scientific reports

OPEN

Generalisation effects of predictive uncertainty estimation in deep learning for digital pathology

Milda Pocevičiūtė[1,2✉], Gabriel Eilertsen[1,2], Sofia Jarkman[2,3] & Claes Lundström[1,2,4]

Deep learning (DL) has shown great potential in digital pathology applications. The robustness of a diagnostic DL-based solution is essential for safe clinical deployment. In this work we evaluate if adding uncertainty estimates for DL predictions in digital pathology could result in increased value for the clinical applications, by boosting the general predictive performance or by detecting mispredictions. We compare the effectiveness of model-integrated methods (MC dropout and Deep ensembles) with a model-agnostic approach (Test time augmentation, TTA). Moreover, four uncertainty metrics are compared. Our experiments focus on two domain shift scenarios: a shift to a different medical center and to an underrepresented subtype of cancer. Our results show that uncertainty estimates increase reliability by reducing a model's sensitivity to classification threshold selection as well as by detecting between 70 and 90% of the mispredictions done by the model. Overall, the deep ensembles method achieved the best performance closely followed by TTA.

The performance of deep learning (DL) has surpassed that of both human experts and other analytic methods on many prediction tasks in computer vision as well as other applications[1,2]. It has also shown great potential in pathology applications, such as breast cancer metastases detection[3–5] and grading of prostate cancer[6–9]. The workflow of a digital pathology laboratory consists of preparation of glass slides, containing tissue or cell sample from a patient, which then are digitalised using specialised scanners to produce whole slide images (WSIs). This allows pathologists to review the patient case using computers instead of microscopes and enables deployment of deep learning systems to assist the review[10].

One remaining challenge for wide clinical adoption of DL in pathology, perhaps the most important one, is that the performance of neural networks (NNs) can deteriorate substantially due to domain shift[11], i.e., differences in the distributions between the training data and the data for which prediction occurs. In digital pathology, differences of WSIs are observed between different centres due to, for instance, medical protocols, tissue preparation processes or scanner types used.

The typical strategy to achieve high generalization capacity of a DL model is to ensure high diversity in the training data. In pathology, collecting data from several sources (care providers) is important. Another approach to reducing the domain shift error is to apply extensive augmentations during the training stage, which is an active research area in the pathology domain[12,13]. Unfortunately, these methods are yet far from completely alleviating the performance drop due to domain shift. Thus, exploring more strategies to ensure a model's robustness when deployed is highly motivated.

The uncertainty of a prediction is a source of information that is typically not used in DL applications. There is, however, rationale indicating that it could be beneficial. By generating multiple, slightly varied, predictions as base for an uncertainty estimate, additional information about the model's sensitivity to changes is made available. This potential added value could be relevant for both the generalisation challenge and for boosting performance overall. In particular, we argue that the added dimension of uncertainty could be utilised as a building block for clinical workflows where pathologists and DL models interact, for instance for triaging WSIs or sorting WSI regions of interest.

In this paper, we explore a number of research questions that are central for understanding how uncertainty can contribute to robustness of DL for digital pathology:

- Can uncertainty estimates add to the predictive capacity?

[1]Department of Science and Technology, Linköping University, Linköping, Sweden. [2]Center for Medical Image Science and Visualization, Linköping University, Linköping, Sweden. [3]Department of Clinical Pathology, and Department of Biomedical and Clinical Sciences, Linköping University, Linköping, Sweden. [4]Sectra AB, Linköping, Sweden. ✉email: milda.poceviciute@liu.se

- Are uncertainty estimates effective for flagging potential prediction errors?
- Is the added value of uncertainty affected by domain shift?
- Are computationally more demanding uncertainty methods, i.e. deep ensembles and TTA, more effective than the softmax score of the NN?
- Do custom architectures, designed to provide uncertainty estimation, outperform model-independent methods?

Our experiments explore two types of domain shift in histology of lymph nodes in breast cancer cases. One shift concerns data coming from two different care providers, in different countries. Another shift concerns the challenge of dealing with cancer subtypes, whether uncertainty estimation could be beneficial when there are rare conditions that may lack sufficient amount of training data. Around 70% of breast cancer cases are ductal carcinomas. The second most common subtype is lobular carcinomas which accounts to roughly 10% of the positive cases. Furthermore, this subtype is often more challenging for a pathologist to detect due to its less obvious infiltrative patterns[14,15]. Therefore, it is reasonable to assume that DL models not specifically trained for lobular carcinomas will have lower performance there. We use the ductal vs lobular carcinoma scenario as a proxy for the general case of subtypes that are under-represented in the training data.

With respect to uncertainty estimation methods, there is an important distinction between methods that influence the choice of NN architecture or training procedure, and methods that are independent of how the DL model was designed. The first category includes MC dropout[16], ensembles[17] and other techniques[18–20]. However, model-independent methods such as Test time augmentations (TTA)[21] have several advantages. One advantage is that constraints on model design may lead to suboptimal performance. Moreover, model independence opens the possibility to benefit from uncertainty estimates for any model—also for the locked-down commercial solutions that are typically deployed in the clinic. Thus, studying the effectiveness of model-independent uncertainty estimation is of a particular interest. To the best of our knowledge, we are the first to compare the effectiveness of TTA, MC dropout, and ensembles on classification tasks in digital pathology.

In this work, we train an NN classifier as a base for our evaluation of uncertainty methods. We contribute to the understanding of uncertainty and deep learning for digital pathology in four ways. First of all, we propose a way of combining the uncertainty measure with the softmax score in order to boost generalisability of the model. Secondly, we measure how well misclassified patches can be deteced by uncertainty methods. Thirdly, we compare the effectiveness of three uncertainty methods (Deep ensembles, MC dropout, and TTA) together with four different metrics utilising the multiple predictions—three established measures (sample variance, entropy and mutual information) and our proposed metric (sample mean uncertainty). Finally, we investigate if uncertainty estimations generalise over a clinically realistic domain shift, and for mitigating the problem of a rare cancer subtype that is under-represented in the training data.

Related work

Uncertainty estimation is an important topic in deep learning research that holds potential in providing more calibrated predictions and increasing the robustness of NNs. The methods can be categorised based on what statistical theory they are grounded on: frequentist approaches, Bayesian neural networks (BNNs) and Bayesian approximations for standard NNs[22]. The methods based on frequentist statistics commonly use ensembles[17,23], bootstrapping[24] and quantile regression[25]. BNNs are based on Bayesian Variational Inference and estimate the posterior distribution for a given task, and thus provide uncertainty distributions over parameters by design. However, currently their adaptation to the medical imaging domain is slow due to the higher computational costs of training and poor uncertainty estimation[22,26]. There is also a more recent line of research showing that certain transformations of the softmax confidence score[27], or some modification to the network architecture[28], may produce a reasonable estimation of uncertainty without any additional computations. To reflect this recent trend we include a comparison of direct uncertainty estimation from softmax score in all of our experiments, as well as an uncertainty estimator based on the sample mean over different network evaluations.

In deep learning applications within the medical domain, most research effort has been devoted to radiology, with MC dropout and Deep ensembles being two common methods compared in the literature. Nair et al.[20] showed that the MC dropout[16] method can improve multiple sclerosis detection and segmentation. They evaluated the uncertainty measures by omitting a certain portion of the most uncertain predictions and comparing the effect on false positive and true positive rates. Kyono et al.[29] evaluated if AI-assisted mammography triage could be safely implemented in the clinical workflow of breast cancer diagnosis. They estimated uncertainty by combining the MC dropout and TTA methods, and concluded that this approach could provide valuable assistance.

Within computational pathology, the most similar previous work is by Thagaard et al.[30], which evaluated the deep ensembles[17], MC dropout[16], and mixup[31] methods for breast cancer metastases detection in lymph nodes. They trained an NN model for breast cancer metastasis detection and evaluated its performance in combination with the three uncertainty estimation methods on several levels of domain shifts: in-domain test data (same hospital, same organ), breast cancer metastases in lymph nodes from a different hospital, colorectal cancer (different hospital and organ), and head and neck squamous cell carcinoma (different hospital, organ and sub-type of cancer) metastases to the lymph nodes. They found that Deep ensembles[17] performed considerably better on most evaluation criteria except for detecting squamous cell carcinoma where mixup[31] showed better results. Similarly, Linmans et al.[19] showed that uncertainties computed by Deep ensembles as well as a multi-head CNN[32] allowed for detection of out-of-distribution lymphoma in sentinel lymph nodes of breast cancer cases.

TTA in medical imaging has successfully been applied for segmentation tasks. Graham et al.[33] improved the performance of gland instance segmentation in colorectal cancer by incorporating TTA uncertainties into the NN system. Wang et al.[34] compared the potential gains from using MC dropout, TTA or a combination of both

on segmentation performance of fetal brains and brain tumours from 2D and 3D magnetic resonance images. They found that the combination of the two methods achieved the best results.

In comparison with previous research efforts, our work brings novel contributions in several ways. This includes evaluating the model-agnostic TTA method for classification in pathology, and making the comparison to model-integrated methods. We introduce an approach to combine a model's softmax score with an uncertainty measure in order to improve the predictive performance. Moreover, we use a broader evaluation scheme for misprediction detection where all classification thresholds are considered instead of a single one. The broad scheme also includes evaluation of three uncertainty estimation methods using four different metrics, whereas previous work mostly has focused on the entropy metric. Finally, in all experiments we include a baseline based on the softmax score from one single model, in order to clearly measure the improvement that can be achieved by the added complexity of uncertainty estimation methods.

Material and methods

In this section we describe the three uncertainty estimation methods and the four uncertainty metrics that are evaluated in our experiments. Then we provide the details about the NN algorithms that we trained for the classification task, the training procedure, and the datasets used for training and evaluation of the uncertainty methods and metrics.

Uncertainty estimation methods.
All of the methods have the same basic principle: to produce multiple predictions for each input. The variation within these predictions shows how uncertain the model is.

MC dropout. We are interested in computing posterior probability distribution $p(W|X, Y)$ over the NN weights W given the input patches X and corresponding ground true labels Y. This posterior is intractable, but it can be approximated using variational inference with some parameterised distribution $q^*(W)$ that minimises the Kullback-Leibler (KL) divergence:

$$q^*(W) = \arg\min_{q(W)} KL(q(W) \| p(W|X, Y)).$$

Gal et al.[16] showed that minimising the cross-entropy loss of an NN with dropout layers, is equivalent to minimising the KL divergence above. Furthermore, the authors show that we can treat the samples obtained by multiple stochastic passes through an NN with the dropout enabled as an approximation of the model's uncertainty. Following Thagaard et al.[30], we added a dropout layer with probability 0.5 in the NN before the logits. During test time, we activated the dropout layer with the same probability and ran 50 stochastic passes for each input.

Deep ensembles. This is a method based on training T identical NNs with different random seeds. During the inference, the T predictions per input are aggregated for uncertainty estimation[17]. Following previous work[19,30], we set $T = 5$.

Test time augmentations. Each input is randomly augmented T times before passing through the trained model. The uncertainty scores are computed from the T predictions. Usually, the test time augmentations are identical to the ones applied during the training of the model[21,34]. In our experiments we set $T = 50$ to match the number of forward passes in the MC dropout method. For a detailed description of the augmentations, refer to Sect. 3.3.

Uncertainty metrics.
Once we obtain the multiple predictions per input, we can compute an uncertainty metric. In this work we compared three well established metrics, sample variance, entropy and mutual information. In addition, we introduce the *sample mean uncertainty* metric which is based on a probabilistic interpretation of the softmax score in a binary classification problem.

Sample mean uncertainty. This metric is based on the mean of the samples generated by an uncertainty estimation method. We define sample mean uncertainty, u_s, as:

$$u_s = 1 - 2(\bar{s} - 0.5)^2,$$

where $\bar{s}$ is the average of softmax scores s_i over T predictions:

$$\bar{s} = \frac{1}{T} \sum_{i=1}^{T} s_i.$$

The value range of the measure is between 0 and 1, and assigns high value for patches that have the mean tumour softmax score around 0.5, indicating that they are potentially more uncertain. Low values are observed when the softmax scores are close to 0 or 1, implying high confidence in the corresponding binary classifications. The measure reflects the general dependence between softmax confidence and uncertainty.

Also, it shares characteristics with the estimator based on max predicted softmax probability for any class, which was evaluated in[27].

Sample variance. This metric is derived by taking the variance across T number of predictions per input produced by each of the uncertainty methods[20].

	Country	Total WSIs	Positive WSIs	Patches	Cancer types
Camelyon16 (test)	Netherlands	129	49	40 940	N/A
BRLN data	Sweden	114	57	39 354	Ductal, lobular
Lobular data	Sweden	71	14	6 960	Lobular
Ductal data	Sweden	96	39	6 960	Ductal

Table 1. Information about the datasets used in the evaluation of the model and the uncertainty methods.

Entropy. For a discrete random variable X, Shannon entropy quantifies the amount of uncertainty inherent in the random variable's outcomes. It is defined as[35]:

$$H(X) = -\sum_i P(x_i) \log P(x_i),$$

which we approximate for each input i as[36]:

$$H(\hat{y}_i|\mathbf{W}, \mathbf{D}) \approx -\sum_{c=1}^{C} \frac{1}{T}\sum_{t=1}^{T}\left[P(\hat{y}_i = c|W_t, D_t) \cdot \log\left(\frac{1}{T}\sum_{t=1}^{T} P(\hat{y}_i = c|W_t, D_t)\right)\right],$$

where T is the number of predictions per input generated by an uncertainty estimation method, C is the number of classes in our data, $\mathbf{D}$ is the dataset, $\hat{y}_i$, $i \in |\mathbf{D}|$ is a prediction by the classifier, and $\mathbf{W}$ are the parameters of the classifier. We refer to this metric as 'entropy'.

Mutual information (MI). The MI metric was first defined by Shannon[35]. It measures how much information we gain for each input by observing the samples produced by an uncertainty estimation method. It is approximated by[36]:

$$MI(\hat{y}_i, \mathbf{W}|\mathbf{D}) \approx H(\hat{y}_i|\mathbf{W}, \mathbf{D}) - E\big[H(\hat{y}_i|W_t, D_t)\big],$$

where $H(\hat{y}_i|\mathbf{W}, \mathbf{D})$ is the entropy of expected predictions. $E\big[H(\hat{y}_i|W_t, D_t)\big]$ is the expected entropy of model predictions across the samples generated by an uncertainty estimation method which can be approximated as[36]:

$$E\big[H(\hat{y}_i|W_t, D_t)\big] \approx -\sum_{c=1}^{C}\frac{1}{T}\sum_{t=1}^{T}\left[P(\hat{y}_i = c|W_t, D_t) \cdot log(P(\hat{y}_i = c|W_t, D_t))\right]$$

Network training. We trained five Resnet18 models[37] with He initialisation[38] and a dropout layer (with probability 0.5)[39] before the logits with five different random seeds. The data augmentations during the training as well as the testing time were based on the work of Tellez et al.[12]. That is, on each input we applied horizontal flip with probability 0.5, 90 degrees rotations, scaling factor between 0.8 and 1.2, HSV colour augmentation by adjusting hue and saturation intensity ratios between [-0.1, 0.1], brightness intensity ration: [0.65, 1.35], contrast intensity ratio: [0.5, 1.5]. We also applied additive Gaussian noise and Gaussian blur, both with $\sigma \in [0.0, 0.1]$.

Each training epoch consisted of 131 072 patches sampled from the training WSIs with equal number of tumour and healthy patches. We used ADAM optimiser with $\beta_1 = 0.9$, $\beta_2 = 0.999$, initial learning rate of 0.01 with learning rate decay of 0.1 applied when the validation accuracy was not improving for 4 epochs. The models were trained until convergence with the maximum limit of 100 epochs. From each training setup, the best performing model in terms of validation accuracy was chosen.

Datasets.
In-domain data in this project is the Camelyon16 dataset[40] which contains 399 whole-slide images (WSI) of hematoxylin and eosin (H&E) stained lymph node sections collected in two medical centres in the Netherlands. The slides were scanned with the 3DHistech Pannoramic Flash II 250 and Hamamatsu NanoZoomer-XR C12000-01 scanners. 270 WSIs from Camelyon16 dataset were used for the training and validation of Resnet18 models . A balanced test dataset was created by extracting the patches from the official Camelyon16 test set which contained 129 WSIs. This was used for the in-domain performance evaluation. In our experiments, 'Camelyon16 data' refers to the set of patches extracted from the Camelyon16 test set unless otherwise noted.

Our out-of-domain class-balanced patch data is extracted from 114 H&E stained WSIs of lymph node sections from a medical center in Sweden which were annotated by a resident pathologist with 4 years of experience aided with immunostained slides. This is a subset of the larger AIDA BRLN dataset[41], which was scanned by Aperio ScanScope AT and Hamamatsu NanoZoomer scanners (XR, S360, and S60). We refer to it as BRLN.

Table 1 lists the four datasets of patches extracted from Camelyon16 and BRLN that were used in our experiments. These datasets were only used for the evaluation. In BRLN data, we have two cancer subtypes: lobular and ductal carcinomas.

In order to study uncertainty effects on generalisation to the lobular cancer subtype, we created two subsets of BRLN data which we call Lobular and Ductal data. They consist of 3480 tumour patches of each cancer subtype

	Camelyon16	BRLN	Lobular	Ductal
ROC-AUC	0.975	0.968	0.899	0.982
PR-AUC	0.981	0.977	0.928	0.987

Table 2. PR and ROC AUC values based on softmax scores (single model) for each of the 4 datasets.

and the same 3480 healthy patches. As all the datasets are publicly available, we did not need to obtain an ethical approval for our study.

Evaluation metrics. We evaluate our results based on area under the curve (AUC) of receiver operating characteristic (ROC) and precision recall (PR). ROC-AUC is the most common metric used to evaluate the performance of a binary classifier[42] and also in uncertainty evaluation in digital pathology[19,30,33]. ROC-AUC captures the trade-off between the true positive rate (TPR), also known as recall, and false positive rate (FPR), also known as 1—specificity:

$$TPR = \text{Recall} = \frac{TP}{TP + FN}$$
$$FPR = 1 - \text{Specificity} = \frac{FP}{TN + FP}$$

The PR curve plots precision against recall where precision is the fraction of positive predictions that are truly positive[43]:

$$\text{Precision} = \frac{TP}{TP + FP}$$

In addition to the AUC measures that aggregate performance across all classification thresholds, we are interested in examining in detail how performance of methods and metrics varies for different choices of classification thresholds. For this comparison, we look at accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Results

Basis for the experiments. *Performance of classifier.* In order to draw any meaningful conclusions on evaluation of uncertainty methods and metrics, we first need to ensure that the base classifier has a reasonable performance.

Table 2 shows ROC-AUC and PR-AUC on the four datasets. Overall, the achieved performance of 0.975 ROC-AUC (0.981 PR-AUC) on in-domain Camelyon16 data indicates that the Resnet18 was a sufficient classifier to perform this detection task. The drop of below 1% in ROC-AUC (and PR-AUC) between Camelyon16 and BRLN data sets is consistent with other work observing that a well-trained model should suffer relatively small decrease in performance under domain shift arising from different medical centers[4].

Increased error for lobular carcinoma. Investigating the model's performance on cancer subtypes within BRLN data, we found that it exhibits a substantially worse result on the lobular carcinoma: 0.889 ROC AUC (0.928 PR AUC) compared to the 0.982 ROC AUC (0.987 PR AUC) on the ductal cancer subtype (see Table 2). This result confirms that there indeed is a domain shift effect due to tumour type, in line with our assumptions.

Boosting metastases detection. Given the multiple predictions provided by three different methods (MC dropout, ensembles, and TTA), the most straightforward method for boosting predictive performance is to utilise traditional ensemble techniques. The most common one is to average the softmax output over the different inference runs/models/augmentations. The results are demonstrated in Figs. 1 and 2 for Camelyon16 and BRLN, respectively, and compared to using a single prediction as baseline. The results show a consistent but small improvement in terms of ROC-AUC for the deep ensembles and TTA. This is also reflected by the accuracy curve, demonstrating how the averaging improves the results by a small margin (~0.3 percentage points) at the optimal classification threshold, for both Camelyon16 and BRLN.

Another strategy for boosting predictive performance is to consider uncertainty estimation and softmax output from a single network as separate entities. Although the uncertainty methods also make use of different combinations of the softmax score, it is interesting to investigate if this approach holds benefits over traditional techniques. We do this by turning the classification task into a two-dimensional thresholding problem, with the softmax score and the uncertainty measure as two separate dimensions.

We observed that for tumour patches it was more common to have a combination of high entropy uncertainty and low softmax score compared to the healthy patches (see Fig. 3a). This inspired us to propose an alternative classification score defined by:
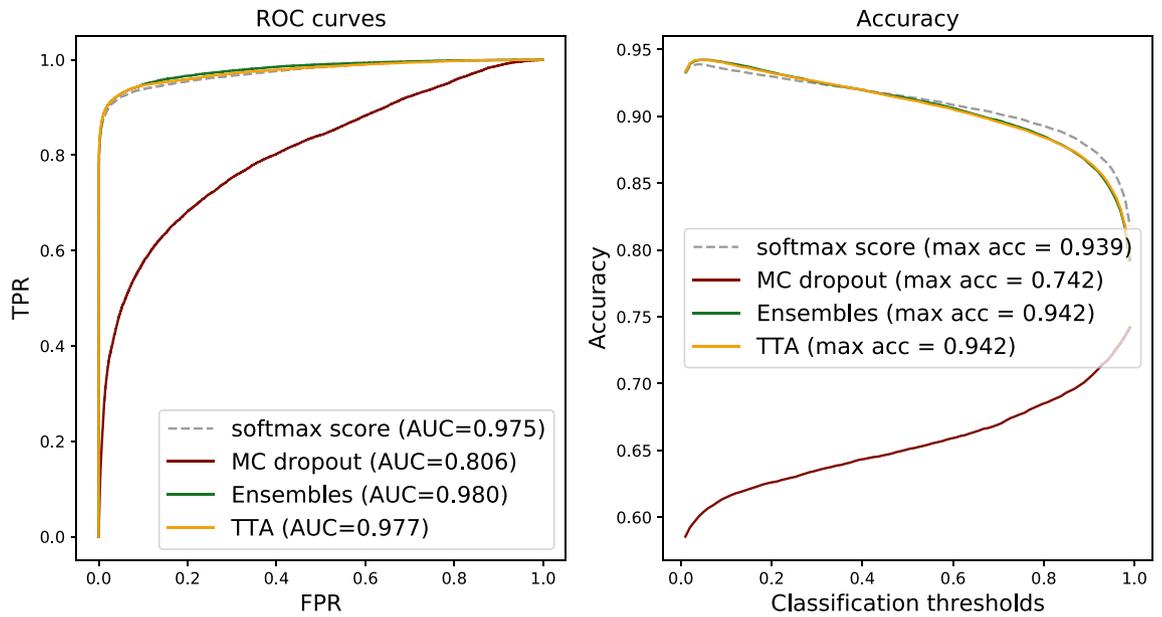
Figure 1. Tumour metastases detection on Camelyon16 data: ROC curves and accuracy of using softmax tumour score from a single NN vs averages of softmax tumour scores (per input) produced by the uncertainty estimation methods.
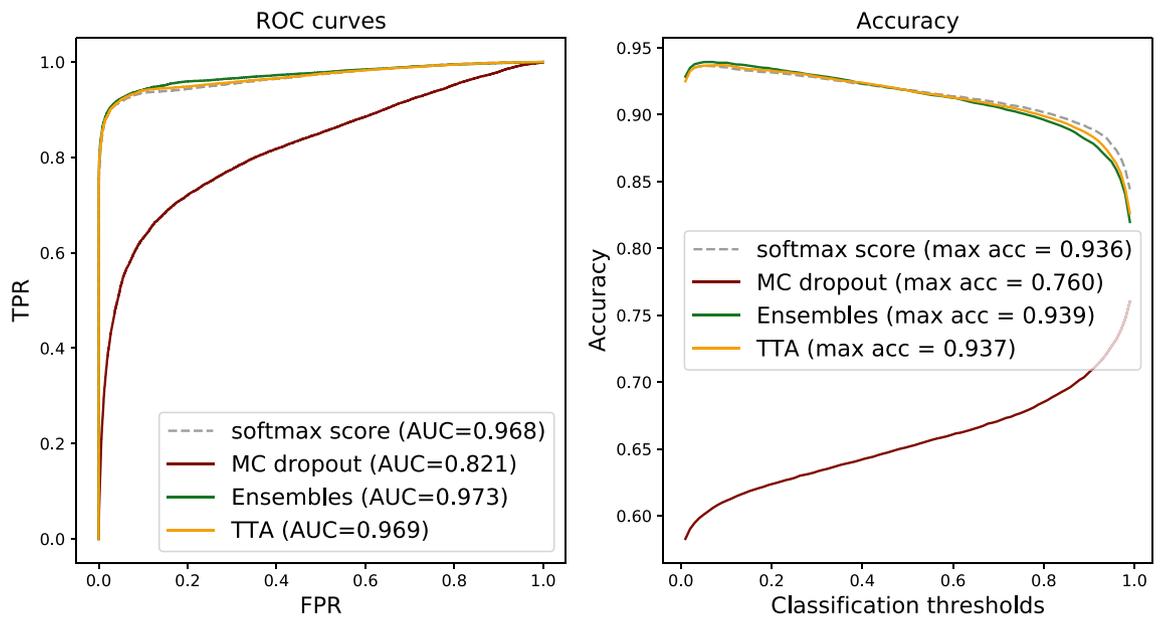

Figure 2. Tumour metastases detection on BRLN data: ROC curves and accuracy of using softmax tumour score from a single NN vs averages of softmax tumour scores (per input) produced by the uncertainty estimation methods.

$$f(u,s) = \left(\left(\frac{u}{P_u} \right)^y + s^y \right)^{\frac{1}{y}},$$

where u is the uncertainty measure, s is the softmax score for the tumour (positive) class from one single NN. The factor P_u is used to normalise the range of uncertainties, and we define this to be the 99th percentile of the uncertainty value range in the data. Based on a specified threshold t, the prediction is positive for $f(u,s) > t$, otherwise negative. The curve $f(u,s) = t$ intersects the axes at t, and the exponent y can be used to control its shape, from circular for $t = 2$ towards square for large t. For all experiments we use $t = 10$. Figure 3 illustrates the 2D space spanned by softmax score and uncertainty estimation, for two different methods, with corresponding 2D decision boundaries, $f(u,s) = t$, for a selection of different t.

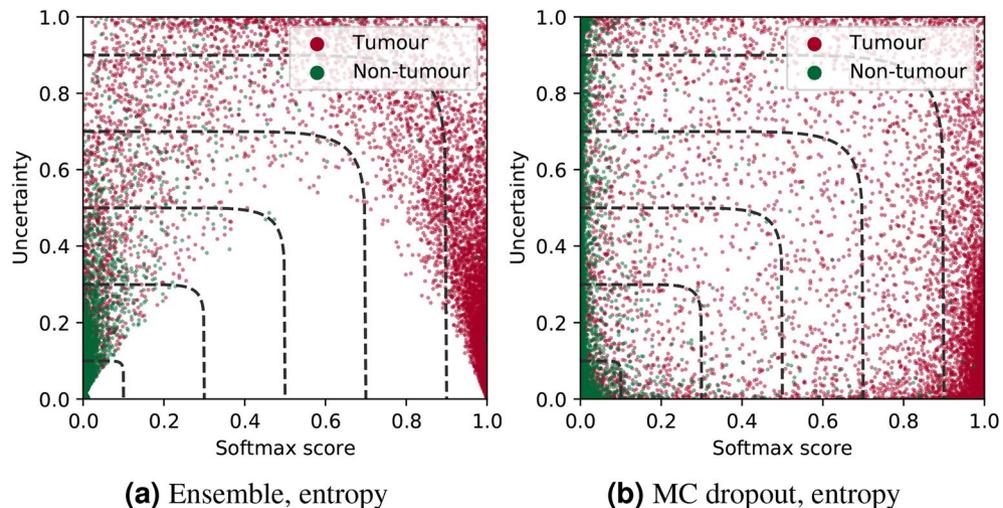**(a)** Ensemble, entropy   **(b)** MC dropout, entropy

Figure 3. Relation between softmax confidence and estimated uncertainty, for two different uncertainty methods. The points show the testset from Camelyon16, with colors encoding ground truth class labels. The dashed lines illustrate the 2D threshold used for classification based on both softmax and uncertainty, for a set of different threshold values.

ROC-AUCs				
	MC dropout	**Ensembles**	**TTA**	**Softmax score**
Sample mean uncertainty				
Camelyon16	0.908	0.979	0.976	0.975
BRLN	0.913	0.971	0.968	0.968
Sample variance				
Camelyon16	0.962	0.975	0.975	0.975
BRLN	0.955	0.968	0.968	0.968
Entropy				
Camelyon16	0.924	0.980	0.977	0.975
BRLN	0.924	0.972	0.969	0.968
Mutual information				
Camelyon16	0.954	0.978	0.976	0.975
BRLN	0.947	0.970	0.968	0.968

Table 3. ROC-AUCs on Camelyon16 and BRLN data: combining softmax tumour score from a single NN with uncertainty estimates. Softmax score refers to using the softmax output alone.

Table 3 summarises the ROC-AUC results on Camelyon16 and BRLN data, for different combinations of uncertainty metrics and methods. We can see that MC dropout is the only method that, independently of metric and data set, achieves worse ROC-AUC scores than the softmax score from a single NN. TTA and deep ensembles exhibit nearly identical performance for each computed metric. The uncertainty-including methods consistently perform at par or better than the baseline of using the softmax score from a single NN, but the improvement is small, below 1 percentage point in terms of ROC-AUC.

Although the ROC-AUC results are similar compared to the traditional ensemble technique (Figs. 1 and 2), another aspect of robustness is how the performance of methods and metrics varies across the range of classification thresholds. In Fig. 4 we can see that when using the sample mean or entropy uncertainty, the shape of the accuracy versus classification threshold curves are considerably different for Camelyon16 data. Instead of a narrow range of peak accuracy, we get high performance over a broader range of thresholds. This indicates that embedding uncertainty information can lessen the sensitivity for how the operating point of the prediction is set, which is one part of the generalisation challenge. Importantly, this finding holds true also under domain shift (Fig. 5).

Misprediction detection. In addition to embedding uncertainty information in the prediction, a straightforward application of the uncertainty estimates is in misprediction detection. Performance for this task also provides a general idea about the capacity of the methods to boost robustness in a deployed diagnostic tool. In this work, we only evaluated how well the methods can detect mispredictions without determining what is the best approach of incorporating this information in the clinical setting. For example, in order to improve perfor-
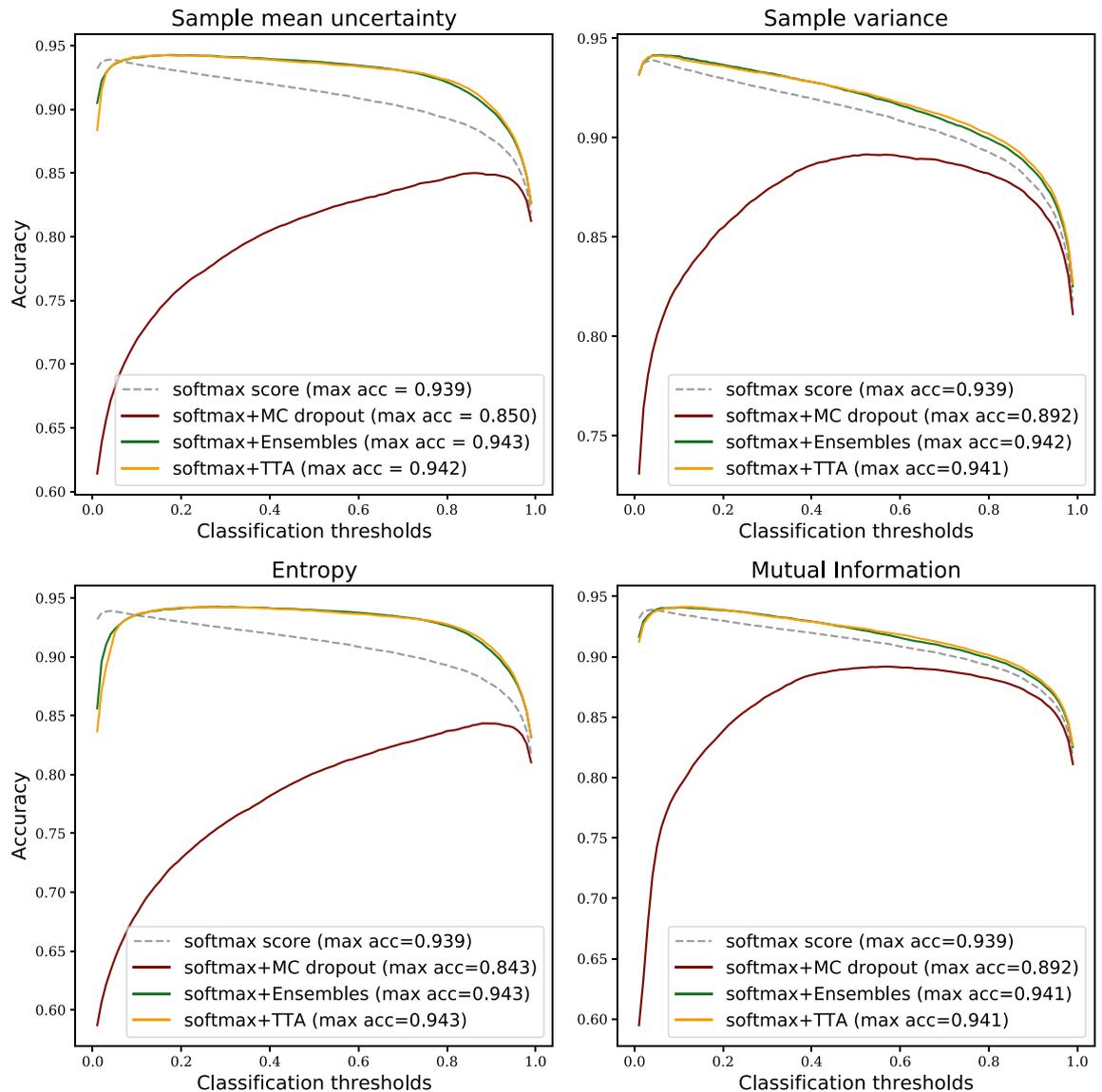
Figure 4. Prediction accuracy on Camelyon16 data when combining softmax tumour score from a single NN with uncertainty estimates.

mance, one could omit the detected mispredictions or adjust their predicted labels, but we leave this direction of research for future work.

We compare the three uncertainty estimation methods incorporating multiple predictions with a baseline uncertainty u_{base} derived from a single softmax value:

$$u_{\text{base}} = 1 - 2(s - 0.5)^2,$$

where s refers to the softmax output for the tumour class of a single NN. The baseline captures the general correlation between softmax score and uncertainty, where uncertainty is maximal at 0.5 and decreases towards 0 and 1, as seen in Fig. 3a.

Evaluating uncertainty methods. From the plots in Fig. 6 showing the ROC-AUC performance for the misprediction detection task, we observe the same tendency as in the experiment of boosting the general performance: ensembles and TTA are substantially better than MC dropout. In fact, MC dropout performs worse than the baseline independently of the chosen metric or classification threshold.

In Table 4, we see that the highest result for all classification thresholds was achieved by ensembles method. TTA performance is midway between the baseline and the ensembles. Comparing with Table 5, we see that domain shift affects misprediction performance in a negative way. Under the domain shift, only ensembles and TTA with sample mean uncertainty consistently achieve improvements over the baseline, whereas other combinations are at par with or below the baseline (see also Fig. 6b).
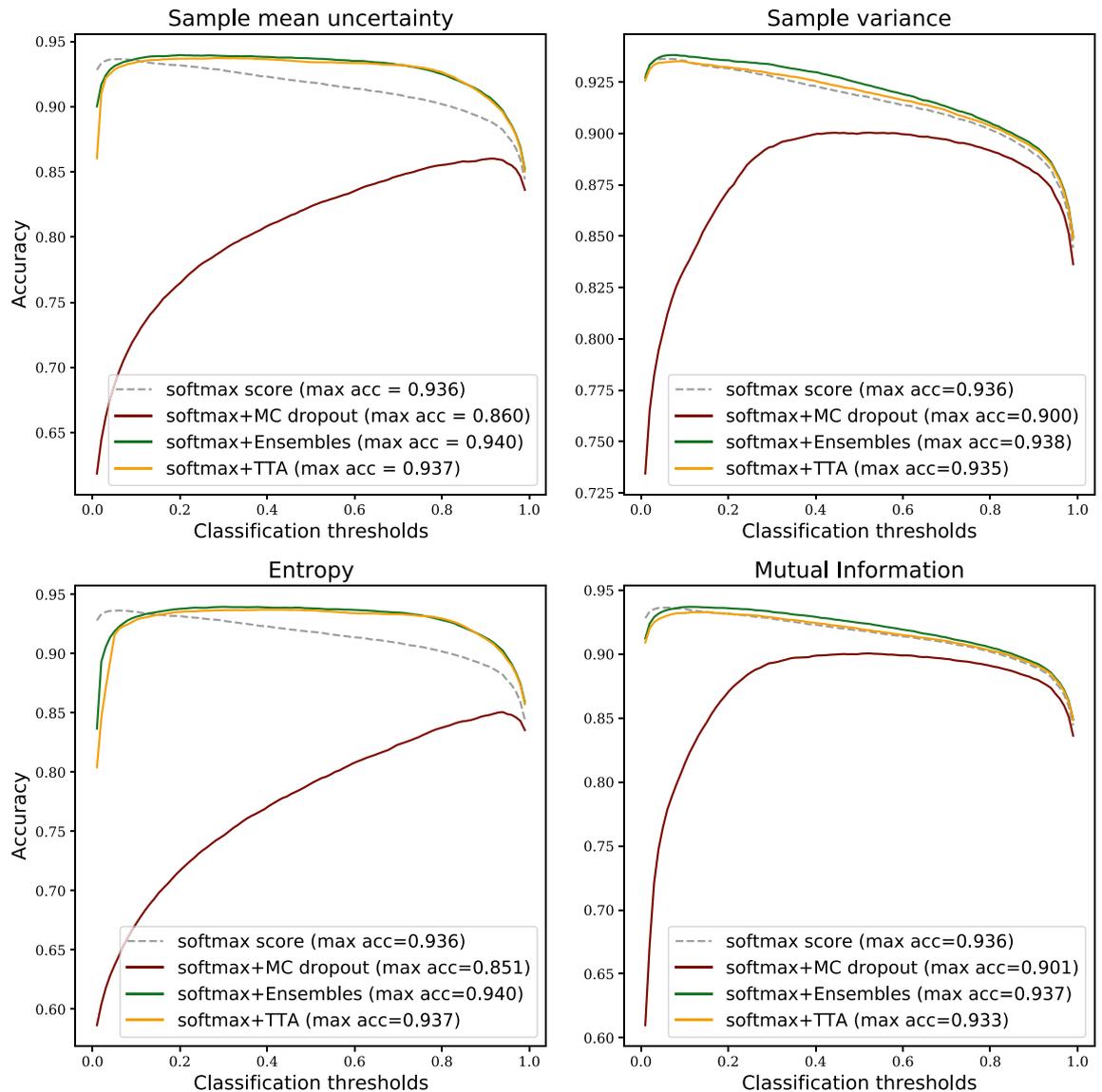
Figure 5. Prediction accuracy on BRLN data when combining softmax tumour score from a single NN with uncertainty estimates.

An interesting observation is that there is a trade-off between how good the uncertainty methods are at misprediction detection and how well the NN performs on its primary task of cancer metastases detection. For higher threshold values, the predictive accuracy of the NN decreases, but the misclassification detection effectiveness increases (Fig. 6). This may suggest that uncertainty estimation is more beneficial for models with weaker predictive performance.

Evaluating uncertainty metrics. In the experiments we also compared the four uncertainty metrics. In Fig. 6a, we observe that on the in-domain data all metrics achieve similar good performance compared to the baseline, when computed from TTA or deep ensembles predictions. From Table 4, MI emerges as the best performing metric, closely followed by the other three.

Sample variance, entropy and MI metric do not generalise well under the domain shift. From Figs. 5 and 6b, we see that sample mean uncertainty is the only metric that performs better than the baseline independently of the classification threshold on the BRLN data (for ensembles and TTA).

Uncertainty and lobular carcinoma. Now we turn to evaluating if uncertainty measures may contribute to boosting the performance on a rare type of data, in our case: lobular carcinoma.

For this experiment, we focus on the consistent good performers in previous experiments: the sample mean uncertainty metric combined with the ensembles and TTA uncertainty estimation methods.
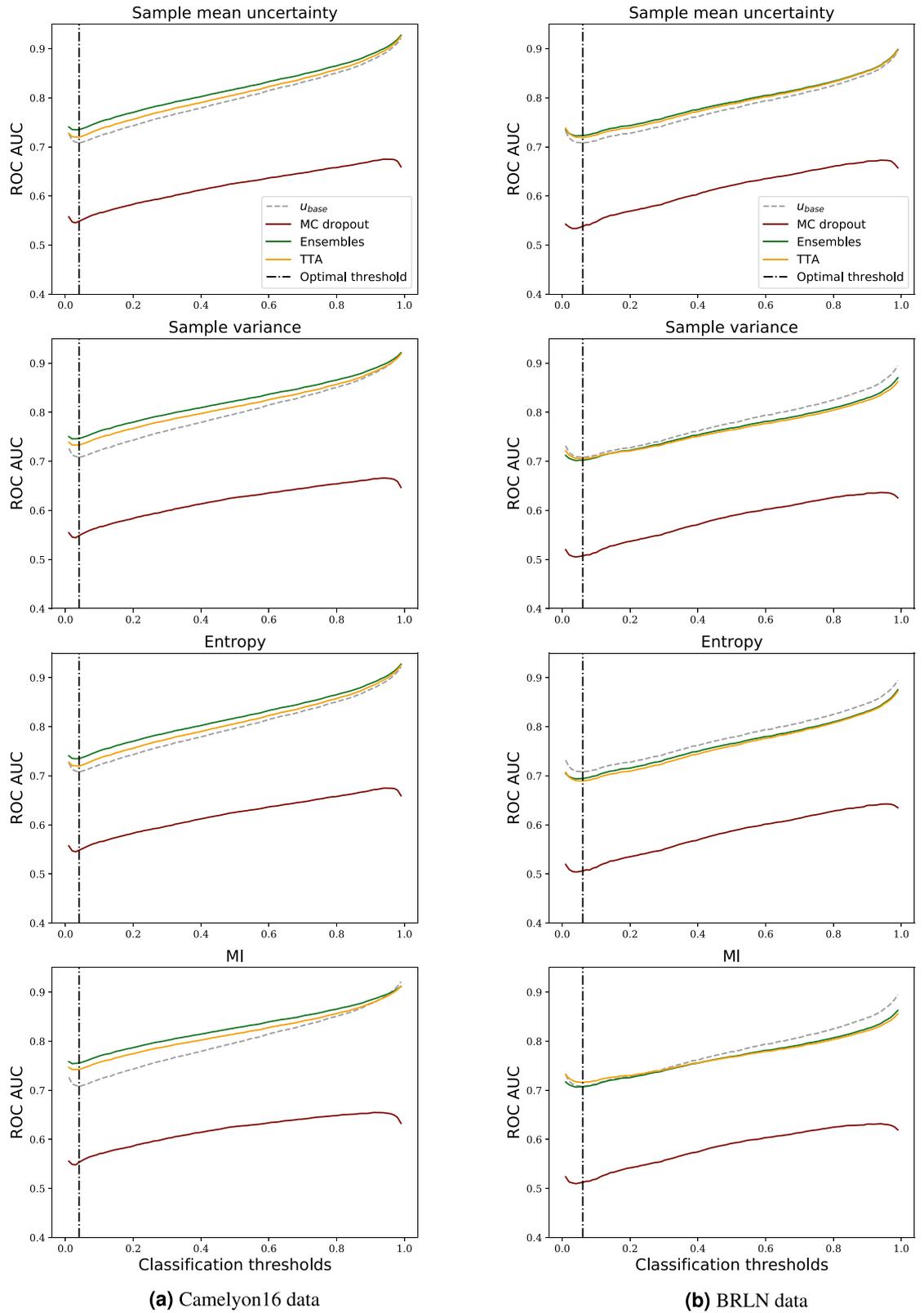
9

(a) Camelyon16 data

(b) BRLN data

Figure 6. ROC-AUCs of misprediction detection on Camelyon16 (in-domain) and BRLN (domain shift) data sets for different thresholds used to differentiate between tumour and non-tumour predictions. The softmax-based baseline uncertainty is the same in all plots.

	MC dropout	Ensembles	TTA	Baseline	Classification Accuracy
Sample mean					
threshold 0.1	0.566	0.751	0.736	0.723	0.935
threshold 0.5	0.626	0.818	0.807	0.798	0.914
threshold 0.9	0.672	**0.891**	0.885	0.879	0.875
Sample variance					
threshold 0.1	0.567	0.762	0.749	0.723	0.935
threshold 0.5	0.627	0.824	0.812	0.798	0.914
threshold 0.9	0.665	0.889	0.882	0.879	0.875
Entropy					
threshold 0.1	0.566	0.751	0.736	0.723	0.935
threshold 0.5	0.626	0.818	0.807	0.798	0.914
threshold 0.9	0.672	**0.891**	0.885	0.879	0.875
Mutual information					
threshold 0.1	0.571	**0.770**	0.757	0.723	0.935
threshold 0.5	0.626	**0.828**	0.816	0.798	0.914
threshold 0.9	0.655	0.886	0.879	0.879	0.875

Table 4. Camelyon16 data: ROC-AUCs of misprediction detection for varying classification thresholds. The highest achieved values per classification threshold are in bold.

	MC dropout	Ensembles	TTA	Baseline	Classification Accuracy
Sample mean uncertainty					
threshold 0.1	0.547	**0.728**	0.724	0.712	0.935
threshold 0.5	0.622	**0.791**	0.788	0.779	0.918
threshold 0.9	0.671	**0.857**	0.856	0.850	0.889
Sample variance					
threshold 0.1	0.514	0.707	0.710	0.712	0.935
threshold 0.5	0.590	0.768	0.764	0.779	0.918
threshold 0.9	0.635	0.831	0.826	0.850	0.889
Entropy					
threshold 0.1	0.513	0.700	0.694	0.712	0.935
threshold 0.5	0.588	0.766	0.761	0.779	0.918
threshold 0.9	0.640	0.833	0.831	0.850	0.889
Mutual information					
threshold 0.1	0.519	0.712	0.719	0.712	0.935
threshold 0.5	0.592	0.769	0.768	0.779	0.918
threshold 0.9	0.631	0.827	0.823	0.850	0.889

Table 5. BRLN data: ROC AUCs of misprediction detection for varying classification thresholds. The highest achieved values per classification threshold are in bold.

ROC-AUC			
	Ensembles	TTA	Baseline
Sample mean uncertainty			
Lobular data	0.908	0.899	0.899
Ductal data	0.984	0.983	0.982

Table 6. Tumour metastases detection on Lobular and Ductal data: ROC AUCs of combination of sample means uncertainty and the softmax score.
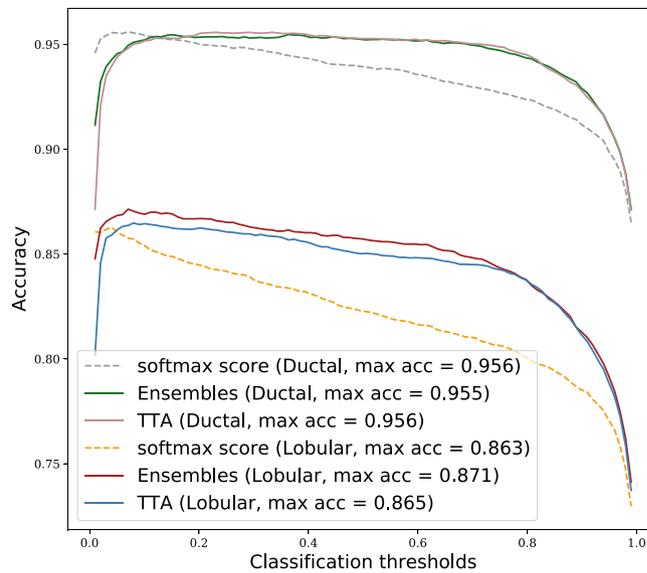
Figure 7. Prediction accuracy on Lobular and Ductal data when combining softmax tumour score from a single NN with sample means uncertainty estimated by ensembles and TTA methods.
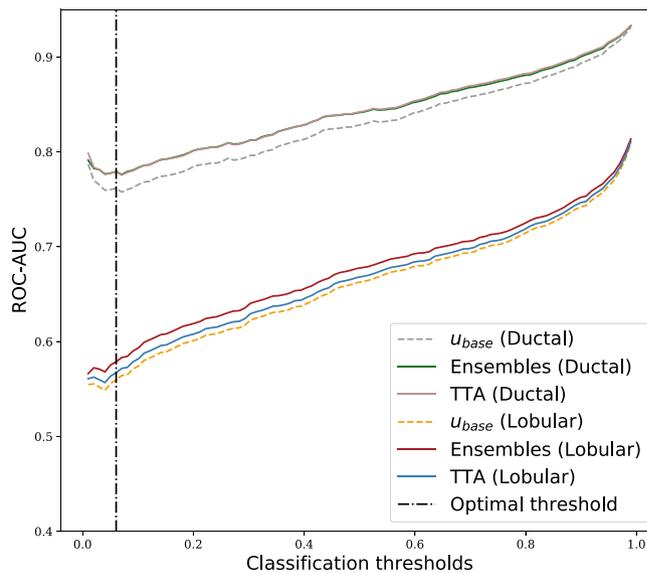

Figure 8. ROC-AUCs of misprediction detection by sample means uncertainty from ensembles and TTA (for varying classification threshold). Baseline is computed from softmax score of a single NN.

Uncertainty for boosting the tumour metastases detection. In Table 6, we see similar results as for the entire BRLN dataset: the ROC-AUCs improve marginally by combining the uncertainty with the softmax score, slightly more improvement for the lobular data. Fig. 7 shows the previously noted effect of a flattened accuracy curve, where the accuracy increase for suboptimal thresholds is more pronounced for the lobular data set.

Uncertainty for misprediction detection. From Fig. 8 we conclude that all methods are substantially better at detecting mispredictions on the ductal cancer subtype than the lobular, meaning that this type of domain shift also has a negative effect on misprediction performance. For the optimal classification threshold, the misprediction performance on lobular data is not much better than a random guess.

From Table 7 we see that the improvement from the baseline for the best performing uncertainty estimation method is similar on ductal and lobular data.

ROC-AUC				
				Classification
	Ensembles	**TTA**	**Baseline**	**Accuracy**
Lobular data				
threshold 0.1	**0.593**	0.581	0.574	0.854
threshold 0.5	**0.678**	0.669	0.663	0.823
threshold 0.9	0.754	**0.756**	0.744	0.784
Ductal data				
threshold 0.1	**0.784**	0.783	0.766	0.954
threshold 0.5	**0.842**	**0.842**	0.829	0.939
threshold 0.9	0.903	**0.904**	0.896	0.910

Table 7. ROC-AUCs of misprediction detection by sample means uncertainty computed from ensembles and TTA methods. The highest achieved values per classification threshold are in bold.

Discussion

The main research question was whether uncertainty estimates can add to the predictive capacity of DL in digital pathology. The results show that uncertainty indeed adds value if good measures and metrics are chosen. The predictive performance can be slightly increased, but a perhaps more important benefit is a lessened sensitivity to the choice of classification threshold—mitigating the infamous AI 'brittleness'. Uncertainty used for misprediction detection is valuable in the sense that performance is far above a random guess. The results also show, however, that the added value of introducing uncertainty over softmax probability is quite limited and it is an open question whether these benefits would make a substantial difference when employed in a full DL solution in a clinical setting.

Drilling down into detailed results, it is clear from the experiments that MC dropout is the least suitable method as the variability in its output has minimal value for boosting the general NN's performance directly or via misprediction detection. This is also apparent from inspecting the relation between softmax confidence and MC dropout uncertainty in Fig. 3b, which show little correlation. In contrast, the TTA and deep ensemble methods outperformed the baseline on both evaluation tasks for most metrics. While deep ensembles exhibited the best performance, the difference to TTA was often negligible. Thus, if the flexibility offered by using a model-agnostic method is important in the scenario considered, TTA could be preferred.

Interestingly, the gains of using ensembles or TTA were larger for the classification thresholds corresponding to high accuracy, at least for the most well-performing metrics. Furthermore, our results demonstrate that misprediction detection is easier when classification is poor. This underlines that misprediction detection should not be considered in isolation, instead the interplay with classification accuracy should always be considered.

The choice of uncertainty metric is not trivial. In our experiments, entropy and sample mean uncertainty can be said to have achieved the best results overall, but the differences are small between all metrics. It is a somewhat surprising result that a mean aggregation performs at par with a metric taking variance into account.

In the out-of-domain experiments we saw some reduction in the performance gains from all combinations of uncertainty estimation methods and metrics. While this is consistent with previous work[30], it is discouraging, as the foremost objective of these approaches is to mitigate the generalisation problem. It seems that the variation of model output is not that different between in-domain and out-of-domain pathology data. In fact, only the sample mean uncertainty sustains a better performance than the simple softmax-based baseline in the out-of-domain case, and the baseline showed the least drop in performance due to domain shift. This is somewhat surprising, as we would have expected the softmax baseline to be more sensitive to domain shift. The reason is likely both that we deal with a smaller, clinically realistic, domain shift and that softmax can behave better than expected in out-of-domain situations[27]. The upside of this result is that even a simple uncertainty measure can exhibit a reasonable performance on misprediction detection.

In the study of detecting mispredictions within a data subtype that is underrepresented in the training set (lobular carcinoma), we observed that uncertainty methods and the baseline are much less effective at this compared to the abundant data subtype (ductal carcinoma). The performance gains from using ensemble and TTA uncertainty estimation had larger margin for the classification thresholds corresponding to the highest accuracy, but less than on the in-domain data.

One of the limitations of this work is that we worked with patches extracted from WSIs. This was essential to investigate the basic properties of the uncertainty in digital pathology, but a study on how this translates to WSI level decisions is necessary. Furthermore, we focused on breast cancer metastases detection in the lymph nodes. More studies should be carried out to confirm that the results hold in other digital pathology applications. Of particular interest is to study prediction tasks with lower accuracy, where our results indicate that the added value of uncertainty may be greater than in this work. Regarding TTA, there may be other types of augmentations that are better suited to the specific objective of estimation of predictive uncertainty. There are also other method parameter options that could be relevant to evaluate. The dropout probability chosen for MC dropout may, for instance, not be optimal for our ResNet18 architecture, but we argue (also in light of previous work) that it is unlikely that the MC dropout performance then would surpass the other methods.

A potential direction for future work could be to do some more extensive tuning of the uncertainty estimation methods. For example, exploring the effects of bagging, boosting or stacking techniques[44] on improving

the diversity of the models in an ensemble which could lead to better uncertainty estimates provided by the deep ensembles method. Alternatively, the focus could be placed on determining if a combination of several uncertainty estimation methods would result in an improved performance.

Conclusion

We conclude that the evaluated uncertainty methods and metrics perform well on in-domain data but are affected by the domain shift due to new medical center as well as the underrepresented subtypes of data in the training set. The softmax score of the target NN can be transformed to provide an uncertainty measure which is less affected by the domain shift than the more established methods. The associated computational costs and NN design constraints indicate that the use of softmax score transformation is an appealing alternative to the uncertainty estimation methods.

Data availability

Camelyon16 dataset[40] generated during and/or analysed during the current study are available in the official GoogleDrive repository,which can be accessed using this web link. The BRLN dataset generated during and/or analysed during the current study are not publicly available due it being used in an ongoing another study but are available from the corresponding author on reasonable request. The dataset is planned to be made publicly available once the study is concluded.

References

1. Alzubaidi, L. *et al.* Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *J. Big Data* **8**, 1–74 (2021).
2. Koumakis, L. Deep learning models in genomics; are we there yet?. *Comput. Struct. Biotechnol. J.* **18**, 1466–1473. https://doi.org/10.1016/j.csbj.2020.06.017 (2020).
3. Babak, E. B. *et al.* Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210. https://doi.org/10.1001/jama.2017.14585 (2017).
4. Campanella, G. *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019).
5. Yun, L. *et al.* Artificial intelligence-based breast cancer nodal metastasis detection: Insights into the black box for pathologists. *Arch. Pathol. Lab. Med.* **143**, 859–868 (2019).
6. Ström, P. *et al.* Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Lancet Oncol.* **21**, 222–232 (2020).
7. Steiner, D. F. *et al.* Evaluation of the use of combined artificial intelligence and pathologist assessment to review and grade prostate biopsies. *JAMA Netw. Open* **3** (2020).
8. Pantanowitz, L. *et al.* An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. *Lancet Dig. Health* **2**, e407–e416 (2020).
9. Wang, Y. *et al.* Improved breast cancer histological grading using deep learning. *Ann. Oncol.* (2021).
10. Kumar, N., Gupta, R. & Gupta, S. Whole slide imaging (WSI) in pathology: current perspectives and future directions. *J. Dig. Imag.* (2020).
11. Wouter, M. K. An introduction to domain adaptation and transfer learning. *ArXiv* **abs/1812.11806** (2018).
12. Tellez, D. *et al.* Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med. Image Anal.* **58** (2019).
13. Stacke, K., Eilertsen, G., Unger, J. & Lundstrom, C. Measuring domain shift for deep learning in histopathology. *IEEE J. Biomed. Health Inform.* **2**, 325 (2021).
14. Li, C. I., Anderson, B. O., Daling, J. R. & Moe, R. E. Trends in incidence rates of invasive lobular and ductal breast carcinoma. *JAMA* **289**, 1421–1424. https://doi.org/10.1001/jama.289.11.1421 (2003).
15. Dossus, L. & Benusiglio, P. Lobular breast cancer: incidence and genetic and non-genetic risk factors. *Breast Cancer Res.* **17** (2015).
16. Gal, Y. & Ghahramani, Z. Dropout as a bayesian approximation: representing model uncertainty in deep learning. In *Proceedings of the 33rd international conference on machine learning, ICML 2016*, vol. 3, 1651–1660 (2016). arXiv:1506.02142v6.
17. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In Guyon, I. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., 2017).
18. Camarasa, R. *et al.* Quantitative comparison of monte-carlo dropout uncertainty measures for multi-class segmentation. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*, 32–41 (Springer, Cham, 2020).
19. Linmans, J., van der Laak, J. & Litjens, G. Efficient out-of-distribution detection in digital pathology using multi-head convolutional neural networks. In Arbel, T. *et al.* (eds.) *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, vol. 121 of *Proceedings of Machine Learning Research*, 465–478 (PMLR, 2020).
20. Nair, T., Precup, D., Arnold, D. L. & Arbel, T. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Med. Image Anal.* **59** (2020).
21. Ayhan, M. S. & Berens, P. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *Medical Imaging with Deep Learning (MIDL)*, Midl, 1–9 (2018).
22. Pocevičiūtė, M., Eilertsen, G. & Lundström, C. *Survey of XAI in Digital Pathology.* (Springer, New York, 2020).
23. Mariet, E. Z., Jenatton, R., Wenzel, F. & Tran, D. Distilling ensembles improves uncertainty estimates. In *Third symposium on advances in approximate bayesian inference* (2021).
24. Osband, I., Blundell, C., Pritzel, A. & Van Roy, B. Deep exploration via bootstrapped dqn. In *Advances in neural information processing systems*, 4033–4041 (2016).
25. Tagasovska, N. & Lopez-Paz, D. Single-model uncertainties for deep learning. In Wallach, H. *et al.* (eds.) *Advances in neural information processing systems 32*, 6414–6425 (Curran Associates, Inc., 2019).
26. Wenzel, F. *et al.* How good is the Bayes posterior in deep neural networks really? In III, H. D. & Singh, A. (eds.) *Proceedings of the 37th International Conference on Machine Learning*, vol. 119 of *Proceedings of Machine Learning Research*, 10248–10259 (PMLR, 2020).
27. Pearce, T., Brintrup, A. & Zhu, J. Understanding softmax confidence and uncertainty (2021). arXiv:2106.04972.

28. Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P. H. S. & Gal, Y. Deterministic neural networks with inductive biases capture epistemic and aleatoric uncertainty (2021). arXiv:2102.11582.
29. Kyono, T., Gilbert, J. F. & van der Schaar, M. Improving workflow efficiency for mammography using machine learning. *J. Am. Coll. Radiol.* **17**, 56–63. https://doi.org/10.1016/j.jacr.2019.05.012 (2020).
30. Thagaard, J. *et al.* Can you trust predictive uncertainty under real dataset shifts in digital pathology?. In *Lecture notes in computer science*, 824–833 (2020).
31. Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations* (2018).
32. Lee, S., Purushwalkam, S., Cogswell, M., Crandall, D. J. & Batra, D. Why M heads are better than one: Training a diverse ensemble of deep networks. *CoRR* **abs/1511.06314** (2015). arXiv:1511.06314.
33. Graham, S. *et al.* Mild-net: minimal information loss dilated network for gland instance segmentation in colon histology images. *Med. Image Anal.* **52**, 199–211 (2019).
34. Wang, G. *et al.* Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing* **338**, 34–45 (2019).
35. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
36. Gal, Y., Islam, R. & Ghahramani, Z. Deep Bayesian active learning with image data. In Precup, D. & Teh, Y. W. (eds.) *Proceedings of the 34th International Conference on Machine Learning*, vol. 70 of *Proceedings of Machine Learning Research*, 1183–1192 (PMLR, 2017).
37. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Computer Vision and Pattern Recognition (CVPR)* 770–778 (2016).
38. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, vol. 2015 International Conference on Computer Vision, ICCV 2015, 1026–1034 (Microsoft Research, 2015).
39. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
40. Litjens, G. *et al.* 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience* **7**, https://doi.org/10.1093/gigascience/giy065 (2018).
41. Jarkman, S. *et al.* Axillary lymph nodes in breast cancer cases. https://doi.org/10.23698/aida/brln (2019).
42. Bekkar, M., Djema, H. & Alitouche, T. Evaluation measures for models assessment over imbalanced data sets. *J. Inform. Eng. Appl.* **3**, 27–38 (2013).
43. Powers, D. M. W. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *Int. J. Mach. Learn. Technol.* 37–63 (2011).
44. Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M. & Suganthan, P. N. Ensemble deep learning: a review. *ArXiv* **abs/2104.02395** (2021).

Acknowledgements

Author contributions

M.P. developed the design and the code for the study, conducted the experiments, analysed the results, and wrote the main manuscript text; G.E. analysed the results and supervised the project; S.J. collected the data and consulted on medical questions; C.L. developed the design of the study and supervised the project. All authors reviewed the manuscript.

Funding

Competing interest

Claes Lundström is an employee and shareholder of Sectra AB. Milda PocevicIute, Gabriel Eilertsen and Sofia Jarkman declare no competing financial and/or non-financial interests.

Additional information

Correspondence and requests for materials should be addressed to M.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.