Medical Image Analysis 83 (2023) 102655

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Predictive uncertainty estimation for out-of-distribution detection in digital pathology

Jasper Linmans [a,*], Stefan Elfwing [b], Jeroen van der Laak [a,c], Geert Litjens [a]

[a] Department of Pathology, Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, The Netherlands
[b] Inify Laboratories AB, Stockholm, Sweden
[c] Center for Medical Image Science and Visualization, Linköping University, Linköping, Sweden

ARTICLE INFO

Keywords:
Deep learning
Histopathology
Out-of-distribution detection
Uncertainty estimation
Ensemble diversity
Multi-heads

ABSTRACT

Machine learning model deployment in clinical practice demands real-time risk assessment to identify situations in which the model is uncertain. Once deployed, models should be accurate for classes seen during training while providing informative estimates of uncertainty to flag abnormalities and unseen classes for further analysis. Although recent developments in uncertainty estimation have resulted in an increasing number of methods, a rigorous empirical evaluation of their performance on large-scale digital pathology datasets is lacking. This work provides a benchmark for evaluating prevalent methods on multiple datasets by comparing the uncertainty estimates on both in-distribution and realistic near and far out-of-distribution (OOD) data on a whole-slide level. To this end, we aggregate uncertainty values from patch-based classifiers to whole-slide level uncertainty scores. We show that results found in classical computer vision benchmarks do not always translate to the medical imaging setting. Specifically, we demonstrate that deep ensembles perform best at detecting far-OOD data but can be outperformed on a more challenging near-OOD detection task by multi-head ensembles trained for optimal ensemble diversity. Furthermore, we demonstrate the harmful impact OOD data can have on the performance of deployed machine learning models. Overall, we show that uncertainty estimates can be used to discriminate in-distribution from OOD data with high AUC scores. Still, model deployment might require careful tuning based on prior knowledge of prospective OOD data.

1. Introduction

Applying Deep Neural Networks (DNNs) in medical imaging introduces challenges related to reliability and interpretability. Such high-stakes environments require DNNs to indicate when they are likely to be incorrect to prevent models from failing silently (Kompa et al., 2021). Correctly quantifying and communicating uncertainty would allow models to flag abnormalities and unseen patterns by signaling "I don't know". However, DNNs are known to produce overconfident predictions (Guo et al., 2017), which can have poor diagnostic consequences, especially on out-of-distribution (OOD) data (Ovadia et al., 2019). Model deployment in digital pathology is particularly challenging due to the heterogeneous, long-tailed data distribution with many pathological abnormalities and artifacts (Schömig-Markiefka et al., 2021). To this end, deep learning applications in digital pathology should report informative uncertainty estimates associated with predictions to flag OOD data and limit models from failing silently.

Due to its relevance to many safety–critical deep learning applications, there is much interest in estimating uncertainty associated with model predictions. A natural approach applies post hoc calibration of softmax values such that its output reflects the probability of the predicted class being correct (Guo et al., 2017). Although this method, referred to as temperature scaling, leads to well-calibrated predictions on held-out test data, it does not usually translate to better calibration on OOD data (Ovadia et al., 2019). Instead, more effective approaches of uncertainty estimation use statistics from a distribution of predictions using methods like Monte Carlo (MC) dropout (Gal and Ghahramani, 2016) or deep ensembles (Lakshminarayanan et al., 2017). Here, the spread of the predictive distribution determines the uncertainty. Although effective at estimating uncertainty in controlled conditions of computer vision datasets (Ovadia et al., 2019), these methods suffer from computational complexity requiring multiple training runs or forward passes at inference. To this end, recent work has developed multi-head ensembles (M-heads): a low-cost ensemble defined by a set of randomly initialized final layers of a network (Rupprecht et al., 2017; Lee et al., 2015; Osband et al., 2016; Linmans et al., 2020).

* Corresponding author.
E-mail address: jasper.linmans@radboudumc.nl (J. Linmans).

https://doi.org/10.1016/j.media.2022.102655
Received 23 July 2022; Received in revised form 26 September 2022; Accepted 12 October 2022
Available online 17 October 2022

The increasing variety of uncertainty estimation methods necessitates large-scale comparative research to better understand existing methods. For this reason, two independent works have recently benchmarked different methods on multiple datasets with favorable results for deep ensembles regarding the quality of the uncertainty estimates (Ovadia et al., 2019; Gustafsson et al., 2020). Although (Fort et al., 2019) attribute the success of deep ensembles to the diversity in predictions, ensembling methods that optimize for diversity like M-heads (Rupprecht et al., 2017) are often lacking from these benchmarks. Furthermore, current benchmarks of uncertainty estimation are often limited to controlled conditions of computer vision datasets with unrealistic or artificial OOD data, such as data from a completely different dataset. Large-scale comparison on more heterogeneous data relevant to real-world deployment is currently lacking.

In contrast to conventional uncertainty benchmark datasets, histopathology data suffers from subtle abnormalities that pose a more significant challenge than complete domain shifts. Prior work has shown promising results on tasks like detecting breast cancer metastasis (Liu et al., 2017) or prostate tumor detection (Bulten et al., 2020). However, DNNs are often trained and evaluated on heavily curated datasets that do not capture the heterogeneous and long-tailed data distribution expected during model deployment. A point estimate prediction from a vanilla DNN might become unreliable in clinical practice (Schömig-Markiefka et al., 2021). Combined, the heterogeneity of histopathology and its adverse effects on deep learning applications signify the need for a suitable uncertainty estimation benchmark relevant to model deployment in digital pathology and the broader machine learning community.

This work addresses the challenge of OOD detection on large-scale, real-world, and clinically relevant histological datasets based on uncertainty estimates associated with model predictions. Uncertainty estimation methods have already been investigated and showed promising results in several medical fields, with recent examples in radiology (Nair et al., 2020), sonography (Karimi et al., 2019), ophthalmology (Leibig et al., 2017), and histopathology (Ianni et al., 2020). However, these works focused on correlating uncertainty estimates with prediction accuracy and did not evaluate the ability to detect OOD data. Related work also evaluated OOD detection in histopathology, but only on data originating from a completely different organ (Thagaard et al., 2020). In contrast, recent work on dermatology and radiology discriminates between two types of OOD data (Guha Roy et al., 2022; Graham et al., 2022). Specifically, *near-OOD*, where the outlier and inlier classes are highly similar, and *far-OOD*, where the outlier is more distinct from the training distribution (Winkens et al., 2020). Overall, evaluating uncertainty estimates for tasks such as OOD detection remains relatively unexplored in histopathology, especially for the more challenging near-OOD problem.

1.1. Contributions

In this paper, we quantitatively compare prevalent uncertainty estimation methods in digital pathology on the task of OOD detection on a whole-slide image (WSI) level. We evaluate models trained using proven network architectures and training regimes with state-of-the-art (SOTA) performance on different segmentation and classification tasks. We aim to better understand existing SOTA models, their applicability to different OOD problems in digital pathology, and the challenges related to model deployment in clinical practice.

We start by training models on lymph node tissue to detect breast cancer metastasis. Afterward, models will be evaluated in their OOD detection performance with whole-slide images containing lymph node tissue diagnosed with diffuse large B-cell lymphoma. When assessing sentinel lymph nodes for breast cancer metastasis, incidental discovery of lymphoma is rare: it is found in only 1.6% of cases (Fox et al., 2010). We consider this a near-OOD problem; lymphoma originates from the same tissue as the benign training class, making this task uniquely challenging. See Fig. 1a for a graphical representation of the different data distributions and example images. Incorrectly classifying unknown malignancies as benign during automated screening could have disastrous clinical consequences. Here, we demonstrate the potential of uncertainty-based approaches to prevent such issues.

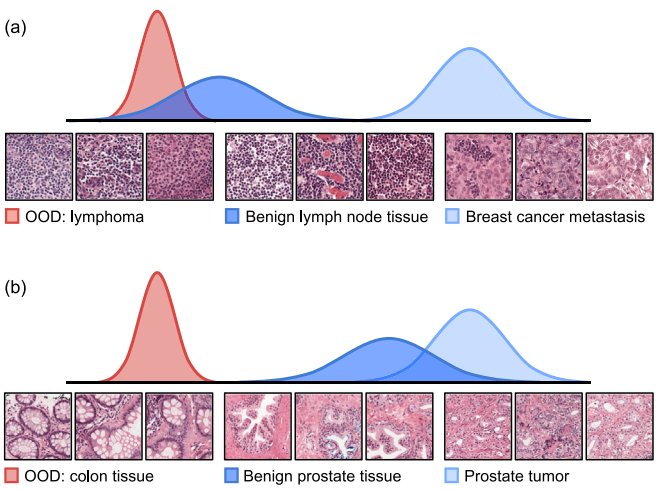

Fig. 1. Graphical representation of the data distributions used throughout this work corresponding to (a) the *near-OOD* detection task, where the OOD class (red) is highly similar to the training data (blue) and (b) the *far-OOD* detection task where the OOD class is more distinct from the training data.

Additionally, we train models to detect prostate tumor in biopsies and evaluate the uncertainty estimates on a dataset containing far-OOD samples, see Fig. 1b. Specifically, these models will be evaluated on prostate biopsies containing colon tissue, which we demonstrate to negatively impact tumor detection performance. Finally, for a more extreme case of far-OOD detection, we will evaluate the prostate tumor detection models on lymph node tissue (and vice versa).

The key contributions of this study are:

- We propose a new way to adopt existing SOTA patch-based trained methods on a whole-slide image level and translate the corresponding uncertainty heatmap to a whole-slide score. We show that this score can detect small OOD regions within a WSI.
- While most related work evaluates OOD detection performance on artificial far-OOD data, we evaluate both near-OOD and far-OOD detection performance in digital pathology. Here we observe that results on conventional OOD benchmarks do not always translate to the medical domain.
- To analyze the effects of ensemble diversity on OOD detection, we train M-heads with a meta-loss function for different levels of diversity and evaluate its impact on OOD detection performance.

2. Materials

Let a dataset be noted as $\mathcal{D}_{train} = \{ (\mathbf{x}_n, y_n) \,|\, \mathbf{x}_n \in \mathcal{X}_{in}, y_n \in \mathcal{Y}_{in} \}_{n=1}^{N_{train}}$ where $\mathcal{X}_{in}$ and $\mathcal{Y}_{in}$ define the in-distribution input and target space for N_{train} amount of data samples. Here, we train various classifiers $p(y \,|\, \mathbf{x})$ and measure their performance on a held-out dataset $\mathcal{D}_{test} = \{ (\mathbf{x}_n, y_n) \,|\, \mathbf{x}_n \in \mathcal{X}_{in}, y_n \in \mathcal{Y}_{in} \}_{n=1}^{N_{test}}$, which we refer to as the *target task* of the model. The goal is not to obtain new state-of-the-art results on the target task, but to perform comparable with the current SOTA while simultaneously providing informative uncertainty estimates for predictions $p(y \,|\, \mathbf{x})$. To evaluate the quality of the uncertainty estimates we measure the OOD detection performance on a dataset containing out-of-distribution samples $\mathcal{D}_{out} = \{ \mathbf{x}_n \,|\, \mathbf{x}_n \in \mathcal{X}_{out} \}_{n=1}^{N_{out}}$ where we expect high uncertainty. Here, the task is to discriminate between samples from $\mathcal{D}_{out}$ and a dataset $\mathcal{D}_{in}$, containing data similar to $\mathcal{D}_{train}$, based on the uncertainty estimates of predictions $p(y \,|\, \mathbf{x})$.

Table 1
Summary of the histopathology datasets that are used in this study. First, we designate 'target task'-specific train $\mathcal{D}_{train}$ and test $\mathcal{D}_{test}$ datasets for breast cancer metastasis detection (C16) and prostate cancer detection (PANDA, Prostate-520 and Prostate-colon). These datasets are use to assess the target task performance relative to OOD detection performance. Subsequently, we define the in-distribution $\mathcal{D}_{in}$ and out-of-distribution $\mathcal{D}_{out}$ datasets, where the CAMELYON datasets (C16 and C17) are considered in-distribution and lymphoma and Prostate-520 are considered OOD, for the lymph node metastases detection models. For the prostate cancer detection models, Prostate-520 is considered in-distribution and Prostate-colon and C16 are considered OOD. Last, we also separately designate subsets of cases that do not contain cancer, the benign subsets.

$\mathcal{D}_{train}$		$\mathcal{D}_{test}$		$\mathcal{D}_{in}$		$\mathcal{D}_{out}$	
Dataset	#WSIs	Dataset	#WSIs	Dataset	#WSIs	Dataset	#WSIs
C16 train set (lymph node tissue)	270	C16	129	C16 C16 (benign) C17 C17 (benign)	129 80 200 66	B-cell lymphoma Prostate-520	26 520
PANDA train set (prostate tissue)	4824	Prostate-520 Prostate-colon	520 112	Prostate-520 Prostate-colon	520 241	Prostate-colon Prostate-colon (benign) C16	112 36 129

To compare this work with related work, we start by evaluating the different uncertainty estimation methods on a conventional computer vision task: train on CIFAR10 (Krizhevsky et al., 2009) and perform inference on SVHN (Netzer et al., 2011), a far-OOD task. Here, the goal is to discriminate between in-distribution images from the CIFAR10 test set and OOD images from the SVHN dataset.

2.1. Lymph node tissue datasets

For the lymph node tissue experiments, we train models to detect breast cancer metastasis in whole-slide images (WSIs) of sentinel lymph node resections using data from the Camelyon16 (C16) challenge (Ehteshami Bejnordi et al., 2017; Litjens et al., 2018). We use the training set $\mathcal{D}_{train}$ (270 WSIs) and test set $\mathcal{D}_{test}$ (129 WSIs) as defined by the challenge organizers. Here, models are trained on randomly selected patches with size 279×279 from a $20\times$ resolution with a pixel spacing of 0.48 μm and a label corresponding to the center pixel of the patch. Following previous work, patches were selected in a ratio of $4:1$ normal to tumor to reduce false-positive detections (Liu et al., 2017). A validation set of 54 WSIs is used to select model checkpoints for inference. To evaluate performance on the target task, we report the challenge metrics on the test-set: the FROC value for tumor localization and the area under the ROC curve for slide-level classification.

The OOD detection performance is evaluated using a dataset $\mathcal{D}_{out}$ containing anomalies not seen during training. In particular, we consider a total of 26 WSIs from 19 patients acquired from one of the centers used in the train set, containing lymph node tissue diagnosed with diffuse large B-cell lymphoma.

To compare the uncertainty estimates of these near-OOD samples with in-distribution samples independent from the training and test set, we also use dataset $\mathcal{D}_{in}$ with data from the Camelyon17 (C17) challenge (Bandi et al., 2018). Specifically, we selected the 200 WSIs acquired by the same centers used in the Camelyon16 dataset from the test set of Camelyon17.

2.2. Prostate tissue datasets

In the prostate tissue experiments, models are trained on the binary task of discriminating between benign and tumor tissue in prostate biopsies, using data from the PANDA challenge (Bulten et al., 2022). Specifically, we train each model on 4342 WSIs ($\mathcal{D}_{train}$) from the Radboud University Medical Center and use an additional 482 WSIs as a validation set to select model checkpoints for inference. This subset of slides contains all WSIs from the PANDA challenge with the same center of origin as the WSIs used during inference, excluding slides containing colon tissue. Here, slide-level colon annotations were provided by the challenge organizers based on manual labeling. Each model is evaluated on a test set $\mathcal{D}_{test}$ containing 520 images taken from the test set of Bulten et al. (2020) where we excluded 15 WSIs due to the presence of colon tissue. Here, the authors of this work examined every case in the test set individually to confirm the presence of colon tissue. We refer to this dataset as *Prostate-520*. During training we select patches with size 279×279 from a $10\times$ resolution with a pixel spacing of 0.96 μm. Furthermore, we select labels in a $2:3$ normal to tumor ratio where normal patches are selected in a ratio of $1:3$ non-epithelium to benign-epithelium.

To measure the quality of the uncertainty estimates, we evaluate each model on $\mathcal{D}_{out}$ defined by 97 WSIs acquired from 52 patients that were excluded from the original PANDA training set and the 15 WSIs excluded from the test set of Bulten et al. (2020). These 112 prostate biopsies were selected based on the presence of colorectal tissue, which co-occurs due to the transrectal procedure in about 15% of the routine histopathological practice (van den Bergh et al., 2009). Incidental discovery of rectal pathologies is found in 17% of prostate biopsies containing colorectal tissue (van den Bergh et al., 2009), emphasizing the importance of detecting OOD regions due to the possibility of rare disease events not seen during training. We refer to this set of WSIs as the *Prostate-Colon* dataset. See Table 1 for an overview of the different histopathology datasets used throughout this work.

3. Methods

In this study, we evaluate different uncertainty estimation methods for the task of OOD detection. We do so by evaluating and comparing the uncertainty of predictions $p(y \mid \mathbf{x})$ on both in-distribution and out-of-distribution data. This section describes these methods, see Fig. 2 for an overview.

3.1. Uncertainty estimation

To asses the predictive uncertainty for each model, we evaluate the entropy of the prediction:

$$\mathcal{H}[p(y \mid \mathbf{x}^*)] = -\sum_{c=1}^{C} p(y_c \mid \mathbf{x}^*) \log p(y_c \mid \mathbf{x}^*) \tag{1}$$

with C the number of classes used for training and $\mathbf{x}^*$ an unseen datapoint. Evaluating the uncertainty of the predictive distribution based on the entropy is widely adopted for classification methods (Ovadia et al., 2019). For regression, the equivalent uncertainty signal would be the variance across predictions.

3.2. Predictive uncertainty estimation methods

As a baseline, we train and evaluate a single CNN in all experiments. In an effort to improve the OOD detection performance of this model, we implement an uncertainty estimation technique based on temperature scaling (Guo et al., 2017). Here, the logit vector (the vector of non-normalized predictions) is rescaled using a single scalar parameter T, following $p(y \mid \mathbf{x}) = \sigma(\mathbf{z}/T)$ with σ the softmax function and $\mathbf{z}$ the logit vector. The temperature parameter, which is optimized with respect to

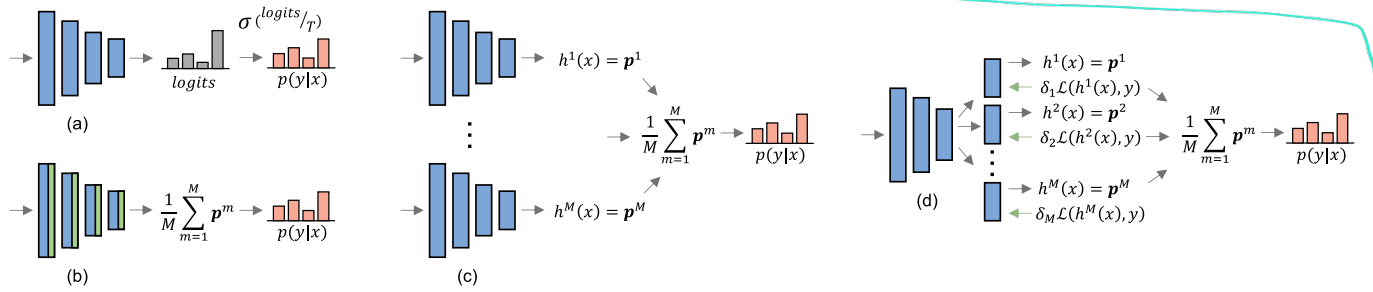

Fig. 2. Architectures of the different uncertainty estimation methods used throughout this work, based on classifiers $p(y|x)$. With (a) a single CNN and temperature scaling, using temperature T and softmax function σ, (b) MC dropout, with green blocks depicting dropout layers, (c) deep ensembles, and (d) M-heads. Arrows denote the flow of operation. Here, $\mathbf{p}^m$ refers to one of M softmax output vectors for methods (b, c, d) based on a distribution of individual predictions $h^m(x)$. For M-heads, a backward pass is made through each head based on the individual loss $\mathcal{L}(h^m(\mathbf{x}), y)$, weighted by δ_m (see Section 3.3).

the negative log-likelihood on the validation set, softens the softmax output for $T > 1$ such that its output better matches the probability of the predicted class being correct.

In addition to the single CNN model and temperature scaling, we also evaluate methods based on a distribution of predictions such as MC dropout and deep ensembles. Here, the final prediction is defined by averaging across the individual predictions, following:

$$p(y \mid \mathbf{x}^*) = \frac{1}{M} \sum_{m=1}^{M} p_{\theta_m}(y \mid \mathbf{x}^*) \tag{2}$$

where M defines the amount of individual predictions and θ_m the parameters used for the m'th prediction. Specifically, MC dropout is implemented by modifying the baseline architecture by adding spatial dropout (Tompson et al., 2015) at different locations in the network. At inference time, Monte Carlo sampling is used to produce 32 individual predictions for each unseen datapoint $\mathbf{x}^*$, with θ_m defining the parameters of the network after applying each individual dropout mask. Here, the entropy of the predictive mean (2) defines the uncertainty.

Similarly, for deep ensembles, the predictions of multiple independently trained vanilla networks are averaged at inference following (2). Here, θ_m defines the parameters of each individual member, and the entropy of the predictive mean defines uncertainty. Each member is trained on the entire dataset, with differences determined by random initialization which has shown to produce a diverse set of ensemble members in recent work (Fort et al., 2019). To evaluate the effects of the size of the ensemble on the performance, ensembles of size five and ten are used throughout the experiments.

3.3. Multi-head convolutional neural networks

To better match the high computational demands of digital pathology, we also implement multi-head ensembles, which mitigate the costs of multiple training and inference runs of the more prevalent deep ensembles. Here, parameters are shared in the early layers of a network, and the ensemble is defined by the heads: a set of randomly initialized final layers. The result is a single function which produces a set of hypotheses of size M, i.e. a function $\mathcal{X} \mapsto \mathcal{Y}^M$. In contrast to deep ensembles, the minimal memory requirements of M-heads enable parallel training of all members using a meta-loss function $\mathcal{M}$ to promote diversity (Rupprecht et al., 2017). Here, the meta-loss function acts on top of any standard loss, e.g. cross-entropy, for a single data-point $(\mathbf{x}, y)$:

$$\mathcal{M}(h(\mathbf{x}), y) = \sum_{m=1}^{M} \delta_m \mathcal{L}(h^m(\mathbf{x}), y) \tag{3}$$

with $\mathcal{L}$ the cross entropy loss and $h^m(\mathbf{x})$ the softmax output of the m'th head, for all M heads and such that

$$\delta_m = \begin{cases} 1-\epsilon & \text{if } m = \arg\min_i \mathcal{L}(h^i(\mathbf{x}), y). \\ \frac{\epsilon}{M-1} & \text{else.} \end{cases} \tag{4}$$

with ϵ the assignment relaxation constant. In other words, δ_m acts as a soft Kronecker delta such that the biggest fraction $1-\epsilon$ of the gradient signal flows through the winning head. Effectively increasing the learning rate for the winning head accordingly will promote specialization. Meanwhile, distributing the remaining loss between the other heads will improve the generalization of all heads to unseen data. To prevent issues of mode-collapse, when a single head dominates training, we add stochasticity by randomly dropping out predictions with a low probability $p_r = 0.05$ (Rupprecht et al., 2017). Note that δ_i is computed per sample, not per batch. Calculating this Kronecker delta per batch limits the upper bound of the batch size since the sample distribution in large batches approaches the full data distribution, crippling specialization. To evaluate the effects of increasing specialization, both $\epsilon = 0.6$ and $\epsilon = 0.1$ were used. Furthermore, to determine the importance of promoting diversity, we train M-heads with a fixed Kronecker delta: $\delta_m = 1/M$, effectively removing the specialization objective from training.

Hyperparameter settings for both p_r and ϵ were selected through tuning on the validation set during preliminary analysis on the lymph node dataset, to minimize issues of mode-collapse while maximizing specialization.

3.4. Patch based training

Throughout the experiments on histopathology, we adopt a fully convolutional approach (Long et al., 2015). This way, we can efficiently train a classifier $p(y \mid \mathbf{x})$ on patches extracted from WSIs during training and support larger inputs at inference. By simply replacing any fully connected layer with an equivalent convolutional layer, the network can operate on an input of arbitrary size. To prevent boundary artifacts caused by differences in padding, when increasing the input size at inference, we use valid padded convolutions for all tumor classification models. Doing so facilitates the use of patch-based classifiers on WSIs efficiently by applying the trained network in a sliding window approach on a WSI. The resulting tumor segmentation mask and its corresponding uncertainty can be evaluated on both a pixel and whole-slide image level. See Fig. 3 for an illustration of the fully convolutional approach.

3.5. Training and implementation details

Throughout all experiments, we train each method using a DenseNet architecture (Huang et al., 2017). Specifically, for the CIFAR10 experiment, we use one of the more parameter-efficient configurations, based on 100 layers with a growth rate of 12, and train using the originally proposed training regime (Huang et al., 2017). However, to better match the computational demands of histopathology datasets, we adopt a more lightweight DenseNet architecture to train the different tumor classification models, which has shown near SOTA performance on Camelyon16 in prior work (Linmans et al., 2020). Each model contains three dense blocks with eight valid padded convolutional layers. In

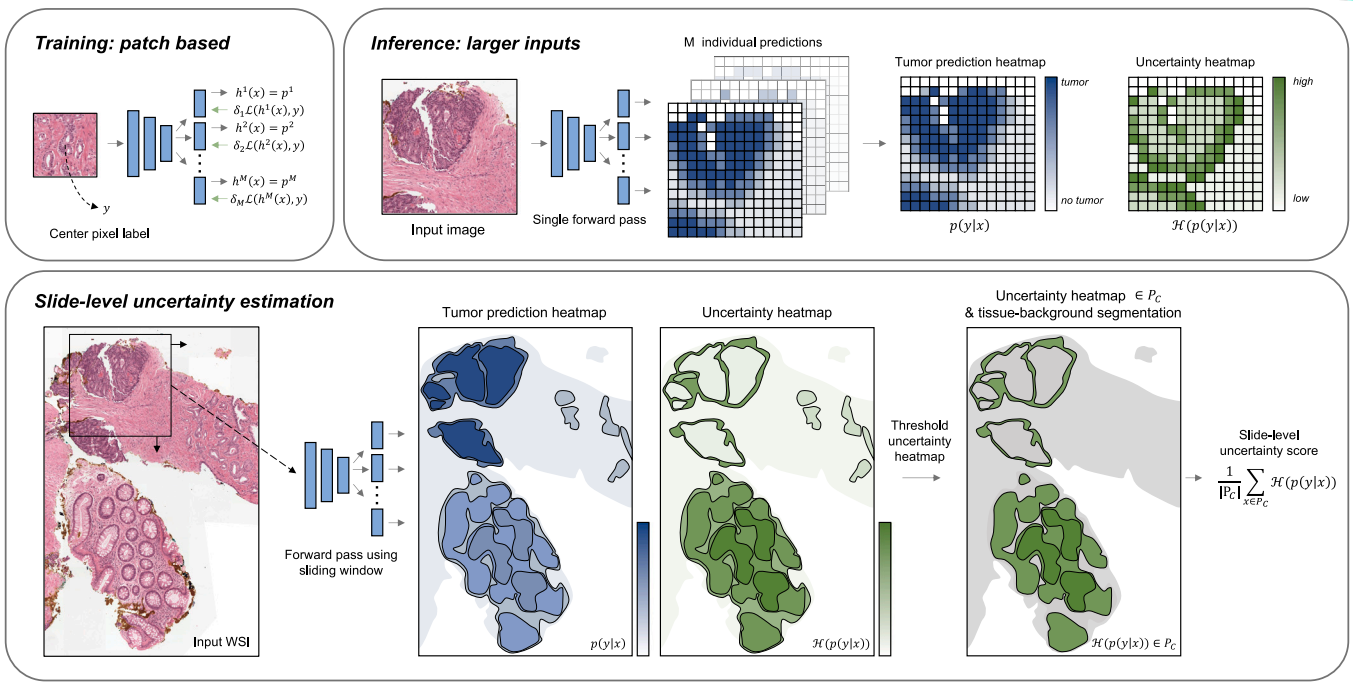

Fig. 3. Illustration of processes involved in training, inference, and slide-level uncertainty estimation, using M-heads as an example. Methods are trained patch-based to predict the center pixel label (*top left*). *Top right*: leveraging fully convolutional networks, we increase the input size during inference and use a sliding window approach to create pixel-level tumor prediction and corresponding uncertainty heatmaps (*bottom*). To determine a slide-level uncertainty score, we evaluate the mean entropy for the set of pixels P_C, containing all pixels with an uncertainty value higher then the C'th percentile of the uncertainty heatmap $\mathcal{H}[p(y \mid \mathbf{x}^*)]$.

total: 27 convolutional layers with 32 initial filters and a growth rate of 32. Each model is trained using the Adam optimizer (Kingma and Ba, 2015) for 90 epochs (defined as 2^{17} patches), with a batch size of 32 and an initial learning rate of $1e{-}4$ (multiplied by 0.1 at both one third and two thirds of training). Preliminary analysis, on a wide array of architectural variations, has shown that these settings lead to the best tumor detection performance on the Camelyon16 dataset at the lowest computational cost. Furthermore, a comparative analysis using the Camelyon17 dataset (Bandi et al., 2018), showed that the training pipeline is more important for the network's performance than the exact architecture, which is why we opted for a competitive but efficient network.

Specifically for MC dropout, the network architecture is modified by adding dropout layers at the end of each bottleneck layer in the DenseNet architecture with a dropout probability of $p = 0.1$. For M-heads, throughout all experiments, we define the depth of the head: the number of convolutional layers defining each head, to be equal to the final dense block of the DenseNet architecture (≈30% of the convolutional layers of the network).

3.6. Whole-slide level uncertainty estimation

By evaluating the entropy (1) value for every prediction $p(y|x)$, we end up with an uncertainty heatmap with equal dimensions to the tumor segmentation (see Fig. 3). This heatmap can help identify ambiguous regions within a WSI. However, discriminating between in-distribution and OOD data on a whole-slide level requires translating the pixel-level uncertainty to a slide-level uncertainty score.

To this end, we propose to perform spatial average pooling on the tissue region of the uncertainty heatmap to calculate a single score defining the slide-level uncertainty. We first apply a tissue-background segmentation algorithm to filter out the background for all WSIs used during inference (Bándi et al., 2019). Afterward, the average pixel-level uncertainty across the tissue is used to define the whole-slide level uncertainty, which we refer to as the whole-slide average (WSA) uncertainty. To allow the detection of smaller OOD regions, such as small colon tissue regions in the prostate datasets, we propose to evaluate a subset of the uncertainty heatmap. Specifically, we evaluate the values exceeding the C'th percentile of the uncertainty heatmap. We refer to this set, containing the highest uncertainty values per WSI, as P_C. Here, C is a hyperparameter of the uncertainty estimation pipeline. The resulting slide-level uncertainty estimate $\mathcal{H}[WSI]$ is defined by:

$$\mathcal{H}[WSI] = \frac{1}{|P_C|} \sum_{x^* \in P_C} \mathcal{H}[p(y \mid x^*)] \tag{5}$$

with x^* representing the individual pixels from the tissue region for the corresponding WSI. Specifically, we use the WSA score for the lymphoma detection experiments, meaning we average the uncertainty across the entire tissue area. For the colon detection experiments, we evaluate both WSA uncertainty and P_{99}, containing all pixels belonging to the 99% percentile of the uncertainty heatmap, and compare the results. Here, the 99% percentile was chosen based on prior knowledge on the prospective OOD data: colon tissue can be present in the form of very small tissue regions within the prostate biopsy. Using P_{99}, we aim to improve the detection of OOD regions which make up only roughly 1% of the total tissue area on the WSI.

Table 2
Target task performance and Out-of-Distribution detection results for the CIFAR10 experiment. Mean and standard deviations are reported for all methods across five independent runs except deep ensembles, which report confidence bounds by 2000-fold bootstrapping.

Model	$\mathcal{D}_{Test}$ (Acc.)	$\mathcal{D}_{Out}$ (AUC)
Single CNN	$94.46_{\pm.100}$	$90.50_{\pm1.47}$
Single CNN (MSP)	$94.46_{\pm.100}$	$90.17_{\pm1.41}$
Single CNN (MaxLogit)	$94.46_{\pm.100}$	$90.61_{\pm2.73}$
Temperature scaling	$94.14_{\pm.213}$	$92.47_{\pm2.09}$
MC dropout	$94.44_{\pm.139}$	$93.06_{\pm.732}$
Deep ensembles 5	95.82 [95.5, 96.2]	94.68 [94.5, 94.9]
Deep ensembles 10	**95.97** [**95.7**, **96.3**]	**95.08** [**94.9**, **95.3**]
M-heads 5 ($\epsilon = 0.1$)	$94.67_{\pm1.940}$	$91.82_{\pm1.75}$
M-heads 10 ($\epsilon = 0.1$)	$94.30_{\pm.185}$	$90.56_{\pm2.94}$

Table 3
Breast cancer metastasis detection performance (target task) on the camelyon16 (C16) test set and near-OOD detection performance: lymphoma detection in lymph node tissue (D_{out}). We measure in-distribution uncertainty estimates on C16 and C17, as well as only the benign cases. Mean and standard deviations of AUC values are reported for all methods across five independent runs except deep ensembles, which report confidence bounds by 2000-fold bootstrapping. M-heads models, trained with increasing levels of implicit head specialization, are included. Results from prior work on tumor detection performance (lacking results on OOD detection performance) are included for comparison.

Model	Target task (D_{Test})		OOD detection (AUC), $\mathcal{D}_{out}$ (B-cell lymphoma) vs. $\mathcal{D}_{in}$:			
	AUC	FROC	C16	C17	C16 (benign)	C17 (benign)
Single CNN	97.57$_{\pm 0.85}$	81.88$_{\pm 1.49}$	54.05$_{\pm 7.54}$	66.16$_{\pm 6.94}$	57.24$_{\pm 7.96}$	68.30$_{\pm 6.60}$
Temperature scaling	97.91$_{\pm 0.64}$	82.62$_{\pm 1.10}$	55.99$_{\pm 7.15}$	64.70$_{\pm 6.65}$	59.57$_{\pm 7.49}$	66.58$_{\pm 6.44}$
MC dropout	95.68$_{\pm 1.49}$	77.49$_{\pm 2.26}$	60.97$_{\pm 2.89}$	71.54$_{\pm 5.73}$	62.74$_{\pm 2.76}$	72.37$_{\pm 5.31}$
Deep ensembles 5	98.12 [95.7, 99.9]	**83.55** [**74.4, 91.9**]	57.81 [44.5, 70.4]	68.23 [57.3, 79.5]	61.03 [47.1, 73.6]	70.44 [59.2, 81.3]
Deep ensembles 10	**98.26** [**96.0, 99.9**]	82.81 [73.1, 91.5]	56.73 [42.7, 69.8]	67.50 [55.7, 78.8]	59.73 [45.5, 73.6]	69.68 [57.4, 80.7]
M-heads 5 ($\epsilon = 0.1$)	96.95$_{\pm 0.82}$	80.48$_{\pm 1.20}$	**70.73**$_{\pm \mathbf{0.97}}$	**81.02**$_{\pm \mathbf{0.65}}$	**73.05**$_{\pm \mathbf{1.09}}$	81.56$_{\pm 0.89}$
M-heads 10 ($\epsilon = 0.1$)	96.50$_{\pm 0.60}$	79.31$_{\pm 1.61}$	68.26$_{\pm 6.91}$	80.82$_{\pm 4.07}$	70.24$_{\pm 7.08}$	**82.30**$_{\pm \mathbf{4.21}}$
M-heads 5 ($\delta_m = 1/M$)	97.10$_{\pm 0.80}$	80.40$_{\pm 1.35}$	48.39$_{\pm 5.04}$	61.74$_{\pm 4.44}$	50.95$_{\pm 5.59}$	63.64$_{\pm 4.36}$
M-heads 5 ($\epsilon = 0.6$)	96.63$_{\pm 1.04}$	80.47$_{\pm 1.96}$	57.01$_{\pm 7.50}$	70.96$_{\pm 3.23}$	60.23$_{\pm 8.03}$	72.79$_{\pm 3.39}$
Liu et al. (2017)	98.6 [96.7, 100]	85.5 [81.0, 89.7]				
camelyon16 winner	99.4	80.7				
Pathologists, Ehteshami Bejnordi et al. (2017)	96.6	73.3				

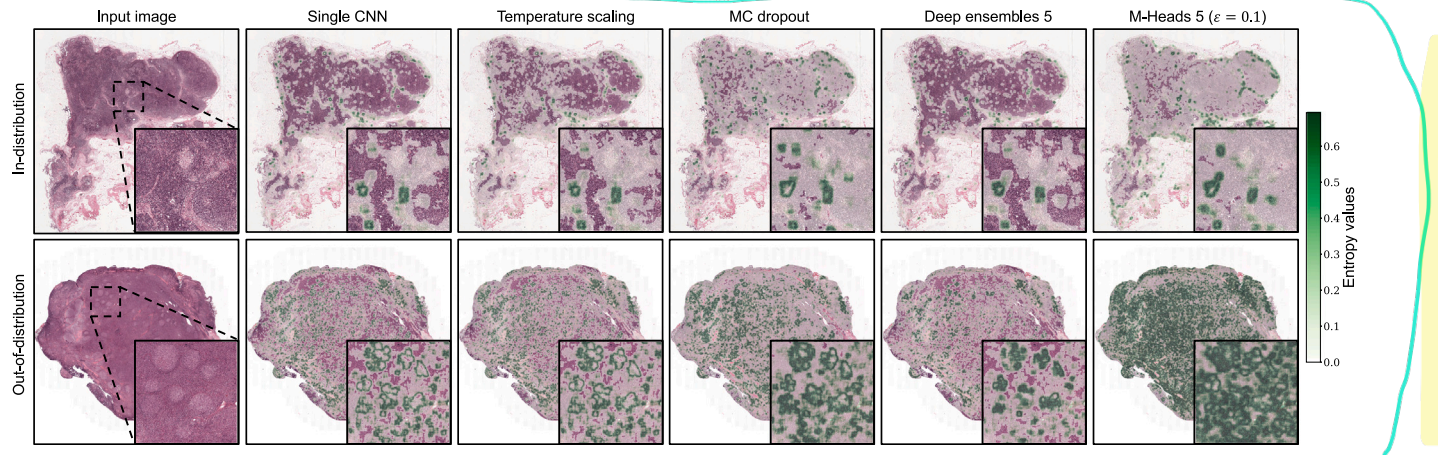

Fig. 4. Uncertainty heatmaps, based on entropy values (1), for benign lymph node tissue and OOD tissue diagnosed with diffuse large B-cell lymphoma.

3.7. OOD detection tasks and evaluation metrics

We evaluate each method using the datasets defined in Section 2 (See Table 1) on the ability to discriminate between samples from $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$ based on the slide-level uncertainty score. We start by evaluating the models trained on lymph node tissue on in-distribution data from the Camelyon challenges and near-OOD data from the lymphoma dataset. For the models trained on prostate tissue, we evaluate their performance using the Prostate-520 and Prostate-Colon datasets as in-distribution and far-OOD respectively. Finally, for a more extreme case of far-OOD data, in line with more conventional computer vision tasks like the CIFAR10 experiment, we also evaluate the prostate tumor detection models on lymph node tissue and vice versa.

During potential pre-screening on a patient or biopsy level, WSIs can be selected for further analysis based on the tumor prediction. However, it might be equally important to flag OOD samples within the set of remaining benign cases due to the potential for rare disease events. Therefore, we also repeat each OOD detection experiment using only the benign cases.

To compare the uncertainty estimates between the different in-distribution and OOD datasets, we perform ROC analysis based on the slide-level uncertainty score (3.6). Doing so, we analyze the ability to discriminate in-distribution and OOD samples independent from the exact decision threshold. For all experiments, we report the mean and standard deviation values for each metric across five independently trained models per method to improve robustness. Except for deep ensembles, instead, we report confidence bounds by 2000-fold bootstrapping as an alternative approach to circumvent the computational demands required to rerun deep ensembles multiple times.

4. Results

In this section, we present the results of the different OOD detection experiments. To provide the context of existing benchmarks on computer vision datasets, we start with the CIFAR10 experiment before analyzing performance on histopathology.

4.1. CIFAR10 vs. SVHN

Table 2 shows the target task and far-OOD detection performance for the CIFAR10 experiment. We observe that the target task performance of a single CNN is comparable with the performance reported by the original DenseNet paper of 94.76 (Huang et al., 2017). The other methods show similar performance except for deep ensembles, which slightly improve, with the best results for the ensemble with ten members. Regarding the OOD detection task, most uncertainty estimation methods perform better than the single CNN baseline, with the best performance by the deep ensembles. Here, we include results based on Maximum Softmax Probability (MSP) (Hendrycks and Gimpel, 2016) and MaxLogit (Hendrycks et al., 2022). Different from the regular single CNN baseline, the ROC analysis is performed based on the maximum softmax or logit output value respectively, instead of the entropy of the predictive distribution. Summarizing, the OOD detection AUC values: MSP (90.17) and MaxLogit (90.61) do not show competitive performance with the Deep Ensembles method (95.08) and only shows marginal improvements over the Single CNN baseline based on the entropy value (90.50). Based on these results, we conclude that the maximum softmax probability and the maximum logit values do not

Table 4
Prostate cancer detection performance (target task) on both the Prostate-520 and Prostate-colon datasets and far-OOD detection performance: colon tissue detection in prostate biopsies. OOD detection experiments are repeated using only the benign WSIs. Slide level uncertainty values are defined by either a Whole-Slide Average (WSA), or the 99th percentile (P_{99}) of the uncertainty heatmap. Mean and standard deviations of AUC values are reported for all methods across five independent runs except deep ensembles, which report confidence bounds by 2000-fold bootstrapping. M-heads models, trained with increasing levels of implicit head specialization, are included. Results from prior work on tumor detection performance on the full test set (lacking results on OOD detection performance) are included for comparison.

Model	Target task (AUC)		Prostate-520 vs. Prostate-colon (AUC)		Prostate-520 (benign) vs. Prostate-colon (benign), (AUC)	
	Prostate-520	Prostate-Colon	WSA	P_{99}	WSA	P_{99}
Single CNN	98.13±0.16	**89.16±0.48**	56.40±0.68	73.69±2.03	69.91±2.22	80.75±2.92
Temperature scaling	98.09±0.14	89.06±0.37	55.95±0.36	**74.42±0.73**	69.16±1.86	81.79±1.88
MC dropout	97.45±0.28	89.01±0.26	53.66±0.37	66.81±2.51	61.60±1.09	70.35±4.04
Deep ensembles 5	**98.19** [97.3, 98.4]	87.94 [80.3, 94.6]	54.35 [50.2, 58.6]	58.14 [54.1, 62.1]	71.48 [63.3, 79.5]	84.47 [78.4, 90.2]
Deep ensembles 10	98.12 [97.2, 98.9]	88.02 [80.1, 94.8]	54.38 [50.4, 58.3]	58.26 [54.5, 62.0]	71.40 [63.0, 79.1]	**84.55** [78.7, 89.9]
M-heads 5 ($\epsilon = 0.1$)	97.21±0.11	87.66±0.50	53.04±0.88	64.54±2.45	62.48±2.27	71.47±3.92
M-heads 10 ($\epsilon = 0.1$)	97.22±0.26	86.35±0.52	54.26±1.39	62.24±2.90	67.72±4.00	69.09±6.93
M-heads 5 ($\delta_m = 1/M$)	98.04±0.18	88.96±0.43	**56.53±0.21**	73.60±1.18	69.04±2.36	81.17±3.25
M-heads 5 ($\epsilon = 0.6$)	98.11±0.09	87.97±0.33	56.17±0.71	69.10±2.17	69.93±1.92	78.15±3.17
Bulten et al. (2020)	99.0 [98.2, 99.6]					

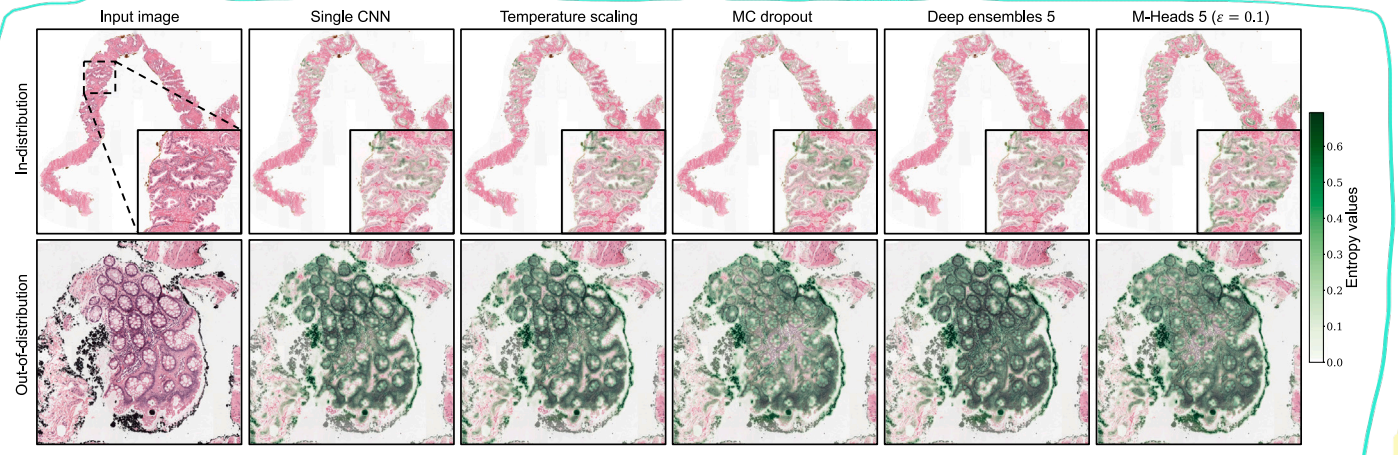

Fig. 5. Uncertainty heatmaps, based on the entropy value (1), for a benign prostate biopsy and Out-of-Distribution colorectal tissue.

contain more discriminative power than the entropy of the predictive distribution. Therefore, and for the consistency between all methods, we only include the results based on the entropy of the predictive distribution for the remainder of the experiments. We observe only slight improvements over the single CNN baseline for the M-heads model with five heads and no performance gains when using ten heads.

4.2. Near-OOD detection performance in lymph node tissue

The first two columns of Table 3 report the performance of the different methods in detecting breast cancer metastasis in lymph node tissue. Evaluations from a pathologist (Ehteshami Bejnordi et al., 2017) and state-of-the-art results (SOTA) from Liu et al. (2017) (at 20× resolution) are included for comparison. Also, the results from the Camelyon16 winner are included. However, it uses 40× resolution and is thus not directly comparable (Wang et al., 2016). These results show that most models are competitive with the current state-of-the-art, despite the relatively lightweight network architecture. To analyze the importance of ensemble diversity, we include results of M-heads with different levels of head specialization. We observe that the M-heads models perform slightly worse than the single CNN baseline, independent of the number of heads or the value of ϵ, defining the amount of specialization during training. The MC dropout model also demonstrates slightly worse performance than the single CNN baseline for both ROC and FROC metrics. Like the CIFAR10 experiment, the deep ensembles achieve the best target task performance with near SOTA results for both challenge metrics.

To analyze near-OOD detection performance, we start by evaluating uncertainty estimates on both the Camelyon datasets (C16 and C17) and the lymphoma dataset. Table 3 shows the results. See Fig. 4 for example uncertainty heatmaps for both in-distribution and OOD whole-slide images. Unlike the CIFAR10 experiment, deep ensembles only show minor performance gains compared to the single CNN baseline. Furthermore, increasing the number of members does not boost OOD detection performance. In contrast, the M-heads models with five and ten heads trained with the highest specialization setting outperform the baseline significantly. The model trained with a fixed Kronecker delta ($\delta_m = 1/M$), effectively neglecting ensemble diversity during training, shows slightly worse performance than the single CNN model. However, increasing ensemble diversity boosts OOD detection performance significantly. Furthermore, we observe more stable OOD detection performance by the most specialized M-heads model through the smaller standard deviations across the five independently trained models. Based on these results, we observe no trade-off between target task performance and increasing specialization to achieve higher near-OOD detection performance. When only considering the benign cases, we see further performance gains for all methods, with the M-heads model with ten heads achieving the highest average AUC score of 82.30. Furthermore, all models also perform better on the Camelyon17 test set.

4.3. Detecting colon tissue in prostate biopsies

The results of the experiments on prostate biopsies, based on the Prostate-520 and Prostate-Colon datasets 2, are reported in Table 4. Included are the results from the original paper containing all 535 WSIs (Bulten et al., 2020). Similar to the target task results in the previous experiment, we see competitive performance of most methods with the current state-of-the-art. Contrary to the results from the

Table 5
Far Out-of-Distribution (foreign tissue) detection performance: prostate tissue vs. lymph node tissue. With $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$: the in-distribution and OOD datasets used in the ROC analysis. Test data from Camelyon challenges (C16 and C17) and the prostate-520 dataset are used for the different OOD detection tasks. Performance using only benign in-distribution WSIs is also reported. Mean and standard deviations of AUC values are reported for all methods across five independent runs except deep ensembles, which report confidence bounds by 2000-fold bootstrapping. M-heads models, trained with increasing levels of implicit head specialization, are included.

Model	$\mathcal{D}_{out}$	Prostate tissue (Prostate-520)				Lymph node tissue (C16)	
	$\mathcal{D}_{in}$	C16	C17	C16 (benign)	C17 (benign)	Prostate-520	Prostate-520 (benign)
Single CNN		97.19±0.86	98.53±0.45	98.16±0.73	99.04±0.43	86.61±5.97	96.62±2.34
Temperature scaling		**97.94±0.82**	**98.82±0.45**	**98.73±0.72**	99.27±0.43	**89.34±3.07**	**97.72±0.59**
MC dropout		96.05±2.01	98.07±8.44	97.03±1.76	98.52±0.74	82.98±6.99	94.88±3.63
Deep ensembles 5		97.60 [96.5, 98.6]	98.74 [98.0, 99.3]	98.60 [97.8, 99.3]	**99.28 [98.8, 99.7]**	78.71 [75.4, 81.8]	97.68 [96.4, 98.9]
Deep ensembles 10		97.45 [96.3, 98.4]	98.68 [97.9, 99.3]	98.47 [97.7, 99.1]	99.26 [98.8, 99.7]	78.67 [75.5, 81.8]	97.67 [96.3, 98.8]
M-heads 5 ($\epsilon = 0.1$)		94.89±1.08	97.35±0.57	96.16±0.96	97.73±0.51	80.70±4.82	95.12±1.59
M-heads 10 ($\epsilon = 0.1$)		93.18±1.62	96.84±0.81	94.73±1.60	97.20±0.88	75.88±2.67	95.30±1.44
M-heads 5 ($\delta_m = 1/M$)		94.79±2.55	97.88±0.93	96.10±2.26	98.49±0.87	87.02±3.74	96.90±0.89
M-heads 5 ($\epsilon = 0.6$)		95.96±0.85	98.01±0.24	97.22±0.65	98.51±0.20	86.30±6.02	96.68±1.76

experiments on lymph node tissue, we observe that decreasing the ϵ parameter for M-heads slightly decreases its performance. When evaluating the target task on the set of WSIs containing colorectal tissue, Prostate-Colon, we see a significant drop in performance for all methods, highlighting the negative impact OOD tissue can have on the performance of DNNs in clinical practice.

To analyze far-OOD detection performance in prostate tissue, we start with a ROC analysis based on the average uncertainty value across the tissue area (WSA, see 3.6). Table 4 shows the results. We observe that the WSA does not contain enough signal to reliably indicate the presence of OOD tissue on a WSI level. However, when only considering values above the 99th percentile P_{99} of the uncertainty heatmap, the AUC increases significantly for each method. In contrast to the previous experiment, the OOD tissue only accounts for a relatively small region within the WSIs such that the subset with the highest uncertainty values becomes a more informative signal. Fig. 5 visualizes uncertainty heatmaps for a benign WSI and an OOD region from Prostate-Colon. Most uncertainty estimation methods perform worse than the single CNN baseline, except for the temperature scaled model and the M-heads model trained without implicit specialization ($\delta_m = 1/M$).

We repeat the OOD detection experiment by only considering the benign cases from both Prostate-520 and Prostate-Colon datasets. The resulting AUC values, reported in the final two columns in Table 4, demonstrate significant performance gains for every method compared to the previous setting containing tumor cases. Overall, we observe that the M-heads model does not benefit from using the meta-loss function during training for detecting foreign tissue in prostate biopsies. Instead, we see that the M-heads model performs better without any implicit specialization ($\delta_m = 1/M$). Furthermore, unlike the results on the other datasets, we observe that the MC dropout model performs significantly worse than the single CNN baseline model. However, deep ensembles demonstrate informative uncertainty estimates with AUC values up to 84.55 based on the 99th percentile for the ensemble based on ten members.

4.4. Foreign tissue detection

Finally, we evaluate a more extreme setting of far-OOD detection by evaluating uncertainty estimates on foreign tissue. Table 5 shows the results. The relatively high AUC scores for all methods across the different datasets indicate a more tractable OOD detection task compared to the previous experiments. Here we observe that overall, the temperature scaled model and the deep ensembles perform best across most datasets. However, similar to the previous experiment, ensembles seem most affected by the presence of tumor tissue for the OOD detection task. Although M-heads demonstrated the best OOD detection performance for detecting lymphoma tissue, its performance in detecting foreign tissue seems worse compared to the other methods. Furthermore, in contrast to the detection of near-OOD data, increasing the specialization setting during training does not lead to improved far-OOD detection. For M-heads specifically, these results are in line with the far-OOD experiment from the previous section, detecting colon tissue in prostate biopsies.

5. Discussion

In this article, we have focused on OOD detection by evaluating the uncertainty estimates for different methods. In our experiments, the baseline defined by a single model achieved reasonable performance for many of the OOD detection tasks, which is in line with previous work (Hendrycks and Gimpel, 2016). However, the different uncertainty based methods outperform the single model baseline in most tasks. Although deep ensembles demonstrate the most consistent performance gains over a single CNN model in most target tasks and OOD detection tasks, the M-heads model trained with high specialization settings performs better on the near-OOD task.

These results demonstrate that the benchmarks for uncertainty estimation on classical computer vision datasets do not fully translate to a more clinically relevant setting within digital pathology. Similar to results on computer vision benchmarks, such as the CIFAR10 experiment, deep ensembles demonstrate the most informative uncertainty estimates for OOD samples that lay far from the training distribution. This is demonstrated by high OOD detection scores for tasks involving foreign tissue, such as detecting colorectal tissue in prostate biopsies or discriminating prostate tissue from lymph node tissue, except when the model is trained on prostate tissue. Furthermore, the temperature scaled model demonstrates competitive performance with deep ensembles in the far-OOD detection experiments. These results show that it could be a cheap alternative for efficient far-OOD detection. However, for the more challenging near-OOD detection task of detecting lymphoma in lymph node resections, the ensembles and the temperature scaled model perform on par with a single CNN baseline. Instead, M-heads with specialized heads seem to outperform all other methods, including the M-heads models trained without the meta-loss or with lower specialization settings. This indicates that training for a diverse ensemble might benefit near-OOD detection performance. The lymphoma tissue is highly similar to the benign class of the training data, originating from the same tissue type, making it more difficult to discriminate it from the in-distribution data. The meta-loss function seems imperative to boost OOD detection performance in this case. These results are in line with recent work on near-OOD detection in dermatology, demonstrating the importance of ensemble diversity for improved OOD detection (Guha Roy et al., 2022).

Throughout the different experiments, we have seen that the OOD detection performance increases when only considering benign cases, indicating higher uncertainty values for tumor slides. This can be explained by the difficulty of accurately identifying the tumor border, but the effect is also inherent to the patch-based classification approach adopted in this work. When classifying the central pixel of a patch at the tumor border, the patch contains both benign and tumor tissue,

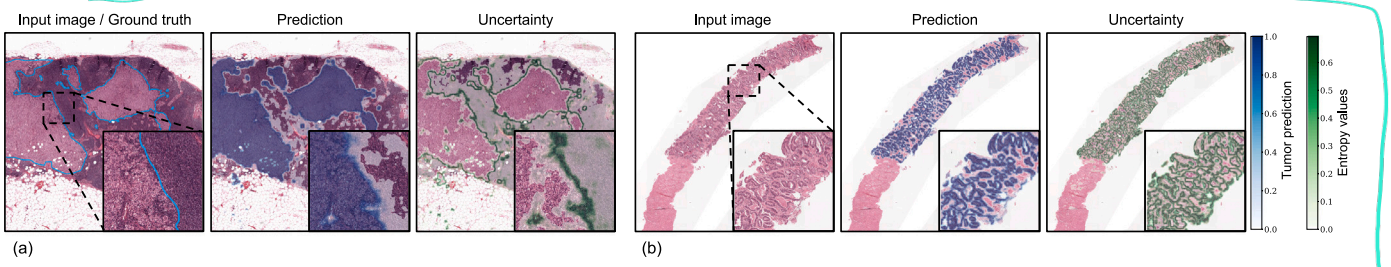

Fig. 6. Two example images demonstrating high uncertainty at the tumor boundary, from the (a) Camelyon16 test set and the (b) Prostate-520 dataset. Pixel-level ground truth annotations are available for C16 and included in the left panel. Prediction and corresponding uncertainty heatmaps are shown for the deep ensemble model as an example, although similar heatmaps are obtained with the other methods. The relative high uncertainty values at tumor boundaries influence the slide-level uncertainty score, explaining the improved performance when only considering benign cases in the OOD detection experiments.

making the classification task more difficult and resulting in higher uncertainty. See Fig. 6 for example prediction and corresponding uncertainty heatmaps. We have seen that deep ensembles are especially affected by this throughout the experiments on prostate biopsies. The relatively large surface area of tumor boundaries from small tumor regions across the prostate biopsies partially explains these results. Furthermore, differences between tumor and benign tissue for the prostate experiment are smaller than in the lymph node experiments, where tumor tissue originates from a completely different tissue type (See Fig. 1). Future work could alleviate these issues by separating different types of uncertainty (Kendall and Gal, 2017). Occasionally, rare morphological cases such as large areas of muscle tissue were flagged by the uncertainty estimates. However, other types of imaging artifacts such as out-of-focus areas or air bubbles, were mostly absent from the carefully designed challenge test sets used in our work, leaving room for future research on the translation to clinical datasets.

The relatively high OOD detection performance on benign cases combined with the tumor detection performance throughout this work suggests a clinical workflow where cases can be selected for further analysis: first based on a high tumor detection score or afterward based on a high uncertainty estimate. With recent developments in unsupervised learning, this second step could also be handled by a completely different module or model to directly detect OOD data. However, this being completely outside the scope of our work, we leave this for future research.

The models trained to detect breast cancer metastasis in lymph node tissue have shown lower uncertainty values on WSIs from the independent dataset from Camelyon17 compared to the WSIs from the actual Camelyon16 test set. This could be explained by the fact that the data from the Camelyon17 set was collected at a later time point with possibly improved data collection pipelines, leading to higher quality WSIs. Regarding the experiment on detecting foreign tissue presented in Table 5, we observe differences between methods trained on the Camelyon datasets and those trained on prostate biopsies. Being symmetrical in setup, we consider both OOD detection experiments far-OOD. We hypothesize that differences here are attributed to the lower amounts of morphology variety present in the lymph node tissue training dataset, leading to a more tight decision boundary and better OOD detection performance on foreign tissue.

Specifically for the MC dropout model, we have seen worse performance than the single CNN baseline on many target and OOD detection tasks throughout this work. These results indicate that applying dropout throughout the network could harm performance. Optimizing the specific location of the dropout layers and the corresponding probability on the validation set might result in improved performance during inference. However, we leave this analysis for future work. Furthermore, adding more members to the different ensembles throughout this work did not always boost performance, indicating diminishing returns for larger ensembles. However, determining the optimal size of the ensembles, similar to other hyperparameters like the specialization settings of M-heads, might be problem-dependent, requiring independent tuning for each task. This could also explain why the specialization settings of M-heads, which were defined based on preliminary analysis on mode-collapse using the validation set of the Camelyon experiment, did not lead to improved OOD detection performance in the prostate biopsy experiment.

Overall, the different uncertainty based methods did not significantly outperform the single CNN baseline in terms of tumor detection performance. As such, we attribute the differences in OOD detection performance mostly to the quality of the uncertainty estimates, not simply to improvements in tumor detection performance. Finally, we note that the ROC analysis used throughout this work evaluates the ability to discriminate in-distribution and OOD WSIs independent from the exact decision threshold on the uncertainty estimate. In clinical practice however, this would require task dependent tuning to determine the optimal decision threshold.

6. Conclusion

We have evaluated and compared multiple prevalent uncertainty estimation methods in digital pathology through various experiments. To rigorously assess the quality of the uncertainty estimates, we performed AUC analyses to discriminate between data similar to the training data and out-of-distribution samples on multiple large-scale datasets with anomalies that can be expected during model deployment. The results show that the aggregate entropy values: either the whole-slide average or the values above a percentile of the uncertainty heatmap, can be used as an informative uncertainty signal on a slide level, including in cases where the OOD region only covers a small part of the WSI. We have seen that tumor detection performance can be negatively impacted by the presence of OOD data, as shown in the prostate biopsy experiment. This, combined with the possibility of falsely reporting lymphoma cases as benign tissue in the lymph node experiment due to morphological similarities, emphasizes the necessity for informative uncertainty estimates in digital pathology as a precondition for model deployment.

Throughout the experiments, we demonstrate performance differences between the near-OOD detection task and the far-OOD detection tasks. We observe that ensemble diversity can increase OOD detection performance on this specific near-OOD detection task but not on far-OOD data. However, more research on other near-OOD detection tasks is required to confirm these findings. Similarly, the translation of the uncertainty estimates on OOD regions of varying size to a slide-level uncertainty score further complicates the design of uncertainty estimation pipelines. As such, problem-dependent tuning of OOD detection pipelines, based on expert knowledge, might be required in clinical practice. However, we have seen that uncertainty estimates can be used to discriminate in-distribution from OOD samples with high AUC scores. The slide-level uncertainty scores could enable practitioners to assess the reliability of the predictions and identify anomalous data on a whole-slide image level. As such, uncertainty estimates can potentially reduce adverse effects of OOD regions on the target task further down the pipeline in a clinical workflow, and could limit models from failing silently on unseen diseases or abnormalities.

CRediT authorship contribution statement

Jasper Linmans: Conceptualization, Methodology, Software, Validation, Investigation, Formal analysis, Writing – original draft, Visualization. **Stefan Elfwing:** Writing – review & editing, Supervision. **Jeroen van der Laak:** Data curation, Funding acquisition, Writing – review & editing, Project administration, Supervision. **Geert Litjens:** Conceptualization, Data curation, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jeroen van der Laak reports financial support was provided by ContextVision AB. Jeroen van der Laak reports a relationship with Philips that includes: board membership and funding grants. Jeroen van der Laak reports a relationship with ContextVision AB that includes: board membership and funding grants. Jeroen van der Laak reports a relationship with Sectra AB that includes: funding grants. Jeroen van der Laak reports a relationship with Aiosyn BV that includes: employment and equity or stocks. Stefan Elfwing reports a relationship with ContextVision AB that includes: employment. Stefan Elfwing reports a relationship with Inify Laboratories AB that includes: employment. Geert Litjens reports a relationship with Philips that includes: funding grants. Geert Litjens reports a relationship with Canon Health Informatics (USA) that includes: consulting or advisory. Geert Litjens reports a relationship with Aiosyn BV that includes: equity or stocks.

Data availability

Data will be made available on request.

Acknowledgments

This study was funded by ContextVision AB, Linköping, Sweden. Jeroen van der Laak and Geert Litjens have also received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 945358. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation program and EFPIA. Geert Litjens also received funding by a grant from the Dutch Cancer Society (KWF), grant number KUN 2015-7970 and the Dutch Research Council (NWO) under Veni grant number 91618152. The Knut and Alice Wallenberg foundation is acknowledged for generous support.

References

Bándi, P., Balkenhol, M., van Ginneken, B., van der Laak, J., Litjens, G., 2019. Resolution-agnostic tissue segmentation in whole-slide histopathology images with convolutional neural networks. PeerJ 7, e8242. http://dx.doi.org/10.7717/peerj.8242.

Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B.E., Lee, B., Paeng, K., Zhong, A., et al., 2018. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. IEEE Trans. Med. Imaging 38 (2), 550–560. http://dx.doi.org/10.1109/TMI.2018.2867350.

Bulten, W., Kartasalo, K., Chen, P.-H.C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D.F., van Boven, H., Vink, R., et al., 2022. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. Nature Med. 1–10. http://dx.doi.org/10.1038/s41591-021-01620-2.

Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van der Laak, J., Hulsbergen-van de Kaa, C., Litjens, G., 2020. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncol. 21, 233–241. http://dx.doi.org/10.1016/S1470-2045(19)30739-9.

Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J., the CAMELYON16 Consortium, 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 318 (22), 2199–2210. http://dx.doi.org/10.1001/jama.2017.14585.

Fort, S., Hu, H., Lakshminarayanan, B., 2019. Deep ensembles: A loss landscape perspective. http://dx.doi.org/10.48550/ARXIV.1912.02757, arXiv preprint arXiv:1912.02757.

Fox, J.P., Grignol, V.P., Gustafson, J., Cheng, P., Weighall, R., Ouellette, J., Hellan, M., Dowdy, Y., Termuhlen, P., 2010. Incidental lymphoma during sentinel lymph node biopsy for breast cancer. J. Clin. Oncol. 28 (15), e11083. http://dx.doi.org/10.1200/jco.2010.28.15_suppl.e11083.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning. pp. 1050–1059, URL: https://proceedings.mlr.press/v48/gal16.html.

Graham, M.S., Tudosiu, P.-D., Wright, P., Pinaya, W.H.L., U-King-Im, J.-M., Mah, Y., Teo, J., Jäger, R.H., Werring, D., Nachev, P., Ourselin, S., Cardoso, M.J., 2022. Transformer-based out-of-distribution detection for clinically safe segmentation. In: Medical Imaging with Deep Learning. URL: https://openreview.net/forum?id=En7660i-CLJ.

Guha Roy, A., Ren, J., Azizi, S., Loh, A., Natarajan, V., Mustafa, B., Pawlowski, N., Freyberg, J., Liu, Y., Beaver, Z., Vo, N., Bui, P., Winter, S., MacWilliams, P., Corrado, G.S., Telang, U., Liu, Y., Cemgil, T., Karthikesalingam, A., Lakshminarayanan, B., Winkens, J., 2022. Does your dermatology classifier know what it doesn't know? Detecting the long-tail of unseen conditions. Med. Image Anal. 75, 102274. http://dx.doi.org/10.1016/j.media.2021.102274.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning. pp. 1321–1330, URL: https://proceedings.mlr.press/v70/guo17a.html.

Gustafsson, F.K., Danelljan, M., Schön, T.B., 2020. Evaluating scalable bayesian deep learning methods for robust computer vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 318–319. http://dx.doi.org/10.48550/ARXIV.1906.01620.

Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., Song, D., 2022. Scaling out-of-distribution detection for real-world settings.

Hendrycks, D., Gimpel, K., 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136. URL: https://arxiv.org/abs/1610.02136.

Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2261–2269. http://dx.doi.org/10.1109/CVPR.2017.243.

Ianni, J.D., et al., 2020. Tailored for real-world: a whole slide image classification system validated on uncurated multi-site data emulating the prospective pathology workload. Sci. Rep. 10 (1), 1–12. http://dx.doi.org/10.1038/s41598-020-59985-2.

Karimi, D., Zeng, Q., Mathur, P., Avinash, A., Mahdavi, S., Spadinger, I., Abolmaesumi, P., Salcudean, S.E., 2019. Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images. Med. Image Anal. 57, 186–196. http://dx.doi.org/10.1016/j.media.2019.07.005.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in Bayesian deep learning for computer vision? In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 5580–5590. http://dx.doi.org/10.48550/ARXIV.1703.04977.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations, San Diego, USA. http://dx.doi.org/10.48550/ARXIV.1412.6980.

Kompa, B., Snoek, J., Beam, A.L., 2021. Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digit. Med. 4, 1–6. http://dx.doi.org/10.1038/s41746-020-00367-3.

Krizhevsky, A., Nair, V., Hinton, G., 2009. Learning Multiple Layers of Features from Tiny Images. Technical Report, CIFAR, URL: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 6402–6413, URL: https://dl.acm.org/doi/10.5555/3295222.3295387.

Lee, S., Purushwalkam, S., Cogswell, M., Crandall, D., Batra, D., 2015. Why M heads are better than one: Training a diverse ensemble of deep networks. http://dx.doi.org/10.48550/ARXIV.1511.06314, arXiv preprint arXiv:1511.06314.

Leibig, C., Allken, V., Ayhan, M.S., Berens, P., Wahl, S., 2017. Leveraging uncertainty information from deep neural networks for disease detection. Sci. Rep. 7 (1), 1–14. http://dx.doi.org/10.1038/s41598-017-17876-z.

Linmans, J., van der Laak, J., Litjens, G., 2020. 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. pp. 465–478, URL: https://openreview.net/forum?id=IdZWfAGuuB.

Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., van de Loo, R., Vogels, R., et al., 2018. H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. GigaScience 7 (6), http://dx.doi.org/10.1093/gigascience/giy065.

Liu, Y., Gadepalli, K.K., Norouzi, M., Dahl, G., Kohlberger, T., Venugopalan, S., Boyko, A.S., Timofeev, A., Nelson, P.Q., Corrado, G., Hipp, J., Peng, L., Stumpe, M., 2017. Detecting cancer metastases on gigapixel pathology images. http://dx.doi.org/10.48550/ARXIV.1703.02442, arXiv preprint arXiv:1511.06314, Also Presented at the 2017 MICCAI Tutorial, Deep Learning for Medical Imaging.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). http://dx.doi.org/10.1109/CVPR.2015.7298965.

Nair, T., Precup, D., Arnold, D.L., Arbel, T., 2020. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Med. Image Anal. 59, 101557. http://dx.doi.org/10.1016/j.media.2019.101557.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y., 2011. Reading digits in natural images with unsupervised feature learning.

Osband, I., Blundell, C., Pritzel, A., Van Roy, B., 2016. Deep exploration via bootstrapped DQN. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, Vol. 29. pp. 4026–4034, URL: https://dl.acm.org/doi/10.5555/3157382.3157548.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J.V., Lakshminarayanan, B., Snoek, J., 2019. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. pp. 14003–14014, URL: https://dl.acm.org/doi/abs/10.5555/3454287.3455541.

Rupprecht, C., Laina, I., DiPietro, R., Baust, M., Tombari, F., Navab, N., Hager, G.D., 2017. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In: Proceedings - 2017 IEEE International Conference on Computer Vision, ICCV 2017. pp. 3591–3600. http://dx.doi.org/10.1109/ICCV.2017.388.

Schömig-Markiefka, B., Pryalukhin, A., Hulla, W., Bychkov, A., Fukuoka, J., Madabhushi, A., Achter, V., Nieroda, L., Büttner, R., Quaas, A., Tolkach, Y., 2021. Quality control stress test for deep learning-based diagnostic model in digital pathology. Mod. Pathol. 34, 1–11. http://dx.doi.org/10.1038/s41379-021-00859-x.

Thagaard, J., Hauberg, S., van der Vegt, B., Ebstrup, T., Hansen, J., Dahl, A., 2020. Can you trust predictive uncertainty under real dataset shifts in digital pathology? In: Proceedings of Medical Image Computing and Computer Assisted Intervention Conference. pp. 824–833. http://dx.doi.org/10.1007/978-3-030-59710-8_80.

Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C., 2015. Efficient object localization using convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 648–656. http://dx.doi.org/10.1109/CVPR.2015.7298664.

van den Bergh, R.C., Wolters, T., Spaander, M.C., Schröder, F.H., van Leenders, G.J., 2009. Non-prostatic pathology on prostate needle-biopsy–colorectal carcinoid: a case report. Cases J. 2 (1), 1–4. http://dx.doi.org/10.1186/1757-1626-2-75.

Wang, D., Khosla, A., Gargeya, R., Irshad, H., Beck, A.H., 2016. Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718. URL: https://arxiv.org/abs/1606.05718.

Winkens, J., Bunel, R., Roy, A.G., Stanforth, R., Natarajan, V., Ledsam, J.R., MacWilliams, P., Kohli, P., Karthikesalingam, A., Kohl, S., Cemgil, T., Eslami, S.M.A., Ronneberger, O., 2020. Contrastive training for improved out-of-distribution detection. http://dx.doi.org/10.48550/ARXIV.2007.05566, arXiv preprint arXiv:2007.05566.