Influence of Classification Task and Distribution Shift Type on OOD Detection in Fetal Ultrasound

Chun Kit Wong[1,2][0000−0001−5528−9727], Anders N. Christensen[1][0000−0002−3668−3128], Cosmin I. Bercea[3,4][0000−0003−2628−2766], Julia A. Schnabel[3,4,5][0000−0001−6107−3009], Martin G. Tolsgaard[6,7][0000−0001−9197−5564], and Aasa Feragen[1,2][0000−0002−9945−981X]

[1] Technical University of Denmark, Kongens Lyngby, Denmark
{ckwo,afhar}@dtu.dk
[2] Pioneer Centre for AI, Copenhagen, Denmark
[3] Technical University of Munich, Munich, Germany
[4] Helmholtz AI and Helmholtz Munich, Munich, Germany
[5] King's College London, London, UK
[6] University of Copenhagen, Copenhagen, Denmark
[7] CAMES Rigshospitalet, Copenhagen, Denmark

Abstract. Reliable out-of-distribution (OOD) detection is important for safe deployment of deep learning models in fetal ultrasound amidst heterogeneous image characteristics and clinical settings. OOD detection relies on estimating a classification model's uncertainty, which should increase for OOD samples. While existing research has largely focused on uncertainty quantification methods, this work investigates the impact of the classification task itself. Through experiments with eight uncertainty quantification methods across four classification tasks on the same image dataset, we demonstrate that OOD detection performance significantly varies with the task, and that the best task depends on the defined ID-OOD criteria; specifically, whether the OOD sample is due to: i) an image characteristic shift or ii) an anatomical feature shift. Furthermore, we reveal that superior OOD detection does not guarantee optimal abstained prediction, underscoring the necessity to align task selection and uncertainty strategies with the specific downstream application in medical image analysis. Code: https://github.com/wong-ck/ood-fetal-us.

Keywords: OOD · uncertainty quantification · fetal ultrasound.

1 Introduction

Out-of-distribution (OOD) detection is crucial for deploying reliable deep learning models in medical image analysis. This is particularly needed in fetal ultrasound, which is ubiquitous in routine maternity check-ups, but also comes with significant heterogeneity in image characteristics due to differences in operator training or maternal Body Mass Index (BMI), and a diverse range of ultrasound scanners. This variability in input image distribution directly impacts the performance of deep learning models, underscoring the need for robust OOD detection to identify distributional shifts of different kinds and ensure diagnostic accuracy.

OOD detection finds wide application in medical imaging [6,26]. For image quality and domain shift detection, [9] attempted to detect distribution shift for surveillance of deployed AI algorithms, and [15] developed calibration technique to estimate performance of a trained model under domain shift. Another task is anatomical shift detection, e.g. identifying unseen pathologies or abnormalities during inference. In chest X-ray analysis, [1] identified fracture cases as OOD examples using a classifier trained to distinguish cardiomegaly from pneumothorax. Similarly, in dermatology, [3] utilized images of unseen skin diseases as OOD sets. In digital pathology, researchers have explored the detection of novel abnormalities as OOD samples [16,23]. Beyond pathology detection, OOD methods are also used to extract clinically relevant frames from ultrasound videos, identifying frames that deviate from expected anatomical content [17].

OOD detection can be formulated via uncertainty quantification (UQ). There is a diverse landscape of UQ techniques applicable to medical imaging [7,13,27], many focusing on image classification [10,22]. Here, OOD detection relies on estimating the predictive uncertainty of a classifier for a given input image, reflecting the model's confidence in its prediction. Higher uncertainty typically indicates a greater likelihood that the input image originates from a distribution different from the training data, suggesting it is out-of-distribution.

While many easily-available UQ algorithms are based on how a particular classifier views the data, being OOD is a natural property of the data itself. This is important from the clinic's point-of-view, where in fetal ultrasound we observe both classical quality shifts such as blur and low resolution, or other shifts in image characteristics such as hue or artifacts – but also anatomical shifts in off-plane images. In the clinic, it is therefore important that OOD detection is robust and consistent for different types of distribution shifts.

Contributions. We study eight common classifier-based UQ techniques for OOD detection. To study robustness of OOD detection across types of distribution shifts, we challenge the idea that the UQ techniques should be based on the **primary classifier of interest**, which for this paper will be an *anatomical plane classifier*. As UQ base classifiers, we train three alternative models to predict image meta-information found in the DICOM header, and study how the resulting four models perform as a basis for the different UQ-methods, validated on OOD distribution for i) image characteristic distribution shifts, and ii) anatomical shifts. Finally, we test how our primary classifier of interest performs in an "abstained prediction" setting using the different OOD models.

We find that the choice of base classifier for the OOD detector has a large effect on OOD performance, and that the primary classifier is not always the best choice. However, we also see that these results are surprisingly not indicative of performance in "abstained prediction" for the primary classifier – where the UQ-methods based on the primary classifier perform better than the competitors. This shows that the choice of base classifiers matters for OOD detection, and suggests that OOD detection and trustworthy classification are not always two sides of the same coin. In particular, our results underscore that classical OOD

detection performance is not always suitable for model selection if the downstream application – as is often relevant in the clinic – is abstained prediction.

2 Method

2.1 Uncertainty Quantification (UQ) Methods

Under a C classes classification setting, a classifier is trained to predict the class label $\mathbf{y} \in \{1 \dots C\}$ given an input image $\mathbf{x} \in \mathbb{R}^2$, with a predictive uncertainty score $u(\mathbf{x})$ that can be estimated using UQ-methods. OOD detection is then achieved by thresholding: If $u(\mathbf{x}) > t$, then $\mathbf{x}$ is flagged as OOD. Inspired by [18], we evaluate eight UQ-methods that do not rely on a hold-out OOD test set during training. We focus on the following, mainly deterministic, methods considering their fast, non-iterative inference procedure:

As a **Baseline**, we trained a ResNet-50 model for the classification task and calculated the entropy of the model's predictive softmax probability as $u(\mathbf{x})$. **Temperature Scaling** [5] adds a calibration step to the softmax probability.

Two alternative methods augment the baseline model with an auxiliary prediction head. **Loss Prediction** [8, 11, 25] trains this head to predict the loss $L(\mathbf{x})$, while **Correctness Prediction** [8, 18] trains it to predict the probability $p(\hat{\mathbf{y}} \neq \mathbf{y}|\mathbf{x})$ of an incorrect prediction, using the predicted values as $u(\mathbf{x})$.

Meanwhile, two methods make use of the feature embedding space density. With a trained classification model, **Deterministic Uncertainty Quantification (DUQ)** [24] and **Deep Deterministic Uncertainty (DDU)** [19] work by first obtaining feature embeddings of all training images, followed by learning a density estimator using these embeddings. To ensure the latent space is well-regularized, DUQ adds a gradient penalty term in the loss function, while DDU applies spectral normalization to the model weights. $u(\mathbf{x})$ is then given by one minus the estimated density of $\mathbf{x}$ in the feature embedding space.

Finally, we also evaluated two probabilistic methods given their popularity in medical image analysis literature [13]. In these methods, multiple predictions are obtained for a given $\mathbf{x}$. For each prediction, **MC-dropout** [4, 21] switches off a random subset of the model activations, while **Ensemble** [12] uses a trained model that is initialized differently. Here, we used a lightweight implementation of ensemble [14], which involves training a single model with multiple randomly-initialized heads. Entropy of these predictions are taken as $u(\mathbf{x})$.

2.2 Dataset

We utilized a combination of two public and one private fetal ultrasound datasets.

The SONAI dataset is a private fetal ultrasound dataset including images from four common fetal anatomical planes: abdomen, brain, femur, and thorax, acquired using advanced ultrasound scanners. Additionally, each image is accompanied by metadata in the form of a DICOM header, which we utilized in designing our classification tasks (see Sec. 2.3). This dataset includes images of

other 'generic' fetal ultrasound planes, which we group into a separate bundle named **the SONAI (Other) dataset**. This will be used as an OOD dataset.

The BCNatal dataset [2] also comprises images from the same four fetal anatomical plane, acquired using advanced ultrasound scanners similar to those used for the SONAI dataset. This dataset also includes images of other 'generic' ultrasound images, where we group into a separate bundle and refer to as **the BCNatal (Other) dataset**, to be used as an OOD dataset.

The African dataset [20] comprises fetal ultrasound images of the same four anatomical planes as the two other datasets, but acquired in resource-constrained settings with less advanced ultrasound scanners. This results in images of lower quality compared to the two other datasets.

2.3 Classification tasks

Four distinct classification tasks were designed using the SONAI dataset, based on the metadata accompanying each image or the anatomy shown.

Plane Classification assigns ultrasound images into four anatomical planes: abdomen (n=1947), brain (n=3059), femur (n=1832), and thorax (n=2125). This task is a core application in fetal ultrasound and serves as a primary in-distribution task for our experiments. **Scanner Classification** identifies the machine used to acquire each image, which can be GE Voluson S (n=3723), V830 (n=2282) or E10 (n=2958). **DICOM Type Classification** predicts image type, which can be a single-region b-mode (n=6786), multi-region b-mode (n=1964), or a color Doppler (n=213) ultrasound image. Finally, **Maternal BMI Group Classification** predicts the BMI group of the pregnant subject undergoing the ultrasound scan, i.e. underweight (BMI $\leq$ 18, n=2452), normal (BMI 19-24, n=2260), overweight (BMI 25-29, n=2230), or obese (BMI $\geq$ 30, n=2021).

3 Experiments and Results

Following [18], we trained 160 classifier models for the four classification tasks (see Sec. 2.3), using the eight UQ-methods (see Sec. 2.1) repeated with five random seeds, i.e. we test both i) the influence of the classification task, and ii) the UQ method independently. All models were based on a ResNet-50 backbone, and trained for 100 epochs. Experiments were conducted on an AlmaLinux 8.7 server with NVIDIA RTX A6000 GPU. All datasets follow a 80:10:10 train-val-test split. Statistical significance was tested using a multi-factor ANOVA model.

3.1 Classification accuracy drops as image characteristics change

Before assessing uncertainty quantification, we validate the accuracy of our UQ-endowed classifiers when trained for our primary classification task: anatomical plane classification. As the models are trained on the SONAI dataset, we expect some distribution shift, and drop in accuracy from the SONAI to BCNatal test sets, as the latter comes from a different site. We expect a further drop for the

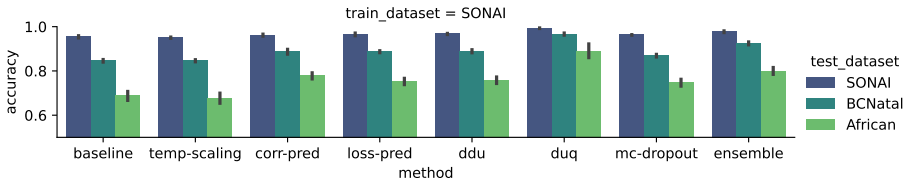

Fig. 1: Accuracy of the anatomical plane classifiers, trained using the eight UQ-methods, in classifying anatomical planes across three test datasets.

African test set, where we also expect a drop in image quality due to a less advanced scanner. Indeed, Fig. 1 shows a significant trend of accuracy dropping from the SONAI, to BCNatal, and to the African test sets for all models.

This demonstrates the vulnerability of deep learning models to shifts in input image distribution, highlighting the importance of reliable OOD detection methods. This also motivates our subsequent experiments to evaluate the effectiveness of different OOD detection methods in identifying these input distribution shifts.

3.2 The effect of base classification task on OOD detection

We now investigate our main research question: *How does the base classification task affect the UQ-methods' performance in OOD detection?* We hypothesize that the task would influence the learned feature representations of the model and, consequently, their ability to identify OOD samples.

Moreover, we study our follow-up research question: *Do "good" classification tasks perform equally well at detecting different plausible distribution shifts?* To answer this, we study how the classification task affects our ability to detect two different families of distribution shifts that occur naturally in fetal ultrasound screening: Shift in image characteristics due to different (quality) scanners, and anatomical shift, which occurs when operators search for a given anatomical plane or feed an incorrect image to the model.

All models below were trained on the SONAI training set for each of the four classification tasks defined in Sec. 2.3.

OOD 1: Shift in image characteristics. Fig. 2 shows how ID vs OOD classification performance, quantified via AUROC, distributes across the eight UQ-methods when built on the four different base classifiers.

First, as expected, we see a significantly higher OOD classification performance for the African dataset than for BCNatal, confirming that most methods can pick up on the expected increased distribution shift for the African dataset.

Second, we observe that for both distribution shifts, and for almost every UQ-method evaluated, models trained on the scanner classification task consistently and significantly outperformed models trained on the other three tasks. This suggests that, in the presence of a strong image quality shift, the scanner classification task led to feature representations that were remarkably more

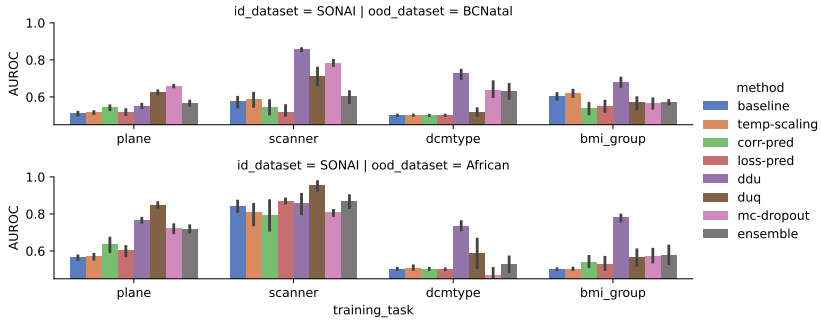

Fig. 2: **Image characteristics shift:** OOD detection performance across UQ-methods, for the four classification tasks. The OOD datasets (BCNatal, African) constitute a shift in image characteristics (image quality, scanner characteristics).

effective at distinguishing OOD data. Notably, models trained on the plane classification task, which represents our primary classification task of interest in the clinic, were almost consistently *not* the best performing OOD detectors. This is particularly interesting, as the primary classifier would normally be the choice of base classifier for OOD detection.

OOD 2: Anatomical shift. Fig. 3 shows how ID vs OOD classification performance, quantified via AUROC, distributes across the eight UQ-methods built on the four base classifiers. We consider two different datasets with anatomical shift from the SONAI dataset; namely SONAI (Other) and BCNatal (Other).

This experiment yields a different picture from our first set of experiments: For detecting SONAI (Others), the UQ-methods built on the primary plane classification task perform best by far. This base classifier is also competitive on BCNatal (Other), although the scanner classifier is, again, slightly better.

3.3 Abstained Prediction

To further investigate the practical implications of task-dependent uncertainty quantification, we explored the performance of our primary (anatomical plane) classifier, using the different UQ-models in an abstained prediction scenario. Abstained prediction is a more stringent evaluation of uncertainty estimation than OOD detection alone. Here, the model abstains from making a prediction when $u(\mathbf{x}) > \tau$ for an input image $\mathbf{x}$ exceeds a predefined τ, and accuracy is evaluated based on the remaining, non-abstained samples.

We first utilized the models trained for the plane classification task with each of the eight UQ-methods. For each plane classification model, we varied the uncertainty threshold τ, and calculated the accuracy of the model only on the samples for which it did not abstain (i.e., $u(\mathbf{x}) \leq \tau$). We repeated this process for a range of thresholds τ to generate an accuracy coverage curve. Next,

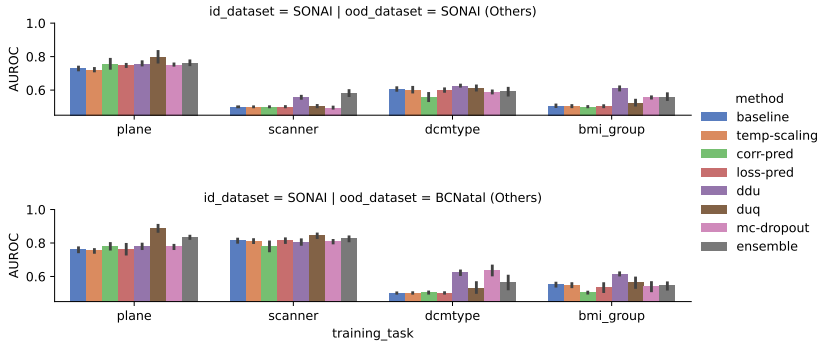

Fig. 3: **Anatomical feature shift:** OOD detection performance across UQ-methods for each of the four classification tasks using the OOD datasets (SONAI (Others), BCNatal (Others)).

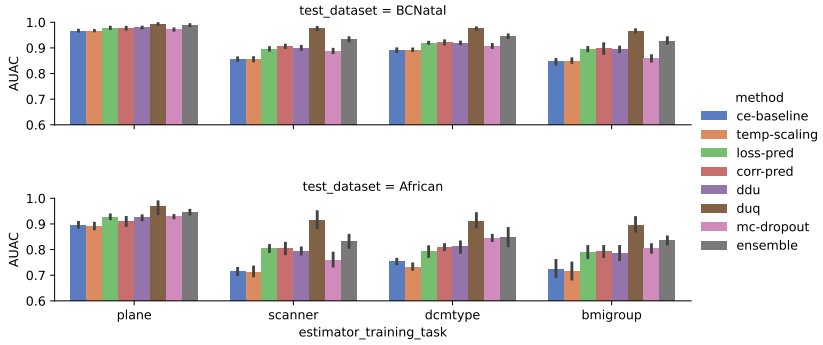

Fig. 4: Performance of "plane classification" with abstained prediction, across base classifier for UQ-methods as well as image characteristic distribution shifts.

we repeated the experiment with $u(\mathbf{x})$ generated by the base models trained for scanner type, dicom type, and bmi group classification instead. Following the convention in [18], we evaluated the abstained prediction performance by calculating the area under the accuracy coverage curve (AUAC).

We had expected that models exhibiting strong OOD detection performance would also excel in abstained prediction. However, Fig. 4 shows that models trained for scanner classification, which demonstrated superior OOD detection capabilities (especially for shifts in image characteristics), performed worse in abstained prediction for the plane classification task than UQ-models using plane classification as a base model: On both image characteristic shifts, uncertainty estimated with plane classification models generally leads to higher AUAC.

4 Discussion and conclusion

Our first result is that **the performance in OOD of a UQ-model depends heavily both the UQ-method's base classification task, and the OOD distribution shift.** Sec. 3.2 suggests that the base classification task plays a critical role in shaping the feature space utilized by the UQ methods. Different tasks lead to distinct learned feature spaces, ultimately impacting OOD detection. For image characteristic shifts (BCNatal and African datasets as OOD), scanner classification models excelled. This task trains models to be sensitive to texture and noise features directly affected by image quality degradation, making them ideal for detecting this type of OOD. For most anatomical shifts (SONAI (Other) and BCNatal (Other) datasets as OOD), on the other hand, plane classification models performed best. Training for anatomical plane recognition makes these models sensitive to broad anatomical differences, which is needed to identify images from entirely different anatomies as OOD. These further reinforce our central finding: OOD detection performance is not solely a function of the chosen uncertainty method, but is also determined by the interplay between the classification task and the specific ID-OOD shift.

Interestingly, scanner classification models also performed well in detecting anatomical shifts with BCNatal (Other) dataset as OOD. We hypothesize that this can be explained by considering the acquisition context of the BCNatal (Other) dataset. Some of the images in the BCNatal (Other) dataset were acquired with different ultrasound probes or acquisition protocols. This leads to detectable differences in image characteristics related to acquisition settings, which the scanner classification models are inherently sensitive to.

Our second important result is that **the best OOD performance does not imply the best abstinence performance.** Our investigation into abstained prediction performance in Sec. 3.3 revealed that optimal OOD detection performance does not necessarily translate to optimal performance for avoiding erroneous predictions in our clinical task of interest. Scanner classification models, which consistently demonstrated superior OOD detection across different ID-OOD shifts, actually underperformed in abstained prediction for our primary plane classification task compared to models trained directly for plane classification. This could reflect that a different datasets do not necessarily imply classifier failure. Models trained directly for plane classification, while potentially less sensitive to broader OOD shifts, appear to develop a more refined understanding of uncertainty within the plane classification task, enabling them to abstain more effectively while maintaining high accuracy on confident samples.

Limitations. First, our choice of UQ methods is limited. We have chosen to focus on UQ-methods that are easily available and widely used, while also being classifier-based to enable consistent experiments. Nevertheless, we hypothesize that our main point stands: There is no universal UQ method to rule them all.

Second, while OOD detection is usually approached using epistemic UQ, we include methods that also measure aleatoric uncertainty. We argue, however, that this makes sense for our applications: While anatomical shifts and image characteristic shifts like artifacts and shadows are clearly epistemic, other image

characteristic shifts, like blur or resolution, could be considered aleatoric. We thus find it appropriate also to include UQ methods with an aleatoric component.

Conclusion. Our findings have significant implications for the practical application and validation of uncertainty quantification. The choice of the most effective UQ-model is not universal and must be guided by the specific downstream application. For data shift monitoring systems, where the goal is to detect when the input data distribution has changed significantly enough to trigger model retraining, OOD detection performance should be prioritized. For clinical referral systems, where the aim is to automate the processing of confident cases while referring uncertain cases to clinicians, abstinence performance becomes the more critical metric for minimizing both errors in automated cases and clinician workload in abstained cases.

Acknowledgments. This work is funded by the Danish Pioneer Centre for AI (DNRF grant number P1) and SONAI - a Danish Regions' AI Signature Project. C.I.B. is funded via the EVUK program ("Next-generation Al for Integrated Diagnostics") of the Free State of Bavaria and partially supported by the Helmholtz Association under the joint research school 'Munich School for Data Science'. This work was in part supported by Berdelle-Stiftung (grant TimeFlow). It was also in part supported by the Munich Center for Machine Learning and the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, both sponsored by the Federal Ministry of Research, Technology and Space.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Berger, C., Paschali, M., Glocker, B., Kamnitsas, K.: Confidence-based out-of-distribution detection: a comparative study and analysis. In: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3. pp. 122–132. Springer (2021)
2. Burgos-Artizzu, X.P., Coronado-Gutiérrez, D., Valenzuela-Alcaraz, B., Bonet-Carne, E., Eixarch, E., Crispi, F., Gratacós, E.: Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. Scientific Reports **10**(1), 10200 (2020)
3. Combalia, M., Hueto, F., Puig, S., Malvehy, J., Vilaplana, V.: Uncertainty estimation in deep neural networks for dermoscopic image classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 744–745 (2020)
4. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
5. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)

6. Hong, Z., Yue, Y., Chen, Y., Cong, L., Lin, H., Luo, Y., Wang, M.H., Wang, W., Xu, J., Yang, X., et al.: Out-of-distribution detection in medical image analysis: A survey. arXiv preprint arXiv:2404.18279 (2024)
7. Huang, L., Ruan, S., Xing, Y., Feng, M.: A review of uncertainty quantification in medical image analysis: probabilistic and non-probabilistic methods. Medical Image Analysis p. 103223 (2024)
8. Kirchhof, M., Mucsányi, B., Oh, S.J., Kasneci, D.E.: Url: A representation learning benchmark for transferable uncertainty estimates. Advances in Neural Information Processing Systems **36**, 13956–13980 (2023)
9. Koch, L.M., Baumgartner, C.F., Berens, P.: Distribution shift detection for the postmarket surveillance of medical ai algorithms: a retrospective simulation study. NPJ Digital Medicine **7**(1), 120 (2024)
10. Kurz, A., Hauser, K., Mehrtens, H.A., Krieghoff-Henning, E., Hekler, A., Kather, J.N., Fröhling, S., von Kalle, C., Brinker, T.J., et al.: Uncertainty estimation in medical image classification: systematic review. JMIR Medical Informatics **10**(8), e36427 (2022)
11. Lahlou, S., Jain, M., Nekoei, H., Butoi, V.I., Bertin, P., Rector-Brooks, J., Korablyov, M., Bengio, Y.: Deup: Direct epistemic uncertainty prediction. arXiv preprint arXiv:2102.08501 (2021)
12. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems **30** (2017)
13. Lambert, B., Forbes, F., Doyle, S., Dehaene, H., Dojat, M.: Trustworthy clinical ai solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis. Artificial Intelligence in Medicine p. 102830 (2024)
14. Lee, S., Purushwalkam, S., Cogswell, M., Crandall, D., Batra, D.: Why m heads are better than one: Training a diverse ensemble of deep networks. arXiv preprint arXiv:1511.06314 (2015)
15. Li, Z., Kamnitsas, K., Islam, M., Chen, C., Glocker, B.: Estimating model performance under domain shifts with class-specific confidence scores. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 693–703. Springer (2022)
16. Linmans, J., van der Laak, J., Litjens, G.: Efficient out-of-distribution detection in digital pathology using multi-head convolutional neural networks. In: Medical Imaging with Deep Learning. pp. 465–478. PMLR (2020)
17. Mishra, D., Zhao, H., Saha, P., Papageorghiou, A.T., Noble, J.A.: Dual conditioned diffusion models for out-of-distribution detection: Application to fetal ultrasound videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 216–226. Springer (2023)
18. Mucsányi, B., Kirchhof, M., Oh, S.J.: Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. arXiv preprint arXiv:2402.19460 (2024)
19. Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P.H., Gal, Y.: Deep deterministic uncertainty: A new simple baseline. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24384–24394 (June 2023)
20. Sendra-Balcells, C., Campello, V.M., Torrents-Barrena, J., Ahmed, Y.A., Elattar, M., Ohene-Botwe, B., Nyangulu, P., Stones, W., Ammar, M., Benamer, L.N., et al.: Generalisability of fetal ultrasound deep learning models to low-resource imaging settings in five african countries. Scientific reports **13**(1), 2728 (2023)

21. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)
22. Tardy, M., Scheffer, B., Mateus, D.: Uncertainty measurements for the reliable classification of mammograms. In: International conference on medical image computing and computer-assisted intervention. pp. 495–503. Springer (2019)
23. Thagaard, J., Hauberg, S., van der Vegt, B., Ebstrup, T., Hansen, J.D., Dahl, A.B.: Can you trust predictive uncertainty under real dataset shifts in digital pathology? In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23. pp. 824–833. Springer (2020)
24. Van Amersfoort, J., Smith, L., Teh, Y.W., Gal, Y.: Uncertainty estimation using a single deep deterministic neural network. In: International conference on machine learning. pp. 9690–9700. PMLR (2020)
25. Yoo, D., Kweon, I.S.: Learning loss for active learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 93–102 (2019)
26. Zadorozhny, K., Thoral, P., Elbers, P., Cinà, G.: Out-of-distribution detection for medical applications: Guidelines for practical evaluation. In: Multimodal AI in healthcare: A paradigm shift in health intelligence, pp. 137–153. Springer (2022)
27. Zou, K., Chen, Z., Yuan, X., Shen, X., Wang, M., Fu, H.: A review of uncertainty estimation and its application in medical imaging. Meta-Radiology p. 100003 (2023)