the most common methods for uncertainty estimation under these aspects. In our experiments, we compare Stochastic Variational Inference, Monte-Carlo Dropout, Deep Ensembles, Test-Time Data Augmentation as well as combinations thereof. We observe that ensembles of methods generally lead to higher accuracies and better calibration and that Test-Time Data Augmentation can be a promising alternative when choosing an appropriate set of augmentations. Across methods, a rejection of the most uncertain tiles leads to a significant increase in classification accuracy on both in-distribution as well as out-of-distribution data. Furthermore, we conduct experiments comparing these methods under varying conditions of label noise. We observe that the border regions of the Camelyon17 dataset are subject to label noise and evaluate the robustness of the included methods against different noise levels. Lastly, we publish our code framework to facilitate further research on uncertainty estimation on histopathological data.

Keywords Deep learning, Uncertainty estimation, Robustness, Histopathology, Domain shift, Label noise

1 Introduction

Deep Neural Networks (DNNs) have been shown to be of equal or even superior performance in studies on many medical tasks [Esteva et al., 2017, Haenssle et al., 2018, Hekler et al., 2019], compared to human practitioners. Nonetheless, they have rarely been adopted in clinical praxis. One commonly given reason [Begoli et al., 2019, Kompa et al., 2021, van der Laak et al., 2021] is their inability to provide well-calibrated estimates of their predictive uncertainty [Guo et al., 2017], thereby prohibiting the practitioner from judging the reliability of the system's decision, which is a necessary condition in areas of high uncertainty like medical decision-making.

[1]Contributed equally
[2]Corresponding author

Ideally, a well-calibrated deep learning system should be able to judge and communicate a correct estimate of its certainty in each prediction, not only informing the human practitioner of its momentary reliability but also enabling the system to automatically reject inputs [Band et al., 2021] and refer them to humans for inspection, thereby enhancing the reliability of the system.

Another problem in practice is the high vulnerability of DNN-based systems to "domain shifts", which are differences between the data distributions in the training and deployment setting. In the general machine learning literature, [Hendrycks and Dietterich, 2019, Ovadia et al., 2019] noticed the vulnerability of deep neural networks to even slight artificial perturbations of images, which result in drastic deterioration of classification performance. In the area of digital pathology, these kinds of shifts are very common [Stacke et al., 2021]. They arise due to different data-acquisition processes across clinics, for example, due to different staining procedures or different scanners but can also be caused by changes in the distribution of patient characteristics (gender, age, etc.). The unpredictable behavior of models in new data regimes limits the ability to put confidence in a system's decision, especially in high-risk environments, an example being decisions that affect a patient's treatment.

Traditional methods for improving the generalization of a machine learning system use augmentations of the input data. This technique is currently heavily investigated in the field of digital histopathology [Tellez et al., 2019]. However, as the topics of uncertainty quantification and generalization under domain shift have received attention in the whole field of machine learning, in recent years many new techniques have been proposed. All of these methods, however, have mostly been evaluated in the general setting of computer vision and object detection and not in the setting of digital pathology.

Whole-Slide-Images (WSIs) are a challenging domain for deep learning, due to the large size of the images, typically in the dimension of Gigapixels, and due to the sparsely available labeled data. Additionally, as the evaluation of a WSI is error-prone, with high inter-observer variability [Karimi et al., 2020], the labels are subject to a high degree of label noise, making the integration of predictive certainty a vital component.

[Linmans et al., 2020] investigated the ability of multi-headed ensemble networks to detect out-of-distribution (OOD) inputs in WSIs. [Thagaard et al., 2020] compared Deep Ensembles [Lakshminarayanan et al., 2017] and Monte-Carlo Dropout (MCDO) [Gal and Ghahramani, 2016] in their ability to estimate an uncertainty on tile-level. To our knowledge, there has been no thorough and reproducible investigation of these methods with respect to histopathological data, especially in the context of domain shift towards other clinics and the general effect of label noise.

As the estimation of uncertainty and robustness to domain shifts and label noise is of utmost importance in digital pathology, we evaluate the performance of the most popular approaches, comparable to the benchmark of [Band et al., 2021] on retina scans. To this end, we carefully compare the robustness of Stochastic Variational Inference (SVI) [Blundell et al., 2015], MCDO [Gal and Ghahramani, 2016], Deep Ensembles [Lakshminarayanan et al., 2017] and Test-Time Data Augmentation (TTA) [Ayhan and Berens, 2018] and combinations thereof under domain shift. We study the impact of the reject option by uncertainty on the performance of the model, observing a significant increase in accuracy in high certainty predictions, and compare multiple metrics of estimating the final certainty from the predictions. Finally, we simulate the effects of label noise in the edge regions of the tumor annotations.

For our experiments, we use the Camelyon17 dataset [Bándi et al., 2019] of breast cancer tissue, which consists of 50 slides with annotated tumor regions. The slides were obtained in five different clinics, using three different scanners, which enables us to evaluate the different approaches under a realistic domain shift scenario.

Our main contributions are:

- A systematic comparison of the most popular uncertainty estimation methods under domain shift in terms of predictive accuracy and network calibration with respect to histopathological data.
- A detailed analysis to what extent rejecting uncertain tiles improves classification performance and an analysis of the distribution of the rejected tiles.
- An investigation of the influence of label noise on the classification of WSIs and the robustness of the included uncertainty estimation methods against it.
- The release of an easily extendible code repository[1] to facilitate further research on the applicability of uncertainty estimation for deep neural networks.

2 Methods

In this section, we describe the methods, uncertainty measures and evaluation settings used in our work.

[1] `https://github.com/DBO-DKFZ/uncertainty-benchmark`

2.1 Uncertainty Estimation Methods

In uncertainty estimation, we want to compute the posterior predictive distribution of the output $\mathbf{y}$, given an input $\mathbf{x}$ and training data $\mathcal{D}$. The distribution can be formalized in terms of Bayesian Model Averaging over the model's parameters $\mathbf{w}$ as

$$p(\mathbf{y}|\mathbf{x}, \mathcal{D}) = \int p(\mathbf{y}|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|\mathcal{D}) \, \mathrm{d}\mathbf{w} \tag{1}$$

For neural networks, the posterior predictive distribution is analytically intractable. In recent years several methods have been proposed that approximate the distribution. In the following, we briefly introduce the most prominent methods.

2.1.1 Stochastic Variational Inference (SVI):

[Blundell et al., 2015] and [Graves, 2011] approximate the posterior distribution by placing a Gaussian distribution over every parameter of the neural network. As an objective for the approximation, the estimated lower-bound (ELBO) is minimized [Blei et al., 2017]. We use the Flipout formulation [Wen et al., 2018] of SVI for stabilizing the training.

2.1.2 Monte-Carlo Dropout (MCDO):

[Gal and Ghahramani, 2016] show that the dropout operation, originally intended as a regularization method to stabilize neural network training, can be used to approximate the true posterior of the neural network. For the approximation, multiple forward passes of the same input, with activated dropout layers, are aggregated during inference. The distribution over the predictions obtained through this method can be seen as samples from an approximation of the posterior distribution.

2.1.3 Deep Ensemble:

A Deep Ensemble [Lakshminarayanan et al., 2017] consists of multiple, in architecture identical, neural networks that are trained from different random initializations. The mean of all ensemble members serves as the prediction during inference. [Ovadia et al., 2019, Ashukha et al., 2020] show that Deep Ensembles outperform many other methods in terms of calibration and robustness under domain shift.

2.1.4 Test-Time Data Augmentations (TTA):

In contrast to Deep Ensembles, which employ multiple models during inference, TTA uses the same model multiple times by augmenting the input in different ways during inference. Generally, the same augmentations as at training time are applied during inference. [Ayhan and Berens, 2018, Ashukha et al., 2020] show the good performance of TTA in terms of robustness and calibration, which can come close to the performance of a Deep Ensemble while requiring less training time, as only one model is trained.

2.1.5 Temperature Scaling:

In their work, [Guo et al., 2017] show that introducing a single scaling factor T on the network's output can improve network calibration. More specifically, the output logits $\mathbf{z}_i$ are divided by T before applying the softmax function to the i-th input sample:

$$\hat{p}_i = \max_{c \in \mathcal{C}} \sigma_{\mathrm{SM}}(\mathbf{z}_i / T)^{(c)} \tag{2}$$

,where c is one class in the set of all possible classes $\mathcal{C}$. The optimal temperature $T > 0$ is computed by minimizing the negative log-likelihood (NLL) on a validation dataset.

2.1.6 Ensemble Variants:

Additionally, ensemble variants of the methods SVI, MCDO and TTA are implemented and tested.

2.2 Uncertainty Metrics

Given the predictions generated by the previously introduced methods, the literature developed multiple metrics to estimate the predictive uncertainty, which we compare in this work.

2.2.1 Confidence:

A commonly used method, especially in the field of calibration [Guo et al., 2017], is to take the maximum of the softmax, which is also called the "confidence" of the prediction. The core idea is that decisions that are considered certain are far away from the decision boundary, while uncertain decisions lie close to the boundary at $1/N$, where N is the number of classes.

2.2.2 Entropy:

Another often used metric [Mobiny et al., 2019, Band et al., 2021] is to take the entropy of the prediction as a measure of uncertainty. The entropy of the model's output probability is computed as

$$H(\mathbf{y}|\mathbf{x}, \mathcal{D}) = -\sum_{c \in \mathcal{C}} p(y = c|\mathbf{x}, \mathcal{D}) \log p(y = c|\mathbf{x}, \mathcal{D}) \tag{3}$$

Since the maximum possible entropy varies with the number of classes, we compute the normed entropy as

$$\begin{aligned} H_{\text{norm}} &= \frac{H}{H_{\max}} \\ \text{with } H_{\max} &= -\sum_{i=1}^{N} \frac{1}{N} \log\left(\frac{1}{N}\right) = \log N \end{aligned} \tag{4}$$

When using the normed entropy, the uncertainty is high when $H_{\text{norm}} \to 1$ and uncertainty is low when $H_{\text{norm}} \to 0$.

Confidence ($conf$) and entropy behave very similarly in a binary classification setting. Given two predictions y_i and y_j the following relation holds between them:

$$H(y_i) < H(y_j) \iff conf(y_i) > conf(y_j) \tag{5}$$

2.2.3 Variance:

Since all included uncertainty estimation methods generate a distribution of predictions, the variance of the distribution can be used as a measure of uncertainty [Lakshminarayanan et al., 2017, Abdar et al., 2021]. If all predictors agree on a result, the variance is zero, whereas a high variance indicates high uncertainty.

2.3 Evaluation Settings

This section covers the different evaluation settings and the performance metrics used in each setting.

2.3.1 Classification:

For the tumor classification on tile-level, we compute accuracy and balanced accuracy. Given the tumor class as the positive class, the balanced accuracy is defined as the arithmetic mean between the true positive rate $\text{TPR} = \text{TP}/\text{P}$ and the true negative rate $\text{TNR} = \text{TN}/\text{N}$:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{\text{TP}}{\text{P}} + \frac{\text{TN}}{\text{N}}\right) \tag{6}$$

,where P is the total number of positive samples and N is the total number of negative samples.

2.3.2 Calibration:

For measuring calibration, we utilize the Expected Calibration Error (ECE) [Guo et al., 2017, Nixon et al., 2019]. Given a prediction for every data point in the dataset, the output probability or "confidence" for each sample should on average match the correctness of the prediction. In other words, we expect a prediction that has a confidence value of 60% to be correct in about 60% of the cases. To validate this intuition, the predictions are split into a predetermined number M of bins B of equal confidence range. Then the absolute difference between the average confidence and accuracy within each bin is summed up:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \Big|\text{conf}(B_m) - \text{acc}(B_m)\Big| \tag{7}$$

Here, $|B_m|$ denotes the number of samples in the m-th bin and n is the total number of samples.

2.3.3 Reject Option:

In a clinical setting, the model should be able to refer predictions with high uncertainty to human practitioners for evaluation. Models suitable for this task should assign higher uncertainties to their wrong predictions than to their correct predictions, thereby allowing to cut off a large number of false predictions, by thresholding at a certainty level. To compare multiple models on this ability, we compute the accuracy-reject curve [Nadeem et al., 2009], plotting the achieved accuracy against the percentage of rejected data points in the dataset.

2.3.4 Label Noise:

Medical annotations are often subject to an unquantified amount of label noise [Joskowicz et al., 2019, Jensen et al., 2019, Karimi et al., 2020], which may deteriorate the performance of supervised machine learning approaches. To our knowledge, the previously described methods have not been compared in their robustness to label noise in the medical domain. We evaluate the effect by creating multiple datasets with increasing levels and different types of label noise and evaluate the methods under these changing conditions.

3 Experiment setup

3.1 Dataset and data processing

We conduct our experiments on the lesion-level annotated slides of the Cameylon17 dataset [Bándi et al., 2019]. This part of the Camelyon17 dataset consists of 50 WSIs of breast lymph node tissue, with the goal of detecting metastatic tissue. The slides were obtained from five different clinics in the Netherlands, using three different scanners, which serves as an ideal setting to estimate the influence of domain shift between clinics on the model performance.

To create a distribution shift between the in-distribution (ID) and out-of-distribution (OOD) domain, we split the dataset into three groups, using the centers 0, 1 and 3 with the 3DHistech scanner as ID data, where the training occurs, while using the 10 slides of centers 2 (Hamamatsu scanner) and 4 (Philips scanner) as OOD data each. By splitting the data between the centers, we induce differences in location and acquisition process between the datasets. We then further split the ID data into the training, validation and test set. We sort the slides of each center by the area of annotated tumor cells they contain and use the two median slides as test set for each center. The training and validation sets are generated by a randomized $75\%/25\%$ split of the tiles.

The tiles themselves are generated following [Khened et al., 2021] with median filtering and Otsu's thresholding of the HSV saturation component of the WSI image, followed by finally applying opening and closing dilation. After that, tiles of the size 256×256 are extracted. The tumor regions on the slide are indicated by polygonal annotations. From the annotations, a tumor coverage per tile is computed. Tiles with more than 25% tumor coverage are counted as tumor tiles and all tiles with 0% tumor coverage are counted as non-tumor tiles. Tumor tiles that are covered by the tumor annotation by less than 25% are discarded for our standard training, to minimize the risk of label noise that can arise due to high inter-observer variability [Jensen et al., 2019, Joskowicz et al., 2019, Karimi et al., 2020].

3.2 Training setup and hyperparameter tuning

We use a ResNet-34 [He et al., 2016] with a batch size of 128 and a learning rate of 0.001 for all our experiments. As optimizer we utilize Adam [Kingma and Ba, 2017], with a reduction of the learning rate by a factor of 10 if the validation loss, which is chosen as the cross-entropy loss, does not decrease for 3 epochs. For data augmentations, we follow [Tellez et al., 2019] applying random crops to size 224×224, random 90° rotations and color jitter (brightness:$\pm 20\%$, contrast:$\pm 30\%$, hue:$\pm 10\%$, saturation:$\pm 10\%$). The inputs are normalized with the mean and variance of the training data. The best model is chosen by balanced accuracy on the validation set. During training, we balance the training set by samples per class, but we do not balance the validation set.

For the Deep Ensemble architecture, we choose $n = 5$ members following recent literature [Linmans et al., 2020, Thagaard et al., 2020]. For MCDO, we place a dropout layer after each ResNet block, with dropout probability $p = 0.3$. This is in contrast to [Linmans et al., 2020, Thagaard et al., 2020], who only place a dropout layer before the last layer, observing no improvement in performance. For inference during testing, we use 10 SVI-, MCDO- and TTA- samples. Following [Wenzel et al., 2020], we use an additional hyperparameter for weighting the influence of the SVI prior (Kullback-Leibler-Divergence to the normal distribution) on the training. We tune this hyperparameter as well as the dropout probability, the dropout layer placement and the learning rate with the python library Optuna [Akiba et al., 2019]. All our experiments are conducted with PyTorch 1.11 on Nvidia GPUs with CUDA 11.3.

4 Results

We trained each method 5 times with different seeds, reporting their average or median performance.

4.1 Tile Classifier Performance

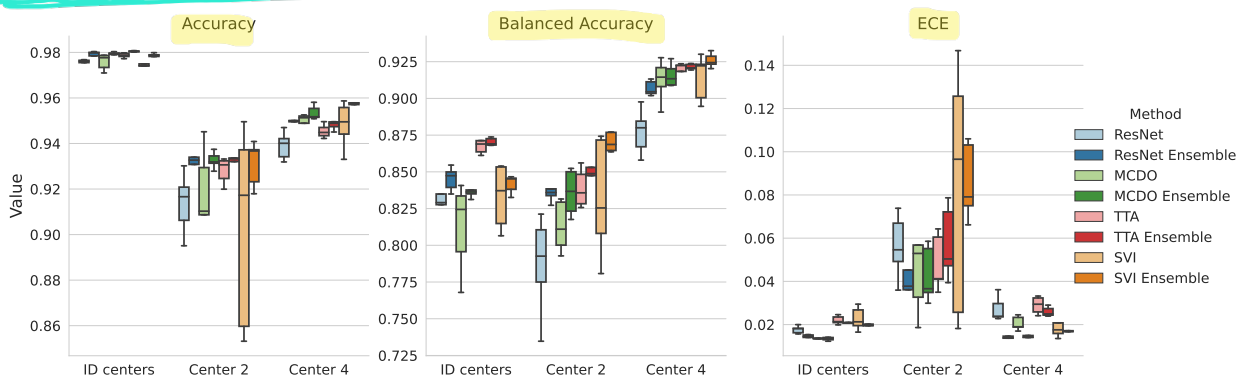

Figure 1: Accuracy, balanced accuracy and estimated calibration error (ECE) of the proposed methods on the in-distribution (ID) centers and out-of-distribution centers 2 and 4.

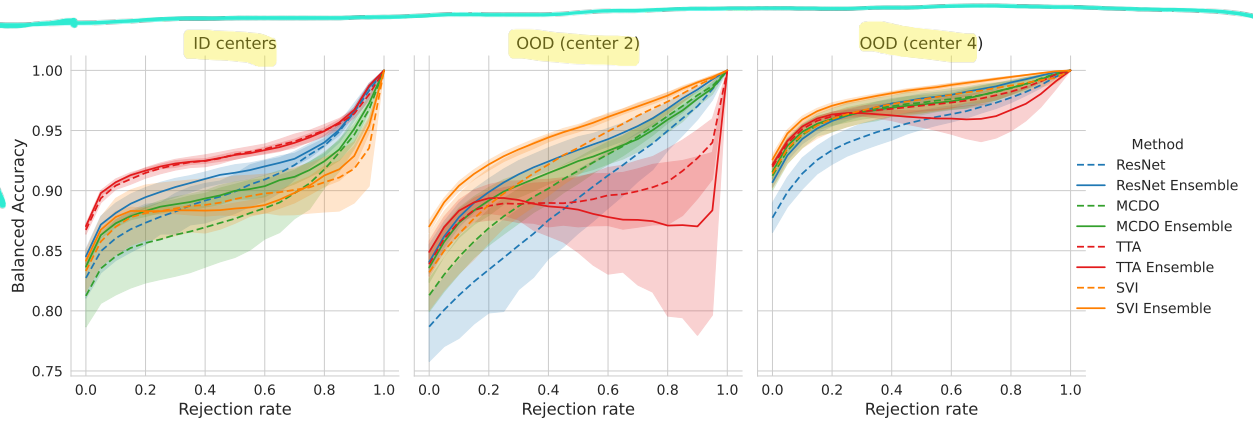

Figure 2: (Balanced) Accuracy-Reject Curves on ID and OOD centers. The x-axis contains the proportion of rejected inputs over the dataset, which is plotted against the reached Balanced Accuracy on the remaining images. The order of the excluded points was determined by the predictive confidence (see 2.2.1). We plot the mean curves over all 5 trials, with the shaded area denoting the 95% confidence interval.

Figure 1 shows the accuracy, balanced accuracy and ECE on the ID data, as well as on the OOD centers 2 and 4. Detailed numeric results for all our experiments can be found in the appendix in Table 2.

We start by analyzing the accuracy values on the three test data splits and between all evaluated uncertainty methods. On the ID test data, we can see that we achieve very high accuracies of about 98% for the classification of tumor and non-tumor tiles. Between the evaluated uncertainty methods, we only observe small differences in accuracy on the ID data. As expected, the ensemble variants outperform their non-ensemble counterparts and TTA performs better than the baseline. However, SVI achieves a slightly lower accuracy (97.4%) than the baseline ResNet, while MCDO can only slightly increase upon the baseline.

On OOD centers 2 and 4, we observe a decrease in accuracy compared to the ID test data, with the decrease on center 4 being smaller than the decrease observed on center 2. This behavior indicates a larger domain shift on center 2, than on center 4. We can observe an increase in performance with almost all uncertainty methods compared to the baseline ResNet, except for SVI and MCDO on center 2, where the results show a large spread in observed values. Especially the ensemble variants (ResNet Ensemble, MCDO Ensemble, TTA Ensemble and SVI Ensemble) consistently achieve higher accuracies than the baseline ResNet classifier.

Table 1: Mean and standard deviation of the area under the curve of the accuracy-reject curves ($AUARC$) presented in Figure 2. We compare the confidence of a single prediction, the confidence of the mean prediction for ensembling-like approaches and the variance over multiple predictions.

Unc. Measure	Unc. Method	ID Centers	OOD (Center 2)	OOD (Center 4)
Confidence	ResNet	$0.906_{\pm 0.017}$	$0.893_{\pm 0.031}$	$0.954_{\pm 0.010}$
Confidence	ResNet Ensemble	$0.918_{\pm 0.007}$	$0.932_{\pm 0.010}$	$0.972_{\pm 0.004}$
	MCDO	$0.886_{\pm 0.030}$	$0.914_{\pm 0.012}$	$0.971_{\pm 0.010}$
	MCDO Ensemble	$0.905_{\pm 0.008}$	$0.924_{\pm 0.012}$	$0.971_{\pm 0.007}$
	TTA	$0.932_{\pm 0.005}$	$0.897_{\pm 0.025}$	$0.970_{\pm 0.004}$
	TTA Ensemble	$\mathbf{0.933}_{\pm 0.002}$	$0.884_{\pm 0.044}$	$0.963_{\pm 0.009}$
	SVI	$0.895_{\pm 0.027}$	$0.931_{\pm 0.025}$	$0.972_{\pm 0.012}$
	SVI Ensemble	$0.896_{\pm 0.006}$	$\mathbf{0.949}_{\pm 0.003}$	$\mathbf{0.981}_{\pm 0.002}$
Variance	ResNet Ensemble	$0.914_{\pm 0.007}$	$0.928_{\pm 0.011}$	$0.972_{\pm 0.004}$
	MCDO	$0.877_{\pm 0.033}$	$0.911_{\pm 0.011}$	$0.968_{\pm 0.012}$
	MCDO Ensemble	$0.898_{\pm 0.010}$	$0.915_{\pm 0.011}$	$0.969_{\pm 0.008}$
	TTA	$\mathbf{0.926}_{\pm 0.005}$	$0.877_{\pm 0.031}$	$0.965_{\pm 0.005}$
	TTA Ensemble	$\mathbf{0.926}_{\pm 0.002}$	$0.861_{\pm 0.053}$	$0.959_{\pm 0.010}$
	SVI	$0.890_{\pm 0.026}$	$0.929_{\pm 0.023}$	$0.970_{\pm 0.013}$
	SVI Ensemble	$0.888_{\pm 0.007}$	$\mathbf{0.942}_{\pm 0.004}$	$\mathbf{0.981}_{\pm 0.001}$

Since we have to deal with a significant class imbalance between tumor and non-tumor tiles, we further investigate balanced accuracy in the center plot of Figure 1. Compared to the previous evaluation of accuracy, we observe lower values concerning balanced accurac/y for all centers. On the ID centers, the baseline ResNet achieves a balanced accuracy of 82.9%, while TTA and TTA Ensemble significantly improve performance (86.9% each), with MCDO even reaching a slightly lower balanced accuracy than the baseline. On the out-of-distribution centers all methods perform better than the baseline ResNet. Of all the methods TTA, TTA Ensemble and SVI Ensemble lead to the largest increases, while MCDO and MCDO Ensemble barely outperformed the pure ResNet Ensemble. Noticeable here is the high performance of SVI Ensemble on the OOD data, which could already be observed in terms of accuracy. SVI alone on the other hand again leads to a very high spread in observed values.

When computing the balanced accuracy, a higher weight is attributed to the under-represented tumor class. The lower values in balanced accuracy compared to standard accuracy indicate that our model performs worse in classifying tumor tiles than non-tumor tiles. Surprisingly, we see higher values in balanced accuracy on OOD center 4 compared to the other two test splits. To explain this behavior, we added an overview of the dataset distribution in the appendix (see B). The center 4 dataset has a larger ratio of tumor tiles to non-tumor tiles because it contains one slide with two large tumor metastases, much larger than the tumor metastases on all other WSIs. Therefore, the balanced accuracy metric is heavily influenced by the large increase in tumor tiles that lie *inside* the area of annotated tumor regions. As we will see in Section 4.3, these tumor tiles lead to higher confidences and better classification results, than tumor tiles that lie at the border of the annotation.

In the plot on the right-hand side of Figure 1, we evaluate model calibration in terms of ECE (see Section 2.3.2). On the ID test data, we observe low ECE values across all evaluated methods, which means that the output confidence values represent a good estimate of the correctness of the predictions. ResNet Ensemble (0.0145), as well as the MCDO approaches (0.0135 and 0.0137), produce better-calibrated results than the baseline, while all TTA and SVI methods slightly worsen the calibration. As shown in [Ovadia et al., 2019], we expect higher ECE values for the domain shift towards OOD centers 2 and 4. On center 2 this is also reflected in our results, as we see higher ECE values and higher interquartile ranges for all evaluated methods compared to the ID test data. Here the relative order remains mostly the same, with TTA and TTA Ensemble now performing better than the ResNet, while the ResNet Ensemble and the MCDO methods still perform best. Center 4 on the other hand shows lower ECE values, comparable to the values on the ID test data. This indicates that with the applied augmentations, we were able to mostly compensate for the domain shift between ID and OOD data, which was already visible in our accuracy results. Here the MCDO methods and the ResNet Ensemble perform better than the baseline method, while the situation of SVI and TTA methods is reversed to that of center 2.

In C, we analyze the influence of data augmentations on network calibration by comparing the results to training with basic augmentations consisting of random crops and flips. There we see higher ECE values for both OOD centers 2 and 4 compared to the ID test data. In the evaluation in the appendix we also include Temperature Scaling as a reference method for improving network calibration.

To summarize the results, we find that data augmentations applied during training have a larger impact on the achieved accuracies and the calibration compared to utilizing one of our evaluated uncertainty methods and as such a large focus should be placed on them. Amongst the methods evaluated in Figure 1, ensembles lead to slightly better-calibrated results than the baseline, while TTA and SVI behave inconsistently and can have a negative impact on the resulting ECE values. In terms of achieved accuracies, the ensemble approaches consistently perform best. The SVI method leads to inconsistent results across the evaluated metrics, therefore no direct conclusion is possible. TTA does not require architectural changes and no extra computational resources during training and leads to a large increase in observed accuracies. If however larger training capacity is available, a TTA Ensemble or even an SVI Ensemble can be recommended. The basic ResNet Ensemble, while not the best in most evaluations, performed the most consistent, always outperforming the baseline method.

4.2 Reject Option

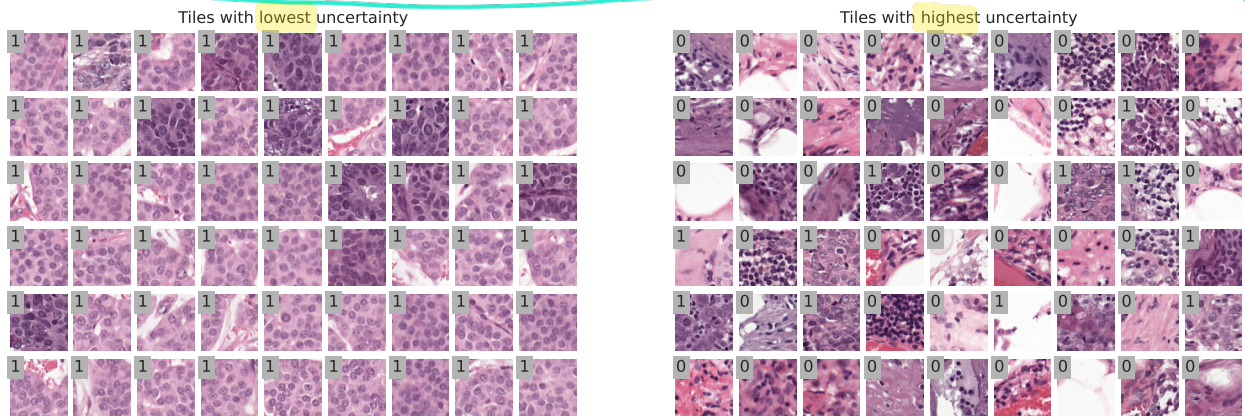

Figure 3: Least and most uncertain tiles from the in-distribution test set computed by the confidence of a ResNet Ensemble. The numbers represent the ground-truth labels: '1' indicates tumor and '0' indicates a non-tumor tile.

We compare the uncertainty methods in terms of their ability to detect mispredictions. For that we compute the (Balanced) Accuracy-Reject Curves (see Section 2.3.3). In Figure 2, we compare the methods and their performance with an increasing ratio of rejected tiles, for the test slides on the ID centers and OOD centers 2 and 4, respectively. To determine the most uncertain tiles that are rejected, we choose the confidence measure (see Section 2.2.1). For each split and method, we see a mostly linear increase of balanced accuracy when rejecting a fraction of the most uncertain tiles. On the ID test data, TTA and the TTA Ensemble consistently achieve the highest balanced accuracies. For the OOD centers, the SVI Ensemble leads to the highest accuracy-reject-curves, while the baseline ResNet performs worst. From our evaluated uncertainty methods, the MCDO methods averaged over all three datasets led to the worst results. While most methods show increasing performance on center 2 when rejecting the most uncertain tiles, TTA and TTA Ensemble show an unexpected stagnation of performance when rejecting more than 20% of uncertain tiles.

In Table 1 we show the area under the curve for the accuracy-reject curves ($AUARC$) across all datasets and compare different uncertainty measures. On the ID data, ResNet Ensemble, TTA and TTA Ensemble lead to an increase of the AUARC over the baseline value, while the MCDO and SVI methods lead to a decrease of said metric. On center 2 and center 4, all methods (other than TTA Ensemble on center 2) lead to higher AUARC values than the baseline. From our experiments, not one single method can be identified, that performs best between all data splits. However, the ResNet Ensemble consistently beats the basic ResNet, while TTA and TTA Ensemble perform at least as well as the baseline or better.

When comparing the scores computed from using the confidence of the predictions or the variance across individual predictions, the relative order of methods remains the same across all datasets. Furthermore, every method performs worse in terms of AUARC, when the uncertainty is calculated by the variance across individual predictions instead of the confidence. Although the disagreement of ensembling methods measured in terms of the variance is often claimed to be a good indicator of uncertainty, in our experiments it consistently performs worse when compared to using the confidence for detecting mispredictions. As using the confidence for misprediction detection by rejection is equivalent to using the entropy in the binary case (see Section 2.2.2), our experiments strongly suggest utilizing confidence or entropy over the variance of predictions.

In Figure 3 we present a collection of the most certain and most uncertain tiles within the ID test data, computed with a ResNet Ensemble. We observe that the neural network appears to be most confident on tumor tiles (label 1), that cover the whole tile and possess a similar cell structure. For the most uncertain tiles on the right side, no comparable structure among the tiles is observable. These tiles both contain tumor and non-tumor tissue and often lie at border regions between tissue types, for example, the border to fat-tissue, as can be identified by the white areas in the H&E stained slides. Many uncertain tiles seem to lie at the border of annotated tumor regions, that we suspect to have a larger degree of label noise, and in regions that have a different appearance than the majority of healthy tissue.

4.3 Label Noise

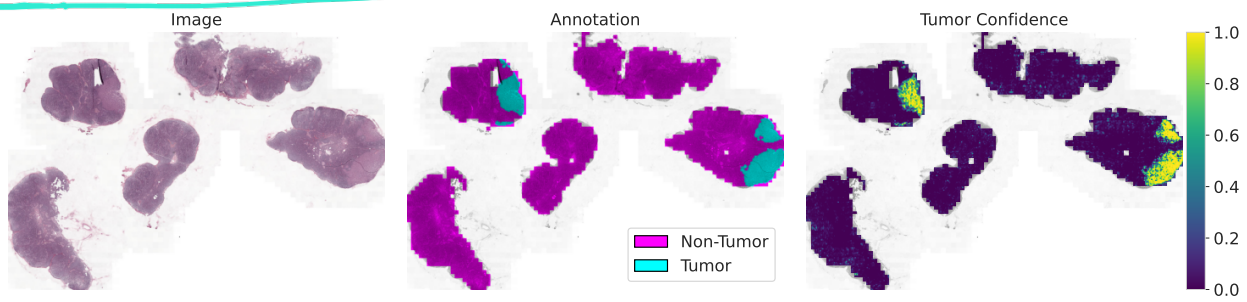

Figure 4: WSI (*patient_017_node_2*) with ground-truth annotations and model predictions. The tumor confidence decreases in areas near the border of the annotation, while the uncertainty thereby increases. This result is consistent over the whole dataset.

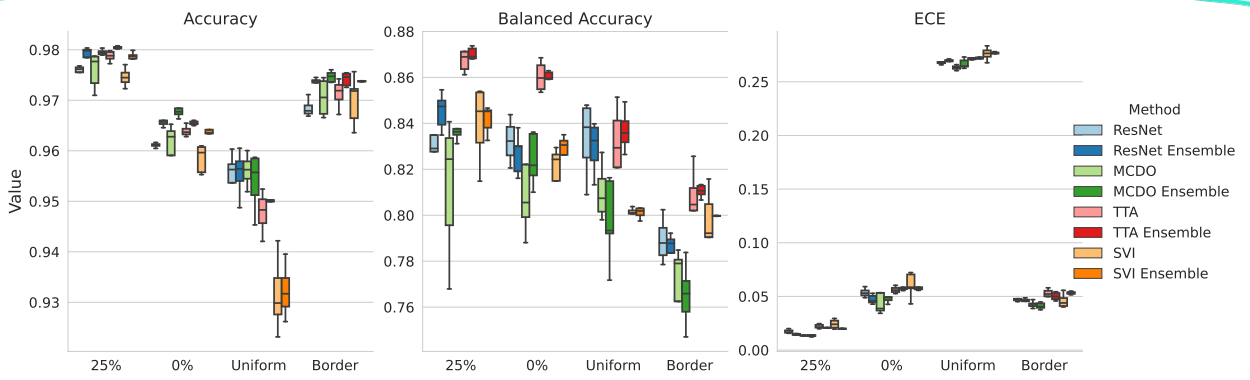

Figure 5: Performance of the proposed methods on different datasets under label noise. We compare the performance on the original dataset (25%), a dataset with a 0% tumor coverage threshold for tumor tiles (0%), applying uniform noise to the tile labels (Uniform) and randomly flipping labels of tiles which are located at the border of the annotation (Border).

For translating our tile-level observations to slide-level, we stitch the tile-level predictions back to a tumor confidence map on slide-level that we show on the example of one slide in Figure 4.

When observing the generated confidence maps, we can see lower tumor confidences at the border of annotated tumor regions. This goes along with the effect of inter-observer variability, where the border of the annotated tumor area is expected to vary between observers.

Building on these observations, in our label noise experiments we investigate the effects of imprecise tumor annotations in the border area of annotated tumor regions. To this end, we define three supplementary datasets by introducing different types of label noise to the annotations. As we suspect the annotations to be rough themselves, we first create a dataset by setting the inclusion threshold by tumor coverage for tumor tiles to 0%. We previously excluded every tile that was covered by less than 25% by tumor annotations (see Section 3.1) as we already suspected a high chance of noisy annotations on these tiles. The other two datasets are created by applying random label noise to the training split of the 0% threshold dataset. First, we apply uniform label noise to the whole slide, with a 25% chance of flipping the tile class. As this type of annotation noise does not reflect real-world inter-observer variability, we next apply

label flipping to the border regions of the annotation. We flip the labels of the tumor tiles, which lie at the border of the annotation polygon and thereby are not fully covered by annotated tumor cells. We set the chance of this event occurring to 25% per tile.

In Figure 5 we show the results of the label noise experiments. Detailed results can be found in Table 3. In terms of accuracy, the ensemble methods outperform their non-ensemble counterparts by a significant margin. Only on the Uniform dataset, ResNet and MCDO perform very similar to their ensemble counterparts ($\sim 95.6\%$ accuracy for each method). When viewing the balanced accuracy metric, TTA and TTA Ensemble exceed every other method by a large margin. SVI and SVI Ensemble are the least robust methods when exposed to label noise, with SVI often performing significantly worse than our single baseline ResNet.

We can conclude that TTA and ensembling approaches are not only robust to domain shifts and image corruptions [Ovadia et al., 2019] but in a similar manner also to label noise, in our case in histopathological images. SVI and MCDO, however, are not fit to deal with label noise often leading to only slightly improved or even worse results.

TTA however does not perform well in terms of calibration error. Here MCDO outperforms TTA and SVI, which produced the overall worst calibrated predictions, in contrast to recent literature [Ashukha et al., 2020, Ayhan et al., 2020]. We can see a large increase in ECE on the dataset with uniform label noise and a slight increase in the miscalibration on the other two datasets with label noise compared to our baseline dataset. Except for the original dataset, no trend of ensembling methods decreasing calibration error is visible. Ensembling does not seem to improve calibration when confronted with larger quantities of label noise, contrary to the setting of domain shift (Figure 1) where ensembling decreased the calibration error.

Finally, we observe that the results for the split with the 0% tumor coverage threshold are worse than the results for our originally selected dataset, containing a tumor coverage of at least 25% per tile and the dataset with added noise in the border regions. We suspect this behavior to be due to a large amount of inherent label noise in the annotations of the Camelyon17 dataset.

5 Discussion

In the previous section, we have gone through an extensive comparison of the most prominent methods for uncertainty estimation under domain shift on histopathological WSIs. In this section, we want to discuss the observations that we have made and we want to formulate recommendations for other researchers that try to integrate uncertainty estimation into digital pathology. Our results show that mispredictions can be detected reliably and that the right methods can increase the robustness to domain shift and label noise, while also providing better-calibrated predictions.

Among the methods for uncertainty estimation, ensembles lead to the most reliable uncertainty estimates and additionally improve classification performance and network calibration. When training an ensemble of multiple networks is too expensive, TTA has, specifically on ID data, shown to improve performance when discarding the most uncertain tiles, while requiring no change in the used architectures. However, if a good calibration is required, TTA can not be recommended. On the other hand, the often used MCDO methodology did not lead to significant improvements in our experiments and can not be recommended. Flipout SVI leads to inconsistent results, being very susceptible to label noise and underperforming in the task of misprediction detection on ID data.

As expected, combining MCDO, TTA and SVI with ensembling leads to further improvements in classification performance, however, it also entails a steep increase in computational requirements, which might not be possible in some medical environments.

In terms of misprediction detection on ID data, all methods provide a reliable improvement in classification performance, with a growing rejection rate making misprediction detection a feasible scenario in a clinical setting. The largest increase for every method is at the beginning of the accuracy-reject-curve, with the slope decreasing rapidly around the 20% reject rate threshold. From our results, it is plausible that in many cases it could be enough to only reject around 20% of most uncertain predictions to receive a significant accuracy boost. Our experiments reconfirm the strong impact of choosing appropriate data augmentations, as has been reported in [Tellez et al., 2019, Stacke et al., 2021]. Choosing the right set of augmentations is a critical factor for OOD performance.

We compared multiple uncertainty metrics, determining the most uncertain tiles in terms of minimum confidence, maximum entropy and maximum variance. From our results, we recommend using the confidence metric as it is defined for every method and performed slightly better in our experiments. It would be of high interest to compare these metrics in a histopathological multi-class setting to generate a stronger recommendation.

By sorting the tile predictions by uncertainty, we observe visually recognizable differences between the most certain and most uncertain tiles, which are consistent over all methods. On the WSIs, the networks make especially confident

predictions for tiles that lie inside tumor metastases, while being the most uncertain in border regions between different tissue types, which can also be seen when visualizing the uncertainty on slide level. Our label noise experiments strengthen this point, by raising the suspicion that the Camelyon17 dataset contains noisy labels in the border area of annotated tumor regions. The inclusion of the likely noisy labels from the dataset immediately leads to a drop in classifier performance. Practitioners should therefore not only concern themselves with the possibility of domain shift but should also put a large focus on extracting pure ground truth labels, dropping labels that are potentially misclassified. In our experiments, the ensemble approaches and especially TTA showed increased resistance to label noise.

Further studies could extend our experiments to a broader range of datasets, as our scope is limited by the availability of tile-level annotated histopathological data in the Camelyon datasets. Moreover, the evaluation of uncertainty on slide level remains an open problem. Additional methods could be compared, for example, integrating multi-head ensembles [Linmans et al., 2020] or deterministic uncertainty methods [Postels et al., 2022], which offer a sampling-free and fast alternative. Our label noise experiments could be extended to more methods and more realistic scenarios.

6 Conclusion

Deployment of AI-based diagnostic systems in the safety-critical area of histopathology demand uncertainty-aware machine learning algorithms, which generate trust in the model's predictions.

To this end, we compared multiple uncertainty estimation methods and uncertainty metrics across domain shift and label noise scenarios in their performance, calibration and ability to detect mispredictions in the histopathological setting. Our results show that on in-domain data, ensembles and TTA are well-performing methods, while under domain shift the relative order and gain of methods is harder to determine. Existing methods are well-capable to detect mispredictions and reject inputs they are unlikely to classify correctly. Furthermore, ensembles generally are better calibrated than their singular counterpart.

As label noise can be a large problem in medical data, practitioners should put great care into identifying potentially mislabeled inputs, as they can lead to large performance decreases. In terms of robustness to label noise, ensembles and TTA also perform best. When comparing the uncertainty metrics of confidence, variance and entropy, we found no significant difference between them and suggest using the predictive confidence, as it is easy to implement and a reliable metric to detect mispredictions.

The ensemble performed the most consistent over all our experiments, TTA can definitely be recommended in addition to ensembling or even by itself, as it is compute-efficient and easy to implement without requiring architectural changes or retraining. MCDO performed worse than TTA, SVI or the ResNet ensemble on most measures, only providing good results in terms of predictive calibration. SVI, on the other hand, could compete with TTA in some scenarios, but is harder to implement and train and in our experiments lead to a higher variability between results.

While no single benchmark can give all-encompassing results and insights, we hope that our evaluation gives guidance for the utilization of uncertainty methods in the area of histopathology. Our published code is designed to be easily reproducible and extendable to further studies.

Acknowledgments

The research is funded by the *Ministerium für Soziales und Integration*, Baden Württemberg, Germany.

References

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, February 2017. ISSN 1476-4687. doi:10.1038/nature21056.

H. A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. Ben Hadj Hassen, L. Thomas, A. Enk, L. Uhlmann, Reader study level-I and level-II Groups, Christina Alt, Monika Arenbergerova, Renato Bakos, Anne Baltzer, Ines Bertlich, Andreas Blum, Therezia Bokor-Billmann, Jonathan Bowling, Naira Braghiroli, Ralph Braun, Kristina Buder-Bakhaya, Timo Buhl, Horacio Cabo, Leo Cabrijan, Naciye Cevic, Anna Classen, David Deltgen, Christine Fink, Ivelina Georgieva, Lara-Elena Hakim-Meibodi, Susanne Hanner, Franziska Hartmann, Julia Hartmann, Georg Haus, Elti Hoxha, Raimonds Karls, Hiroshi Koga, Jürgen Kreusch, Aimilios Lallas, Pawel Majenka, Ash Marghoob, Cesare Massone, Lali Mekokishvili, Dominik Mestel, Volker Meyer, Anna Neuberger, Kari Nielsen, Margaret Oliviero, Riccardo Pampena, John Paoli, Erika Pawlik, Barbar Rao, Adriana Rendon, Teresa Russo, Ahmed Sadek, Kinga Samhaber, Roland Schneiderbauer, Anissa Schweizer, Ferdinand Toberer, Lukas Trennheuser, Lyobomira Vlahova, Alexander Wald, Julia Winkler, Priscila Wölbing, and Iris Zalaudek. Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology: Official Journal of the European Society for Medical Oncology*, 29(8):1836–1842, August 2018. ISSN 1569-8041. doi:10.1093/annonc/mdy166.

Achim Hekler, Jochen S. Utikal, Alexander H. Enk, Axel Hauschild, Michael Weichenthal, Roman C. Maron, Carola Berking, Sebastian Haferkamp, Joachim Klode, Dirk Schadendorf, Bastian Schilling, Tim Holland-Letz, Benjamin Izar, Christof von Kalle, Stefan Fröhling, Titus J. Brinker, Laurenz Schmitt, Wiebke K. Peitsch, Friederike Hoffmann, Jürgen C. Becker, Christina Drusio, Philipp Jansen, Joachim Klode, Georg Lodde, Stefanie Sammet, Dirk Schadendorf, Wiebke Sondermann, Selma Ugurel, Jeannine Zader, Alexander Enk, Martin Salzmann, Sarah Schäfer, Knut Schäkel, Julia Winkler, Priscilla Wölbing, Hiba Asper, Ann-Sophie Bohne, Victoria Brown, Bianca Burba, Sophia Deffaa, Cecilia Dietrich, Matthias Dietrich, Katharina Antonia Drerup, Friederike Egberts, Anna-Sophie Erkens, Salim Greven, Viola Harde, Marion Jost, Merit Kaeding, Katharina Kosova, Stephan Lischner, Maria Maagk, Anna Laetitia Messinger, Malte Metzner, Rogina Motamedi, Ann-Christine Rosenthal, Ulrich Seidl, Jana Stemmermann, Kaspar Torz, Juliana Giraldo Velez, Jennifer Haiduk, Mareike Alter, Claudia Bär, Paul Bergenthal, Anne Gerlach, Christian Holtorf, Ante Karoglan, Sophie Kindermann, Luise Kraas, Moritz Felcht, Maria R. Gaiser, Claus-Detlev Klemke, Hjalmar Kurzen, Thomas Leibing, Verena Müller, Raphael R. Reinhard, Jochen Utikal, Franziska Winter, Carola Berking, Laurie Eicher, Daniela Hartmann, Markus Heppt, Katharina Kilian, Sebastian Krammer, Diana Lill, Anne-Charlotte Niesert, Eva Oppel, Elke Sattler, Sonja Senner, Jens Wallmichrath, Hans Wolff, Anja Gesierich, Tina Giner, Valerie Glutsch, Andreas Kerstan, Dagmar Presser, Philipp Schrüfer, Patrick Schummer, Ina Stolze, Judith Weber, Konstantin Drexler, Sebastian Haferkamp, Marion Mickler, Camila Toledo Stauner, and Alexander Thiem. Superior skin cancer classification by the combination of human and artificial intelligence. *European Journal of Cancer*, 120:114–121, October 2019. ISSN 0959-8049. doi:10.1016/j.ejca.2019.07.019.

Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, January 2019. ISSN 2522-5839. doi:10.1038/s42256-018-0004-1.

Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. Second opinion needed: Communicating uncertainty in medical machine learning. *npj Digital Medicine*, 4(1):1–6, January 2021. ISSN 2398-6352. doi:10.1038/s41746-020-00367-3.

Jeroen van der Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: The path to the clinic. *Nature Medicine*, 27(5):775–784, May 2021. ISSN 1546-170X. doi:10.1038/s41591-021-01343-4.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330. PMLR, July 2017.

Neil Band, Tim G J Rudner, Qixuan Feng, Angelos Filos, Zachary Nado, Michael W Dusenberry, Ghassen Jerfel, Dustin Tran, and Yarin Gal. Benchmarking Bayesian Deep Learning on Diabetic Retinopathy Detection Tasks. In *NeurIPS 2021*, page 15, 2021.

Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *ICLR 2019*, page 16, 2019.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model' s uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundström. Measuring Domain Shift for Deep Learning in Histopathology. *IEEE Journal of Biomedical and Health Informatics*, 25(2):325–336, February 2021. ISSN 2168-2208. doi:10.1109/JBHI.2020.3032060.

David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58:101544, December 2019. ISSN 1361-8415. doi:10.1016/j.media.2019.101544.

Davood Karimi, Haoran Dou, Simon K. Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759, October 2020. ISSN 1361-8415. doi:10.1016/j.media.2020.101759.

Jasper Linmans, Jeroen van der Laak, and Geert Litjens. Efficient Out-of-Distribution Detection in Digital Pathology Using Multi-Head Convolutional Neural Networks. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, pages 465–478. PMLR, September 2020.

Jeppe Thagaard, Søren Hauberg, Bert van der Vegt, Thomas Ebstrup, Johan D. Hansen, and Anders B. Dahl. Can You Trust Predictive Uncertainty Under Real Dataset Shifts in Digital Pathology? In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, Lecture Notes in Computer Science, pages 824–833, Cham, 2020. Springer International Publishing. ISBN 978-3-030-59710-8. doi:10.1007/978-3-030-59710-8_80.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1050–1059. PMLR, June 2016.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Networks. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1613–1622. PMLR, June 2015.

Murat Seçkin Ayhan and Philipp Berens. Test-time Data Augmentation for Estimation of Heteroscedastic Aleatoric Uncertainty in Deep Neural Networks. In *Proceedings of the First Conference on Medical Imaging with Deep Learning*, page 9, 2018.

Péter Bándi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders B. Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Çetin, Eren Halıcı, Hunter Jackson, Richard Chen, Fabian Both, Jörg Franke, Heidi Küsters-Vandevelde, Willem Vreuls, Peter Bult, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens. 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, February 2019. ISSN 1558-254X. doi:10.1109/TMI.2018.2867350.

Alex Graves. Practical Variational Inference for Neural Networks. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, April 2017. ISSN 0162-1459. doi:10.1080/01621459.2017.1285773.

Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse. Flipout: Efficient Pseudo-Independent Weight Perturbations on Mini-Batches. In *ICLR 2018*, 2018.

Arsenii Ashukha, Dmitry Molchanov, Alexander Lyzhov, and Dmitry Vetrov. Pitfalls of In-Domain Uncertainty Estimation and Ensembling in Deep Learning. In *ICLR 2020*, page 30, 2020.

Aryan Mobiny, Aditi Singh, and Hien Van Nguyen. Risk-Aware Machine Learning Classifier for Skin Lesion Diagnosis. *Journal of Clinical Medicine*, 8(8):1241, August 2019. ISSN 2077-0383. doi:10.3390/jcm8081241.

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76: 243–297, December 2021. ISSN 1566-2535. doi:10.1016/j.inffus.2021.05.008.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring Calibration in Deep Learning. In *CVPR Workshop*, page 4, 2019.

Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. Accuracy-Rejection Curves (ARCs) for Comparing Classification Methods with a Reject Option. In *Proceedings of the Third International Workshop on Machine Learning in Systems Biology*, pages 65–81. PMLR, March 2009.

Leo Joskowicz, D. Cohen, N. Caplan, and J. Sosna. Inter-observer variability of manual contour delineation of structures in CT. *European Radiology*, 29(3):1391–1399, March 2019. ISSN 1432-1084. doi:10.1007/s00330-018-5695-5.

Martin Holm Jensen, Dan Richter Jørgensen, Raluca Jalaboi, Mads Eiler Hansen, and Martin Aastrup Olsen. Improving Uncertainty Estimation in Convolutional Neural Networks Using Inter-rater Agreement. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Lecture Notes in Computer Science, pages 540–548, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32251-9. doi:10.1007/978-3-030-32251-9_59.

Mahendra Khened, Avinash Kori, Haran Rajkumar, Ganapathy Krishnamurthi, and Balaji Srinivasan. A generalized deep learning framework for whole-slide image segmentation and analysis. *Scientific Reports*, 11(1):11579, June 2021. ISSN 2045-2322. doi:10.1038/s41598-021-90444-8.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE. ISBN 978-1-4673-8851-1. doi:10.1109/CVPR.2016.90.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR 2015*. arXiv, January 2017.

Florian Wenzel, Kevin Roth, Bastiaan S. Veeling, Jakub Świątkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How Good is the Bayes Posterior in Deep Neural Networks Really? In *ICML 2020*, July 2020.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 2623–2631, New York, NY, USA, July 2019. Association for Computing Machinery. ISBN 978-1-4503-6201-6. doi:10.1145/3292500.3330701.

Murat Seçkin Ayhan, Laura Kühlewein, Gulnar Aliyeva, Werner Inhoffen, Focke Ziemssen, and Philipp Berens. Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. *Medical Image Analysis*, 64:101724, August 2020. ISSN 1361-8415. doi:10.1016/j.media.2020.101724.

Janis Postels, Mattia Segù, Tao Sun, Luca Daniel Sieber, Luc Van Gool, Fisher Yu, and Federico Tombari. On the Practicality of Deterministic Epistemic Uncertainty. In *Proceedings of the 39th International Conference on Machine Learning*, pages 17870–17909. PMLR, June 2022.

A Detailed Results

Table 2: Detailed results for tile classifier performance on Camelyon17. We report median and interquartile range as shown in the boxplots.

Split	Method	Accuracy ↑	Balanced Accuracy ↑	ECE ↓
ID	ResNet	$0.9762_{0.0010}$	$0.8291_{0.0073}$	$0.0181_{0.0020}$
	ResNet Ensemble	$0.9798_{0.0015}$	$0.8474_{0.0105}$	$0.0145_{0.0011}$
	MCDO	$0.9777_{0.0053}$	$0.8244_{0.0380}$	$\mathbf{0.0135}_{0.0001}$
	MCDO Ensemble	$0.9794_{0.0006}$	$0.8363_{0.0026}$	$0.0137_{0.0010}$
	TTA	$0.9788_{0.0013}$	$0.8689_{0.0076}$	$0.0214_{0.0026}$
	TTA Ensemble	$\mathbf{0.9805}_{0.0003}$	$\mathbf{0.8690}_{0.0043}$	$0.0207_{0.0003}$
	SVI	$0.9743_{0.0009}$	$0.8372_{0.0385}$	$0.0213_{0.0073}$
	SVI Ensemble	$0.9787_{0.0007}$	$0.8452_{0.0076}$	$0.0196_{0.0007}$
OOD (Center 2)	ResNet	$0.9166_{0.0146}$	$0.7926_{0.0355}$	$0.0546_{0.0177}$
	ResNet Ensemble	$0.9327_{0.0032}$	$0.8359_{0.0046}$	$0.0378_{0.0091}$
	MCDO	$0.9103_{0.0207}$	$0.8110_{0.0293}$	$0.0529_{0.0241}$
	MCDO Ensemble	$0.9319_{0.0032}$	$0.8367_{0.0268}$	$\mathbf{0.0366}_{0.0203}$
	TTA	$0.9306_{0.0077}$	$0.8357_{0.0200}$	$0.0412_{0.0196}$
	TTA Ensemble	$0.9333_{0.0016}$	$0.8488_{0.0056}$	$0.0504_{0.0249}$
	SVI	$0.9173_{0.0775}$	$0.8254_{0.0636}$	$0.0965_{0.1000}$
	SVI Ensemble	$\mathbf{0.9367}_{0.0141}$	$\mathbf{0.8687}_{0.0123}$	$0.0791_{0.0281}$
OOD (Center 4)	ResNet	$0.9401_{0.0080}$	$0.8800_{0.0174}$	$0.0238_{0.0061}$
	ResNet Ensemble	$0.9500_{0.0005}$	$0.9045_{0.0079}$	$\mathbf{0.0139}_{0.0009}$
	MCDO	$0.9513_{0.0032}$	$0.9145_{0.0129}$	$0.0190_{0.0045}$
	MCDO Ensemble	$0.9517_{0.0039}$	$0.9134_{0.0110}$	$0.0142_{0.0009}$
	TTA	$0.9448_{0.0036}$	$0.9225_{0.0041}$	$0.0294_{0.0065}$
	TTA Ensemble	$0.9489_{0.0021}$	$0.9208_{0.0028}$	$0.0249_{0.0030}$
	SVI	$0.9495_{0.0117}$	$0.9222_{0.0227}$	$0.0176_{0.0049}$
	SVI Ensemble	$\mathbf{0.9577}_{0.0007}$	$\mathbf{0.9244}_{0.0054}$	$0.0169_{0.0004}$

Table 3: Detailed results for our experiments under label noise. We report median and interquartile range as shown in the boxplots.

Experiment	Method	Accuracy ↑	Balanced Accuracy ↑	ECE ↓
25%	ResNet	$0.9762_{0.0010}$	$0.8291_{0.0073}$	$0.0181_{0.0020}$
	ResNet Ensemble	$0.9798_{0.0015}$	$0.8474_{0.0105}$	$0.0145_{0.0011}$
	MCDO	$0.9777_{0.0053}$	$0.8244_{0.0380}$	$\mathbf{0.0135}_{0.0001}$
	MCDO Ensemble	$0.9794_{0.0006}$	$0.8363_{0.0026}$	$0.0137_{0.0010}$
	TTA	$0.9788_{0.0013}$	$0.8689_{0.0076}$	$0.0214_{0.0026}$
	TTA Ensemble	$\mathbf{0.9805}_{0.0003}$	$\mathbf{0.8690}_{0.0043}$	$0.0207_{0.0003}$
	SVI	$0.9745_{0.0019}$	$0.8453_{0.0219}$	$0.0240_{0.0066}$
	SVI Ensemble	$0.9787_{0.0007}$	$0.8452_{0.0076}$	$0.0196_{0.0007}$
0%	ResNet	$0.9612_{0.0004}$	$0.8323_{0.0124}$	$0.0531_{0.0046}$
	ResNet Ensemble	$0.9658_{0.0007}$	$0.8299_{0.0110}$	$0.0460_{0.0061}$
	MCDO	$0.9628_{0.0048}$	$0.8055_{0.0229}$	$\mathbf{0.0387}_{0.0165}$
	MCDO Ensemble	$\mathbf{0.9677}_{0.0012}$	$0.8217_{0.0181}$	$0.0491_{0.0030}$
	TTA	$0.9637_{0.0011}$	$\mathbf{0.8598}_{0.0105}$	$0.0558_{0.0038}$
	TTA Ensemble	$0.9654_{0.0005}$	$0.8596_{0.0029}$	$0.0570_{0.0014}$
	SVI	$0.9596_{0.0050}$	$0.8243_{0.0115}$	$0.0586_{0.0129}$
	SVI Ensemble	$0.9635_{0.0008}$	$0.8306_{0.0062}$	$0.0571_{0.0020}$
Uniform	ResNet	$0.9563_{0.0037}$	$\mathbf{0.8383}_{0.0215}$	$0.2678_{0.0012}$
	ResNet Ensemble	$0.9564_{0.0039}$	$0.8325_{0.0143}$	$0.2698_{0.0013}$
	MCDO	$0.9563_{0.0034}$	$0.8074_{0.0143}$	$\mathbf{0.2633}_{0.0024}$
	MCDO Ensemble	$\mathbf{0.9557}_{0.0073}$	$0.7934_{0.0228}$	$0.2644_{0.0057}$
	TTA	$0.9483_{0.0047}$	$0.8294_{0.0204}$	$0.2712_{0.0013}$
	TTA Ensemble	$0.9500_{0.0003}$	$0.8358_{0.0092}$	$0.2719_{0.0012}$
	SVI	$0.9299_{0.0072}$	$0.8010_{0.0014}$	$0.2765_{0.0059}$
	SVI Ensemble	$0.9317_{0.0057}$	$0.8019_{0.0030}$	$0.2770_{0.0011}$
Border	ResNet	$0.9679_{0.0016}$	$0.7879_{0.0119}$	$0.0468_{0.0015}$
	ResNet Ensemble	$0.9739_{0.0005}$	$0.7878_{0.0060}$	$0.0463_{0.0015}$
	MCDO	$0.9706_{0.0065}$	$0.7790_{0.0182}$	$0.0420_{0.0025}$
	MCDO Ensemble	$\mathbf{0.9747}_{0.0018}$	$0.7659_{0.0134}$	$\mathbf{0.0405}_{0.0042}$
	TTA	$0.9719_{0.0029}$	$0.8047_{0.0098}$	$0.0523_{0.0053}$
	TTA Ensemble	$0.9746_{0.0024}$	$\mathbf{0.8107}_{0.0038}$	$0.0500_{0.0059}$
	SVI	$0.9718_{0.0058}$	$0.7922_{0.0144}$	$0.0439_{0.0074}$
	SVI Ensemble	$0.9738_{0.0001}$	$0.7997_{0.0002}$	$0.0532_{0.0017}$

B Dataset statistics

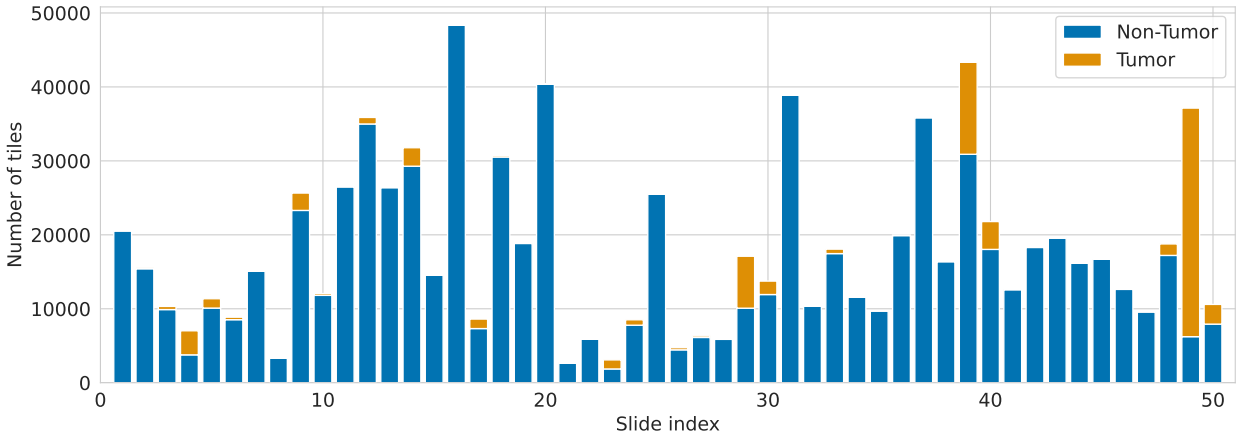

Figure 6: Tumor vs. Non-Tumor ratio for all lesion-level annotated slides. Slides are sorted by center with 10 slides belonging to each center. This means slides with index $[21, 30]$ belong to OOD center 2 and slides with index $[41, 50]$ belong to OOD center 4. All other slides are part of the in-distribution dataset consisting of center 0, 1 and 3 which use the same slide scanner.

C Effects of data augmentations on network calibration

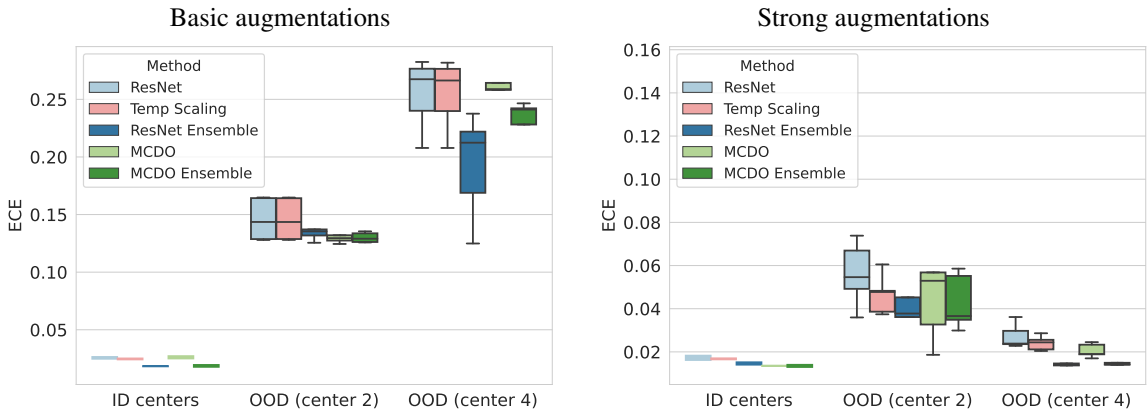

Table 4: Expected calibration error between data splits for training with two different settings of augmentations. For basic augmentations, we see increasing calibration error accross methods on both OOD centers. The plot on the right represents the augmentations used in the main part. As uncertainty methods we include MCDO and Deep Ensembles and we additionally include Temperature Scaling as another method to improve network calibration.