Benchmarking Uncertainty and its Disentanglement in multi-label Chest X-Ray Classification

Simon Baur[1], Wojciech Samek[1,2,3], Jackie Ma[1]

[1] Fraunhofer Heinrich-Hertz-Institut, 10587 Berlin, Germany
{simon.baur, wojciech.samek, jackie.ma}@hhi.fraunhofer.de
[2] Technische Universität Berlin, 10623 Berlin, Germany
[3] The Berlin Institute for the Foundations of Learning and Data (BIFOLD), 10587 Berlin, Germany

Abstract. Reliable uncertainty quantification is crucial for trustworthy decision-making and the deployment of AI models in medical imaging. While prior work has explored the ability of neural networks to quantify predictive, epistemic, and aleatoric uncertainties using an information-theoretical approach in synthetic or well defined data settings like natural image classification, its applicability to real life medical diagnosis tasks remains underexplored. In this study, we provide an extensive uncertainty quantification benchmark for multi-label chest X-ray classification using the MIMIC-CXR-JPG dataset. We evaluate 13 uncertainty quantification methods for convolutional (ResNet) and transformer-based (Vision Transformer) architectures across a wide range of tasks. Additionally, we extend Evidential Deep Learning, HetClass NNs, and Deep Deterministic Uncertainty to the multi-label setting. Our analysis provides insights into uncertainty estimation effectiveness and the ability to disentangle epistemic and aleatoric uncertainties, revealing method- and architecture-specific strengths and limitations.

Keywords: Uncertainty Quantification · Disentanglement · Radiology

1   Introduction

Deep neural networks have widely been used to solve medical image analysis tasks, such as image classification and segmentation [17,30]. However, in high-stakes applications such as radiology, models must meet rigorous standards to ensure reliability and trustworthiness. These requirements are developed from multiple perspectives, including the needs of patients, medical professionals, and regulatory agencies. From a technical standpoint, comprehensive benchmarks are essential to guide the development of safe and interpretable AI models. In radiology, AI-assisted diagnosis can reach human-expert level performance [22], but the expressivity of uncertainty of such models is often not intensively studied. A false negative AI classification of pneumonia, for example, could delay life-saving

treatment, while overconfident AI predictions could lead to unnecessary interventions. Regulatory agencies such as the FDA [26] emphasize the importance of uncertainty-aware AI systems, yet robust uncertainty quantification remains underexplored in multi-label medical imaging [16]. Despite the increasing focus in benchmarking AI models on performance, robustness, and explainability [24,4], uncertainty quantification (UQ)[9,1] remains an underexplored aspect of model evaluation. Reliable UQ methods are crucial for assessing predictive confidence, and ultimately improving model trustworthiness in clinical practice [2,8]. Our work addresses this research gap by systematically benchmarking a diverse set of uncertainty quantification methods for multi-label chest X-ray classification. We provide a comprehensive evaluation across convolutional and transformer-based architectures, offering practical insights into their reliability and applicability in real-world medical imaging workflows.

1.1   Related Work

Benchmarking Uncertainty Quantification. Our work follows the benchmarking setup recently proposed by Mucsányi et al. [19] for uncertainty quantification in natural imaging, which we adapt to a multilabel setting. Their framework benchmarks UQ methods across five tasks, assessing their ability to disentangle uncertainty and evaluating various aggregation strategies.

Distributional Methods. Distributional methods model uncertainty by learning a second-order predictive distribution $p(y \mid x, \theta)$, where the model outputs a distribution over predictions rather than a point estimate. Practically, M forward passes are sampled for each input x to approximate $p(y \mid x, \theta)$ [27,28]. We evaluate a diverse set of uncertainty quantification methods, including Latent Heteroscedastic Classifier (HET-XL) [5], Swag [18], MC Dropout (MC-D) [25], Heteroscedastic Classification Neural Network (Het-NN) [6], Deep Ensemble (D-Ens) [14], Shallow Ensemble (SE) [15], Evidential Deep Learning (EDL) [23], Masked Attention (M-Attn) [3], and Gumbel Softmax (GS-Attn) [21]. The latter two are only applicable to transformers.

Deterministic Methods. Deterministic uncertainty methods estimate uncertainty without requiring a distribution over predictions. These approaches often rely on model-internal features to assess confidence in predictions. We evaluate several deterministic methods, namely Loss Prediction (LP) [29], Correctness Prediction (CP), Temperature Scaling (Temp) [10] and Deep Deterministic Uncertainty (DDU) [20].

Information-Theoretical Approach and Disentanglement. The information-theoretical (IT) formulation of uncertainty [7,11] models predictive uncertainty as a combination of epistemic uncertainty (EU), arising from limited data or model capacity, and aleatoric uncertainty (AU), reflecting inherent noise in the

data. While this decomposition has been thoroughly investigated in controlled and synthetic settings, recent studies [13,19] raise concerns about the practical effectiveness of the IT framework in complex, real-world datasets, questioning its ability to reliably disentangle epistemic and aleatoric uncertainties, and therefore its effectiveness of modeling uncertainty in total.

1.2   Contribution

Our key contributions can be summarized as follows:

1. We present the first extensive empirical benchmark of 13 uncertainty methods on chest X-ray data, spanning multiple tasks and both ConvNets and Vision Transformers.
2. We adapt three UQ methods—EDL, Het NN, and DDU—from multiclass to multilabel classification.
3. Our analyses highlight limitations in disentangling EU and AU using the IT framework. We infer that the IT framework is to be used with caution.

2   Theoretical Framework and Methods

In radiology, epistemic uncertainty stems from limited knowledge or distribution shifts, affecting underrepresented or unseen cases and signaling the need for expert review or more data. Aleatoric uncertainty, originating from inherent data variability (e.g. motion artifacts), can be used to flag images for re-acquisition. Methods that are reliably able to distinguish EU and AU could therefore significantly advance AI deployment in diagnostic applications.

2.1   Uncertainty Disentanglement

To derive uncertainty estimates $u(x)$ for an input x we focus on the information-theoretical (IT) [7,11] approach of decomposing total predictive uncertainty (PU) into aleatoric uncertainty (AU) and epistemic uncertainty (EU). The model's predictions are obtained via the Bayesian-Model-Average (BMA):

$$\text{BMA} \equiv \bar{\pi}(x) \equiv \frac{1}{M} \sum_{m=1}^{M} \pi^{(m)}, \tag{1}$$

$$\text{PU} \equiv \mathbb{H}\left(\bar{\pi}(x)\right), \tag{2}$$

$$\text{AU} \equiv \frac{1}{M} \sum_{m=1}^{M} \mathbb{H}\left(\pi^{(m)}\right), \tag{3}$$

$$\text{EU} \equiv \mathbb{H}\left(\bar{\pi}(x)\right) - \frac{1}{M} \sum_{m=1}^{M} \mathbb{H}\left(\pi^{(m)}\right), \tag{4}$$

where M is the number of sampled forward passes, π is the class probability output vector and $\mathbb{H}(x)$ is the entropy.

2.2   Multilabel Adaptions

We adapt EDL, Het-NN and DDU, methods originally designed for multiclass tasks, to the multilabel setting. For a detailed mathematical description see **??**.

3   Tasks, Experimental Setting and Results

3.1   Tasks

We evaluate all methods mentioned in 1.1 on 6 different tasks on the MIMIC-CXR-JPG [12] dataset (official train/val/test split) which is annotated for 14 pathologies with labels being either 0 (negative), 1 (positive), or -1 (uncertain). We further provide correlation analysis of aleatoric and epistemic uncertainty as well as insights into capabilities of uncertainty disentanglement of each method.

Task 1: OOD-Detection. Out-of-distribution detection assesses a model's ability to capture epistemic uncertainty. We construct an OOD dataset by randomly sampling unaltered images from MIMIC-CXR-JPG (assigning label 0, marking ID data) and applying a combination of transformations (Gaussian noise, motion blur, vignettes, mask occlusion, etc.) to an equal number of images (assigning label 1, marking OOD data). These transformations induce a performance drop of $\sim 20\%$ on validation AUROC, similar to the difference between ImageNet and ImageNet-C. OOD detection performance is evaluated using binary AUROC, where a higher score reflects better separation between ID and OOD samples.

Task 2: Uncertainty Label Prediction. MIMIC-CXR-JPG includes class labels for each class annotated as negative (0), positive (1), or uncertain (-1), where the latter is therefore providing a source of expert-labeled ground truth uncertainties. To assess aleatoric uncertainty, we evaluate how well the model's uncertainty scores align with these ground truth uncertainties. Specifically, we filter the test set to include only samples with at least one uncertain (-1) label across all classes, mapping -1 to 1 (effectively marking the class as uncertain) and both 0 and 1 to 0 (effectively marking the class as certain). We then compute the binary AUROC per class and report the macro-average AUROC, where higher values indicate better identification of ground truth uncertainties.

Task 3: Correctness Prediction. To assess the model's ability to judge its own prediction reliability, we infer correctness labels through assigning 1 to correct predictions (threshold 0.5 for positives) and 0 to incorrect ones. We then compute binary AUROC per class between estimated UQ scores and correctness labels, and report the macro-average AUROC, where a higher score indicates better distinction between correct and incorrect predictions.

Task 4: Abstained Prediction. To assess the impact of uncertainty on accuracy, we iteratively remove the 5% most uncertain samples (as determined by estimated UQ scores) and track accuracy, effectively making the test set more certain in each iteration. We compute the Area Under the Accuracy Coverage Curve (AUAC), where higher values indicate better uncertainty-informed abstention.

Task 5 and Task 6: Calibration. For calibration assessment, we compute Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) per class. We report the macro-average ECE and MCE across all classes, where lower values indicate better calibration.

3.2   Experimental Setting.

We trained each model for 50 epochs with early stopping of 5 based on validation AUROC. Learning rate was initialized with 1e-4 and decreased with a cosine annealing learning rate scheduler. We used the AdamW optimizer with a weight decay of 0.1. We trained five models per method with different seeds. For distributional methods, we performed M=5 forward passes per model, totaling 25 samples per input. We report the mean and standard error across all models and forward passes. We chose $M = 5$ to keep our experiments computationally efficient and because higher Ms did not show significant improvements [19]. We chose ViT-tiny and ResNet18 models due to their comparable parameter sizes, not utilizing larger models as they yielded only minor performance improvements on validation metrics. We initialized all models with weights pretrained on ImageNet. We retained method-specific hyperparameters as reported from the best-performing runs in [19].

3.3   Results

Fig.1 summarizes results across Tasks 1-6 and reports the best performing of PU, AU and EU scores for each method. Task 1 exhibits a wide performance range (0.55–0.99), implicating a strong impact of method choice on OOD detection. For ViT, S-Ens and D-Ens perform best (only two methods achieving > 0.9), while ResNet benefits most from S-Ens, D-Ens, EDL, and SWAG, all achieving scores above 0.9. Distributional methods in general are outperforming deterministic methods across model architectures. Task 2 scores are varying less (almost all methods scoring between 0.58–0.62) compared to Task 1, with D-Ens and DDU being the top performing. Swag on ResNet is an outlier towards the bottom (0.52), indicating poor capability of capturing aleatoric uncertainties. Notably more deterministic methods lie within the upper performance range compared to Task 1. Task 3 again displays a narrow range for most methods (0.7-0.75). Again, Ensemble based methods (D-Ens, S-Ens) are among top scorers. CP, which specifically is designed to solve Task 3 shows high performance, and therefore aligns well with it's intention. DDU and Swag for ResNet show
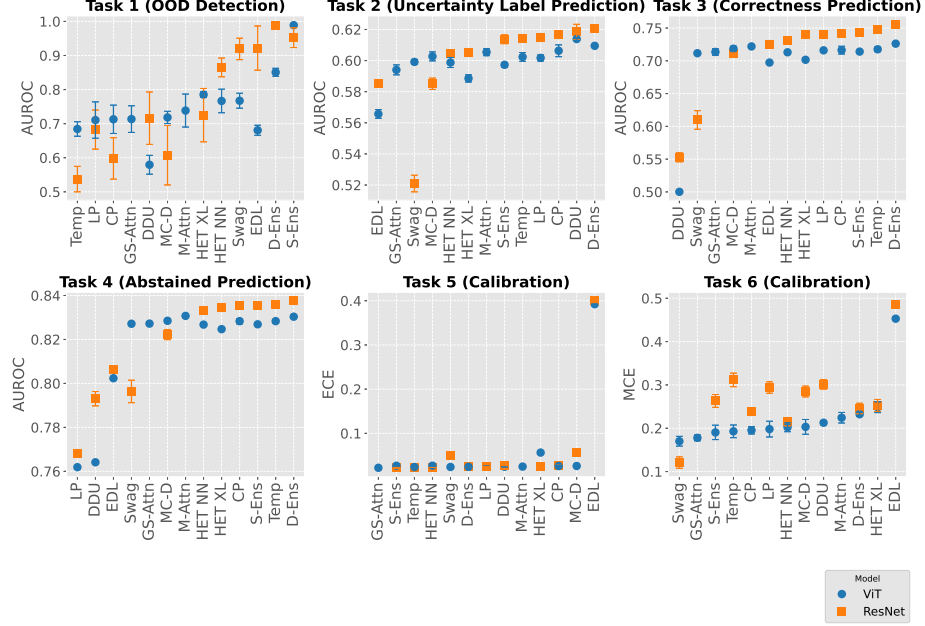
Fig. 1. Results for Tasks 1-6: ViT-Tiny (blue) and ResNet-18 (orange). For Tasks 1–4, we computed three uncertainty scores per method—predictive (PU), epistemic (EU), and aleatoric (AU)—as defined in 2.1. Each score yields a separate AUROC, and we report only the highest per method. Task 5 and 6 are evaluated based on BMA scores.

significantly lower performance than other methods here. Task 4 paints a similar picture as Task 3 (most ranging from 0.82–0.84), this time lower outliers being LP (both) and DDU (ViT), indicating weak alignment with accuracy for those methods. Again, D-Ens and S-Ens are top scorers. Results for Task 5 suggest almost all methods are well calibrated w.r.t ECE,(ECE < 0.04), except EDL (ECE = 0.4) which is strongly ill-calibrated for both models. Calibration in terms of MCE paints a more diverse picture: for ViT, most methods align around 0.2, whereas ResNet shows strong differences, with Swag being best (0.12) and almost all other methods displaying a MCE > 0.25. EDL is again heavily miscalibrated. Results across both calibration tasks therefore imply that swag is leading across calibration metrics and architectures, while EDL has significant trouble in achieving good calibration.

We now go on to analyze the promised disentanglement capabilities of the IT approach. As a first step we investigate rank correlations of EU and AU scores on Task 1 and 2, which represent tasks aimed to be solved by the respective scores. If the promise of the IT approach holds, EU and AU scores should not display significant correlations. However, almost all methods show significant correlations between EU and AU. See **??** for detailed correlation metrics. We therefore can confirm prior findings of e.g. [19] that UQ scores as disentangled

by the IT approach do not keep their promise of uncorrelated uncertainties. Given strong correlation between EU and AU scores —despite their intended
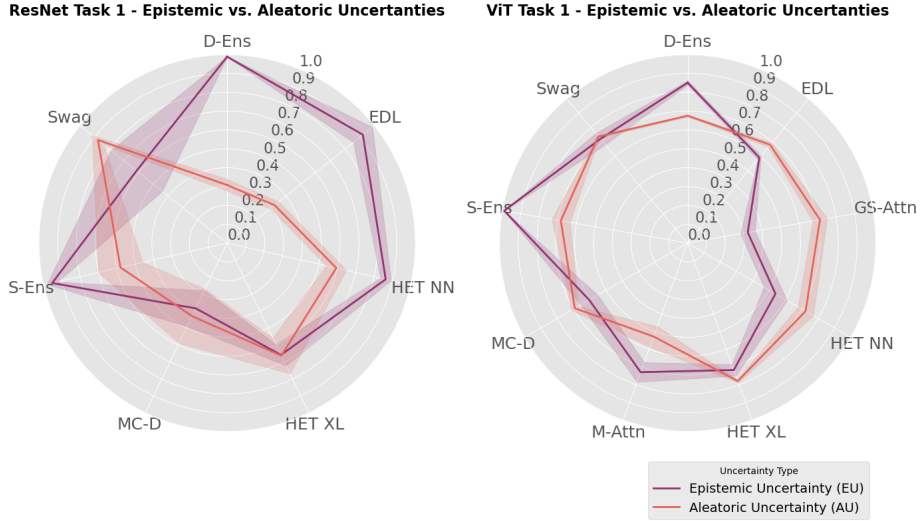
Fig. 2. Task 1 (OOD-Detection) AUROC scores for epistemic (purple) and aleatoric (red) uncertainty. Capability to disentangle is greater when epistemic uncertainty outperforms aleatoric uncertainty (i.e. purple line is further outwards than red line).

non-correlation— we evaluate how well uncertainty estimates perform on their respective target tasks as final step of our analysis. We do this exemplary for Task 1 (OOD detection). Fig.2 compares EU and AU estimators on Task 1. Only distributional methods are shown, as deterministic methods yield EU = 0 per 2.1. Note that AUROC scores here do not necessarily align with AUROC scores of Task 1 in Fig.1, as it can also include AUROC derived from PU estimates, whereas in Fig.2 we solely take EU and AU scores into account. Ideally, EU should significantly outperform AU in this task. However, many methods fail to disentangle uncertainty types, with AU estimates performing comparably or better. For ViT, only D-Ens, S-Ens, and M-Attn outperform aleatoric estimates. In ResNet, D-Ens, S-Ens, EDL, and Het-NN show significantly stronger epistemic performance. To synthesize our prior analyses, we aggregate metrics from Fig.1 and Fig.2 as well as rank correlations into an overview given in Fig.3. Ensemble-based methods (D-Ens, S-Ens) consistently perform best across architectures, with HET-NN (ResNet) and M-Attn (ViT) coming closest in performance. Our results motivate several practical implications, discussed in the following section 4.
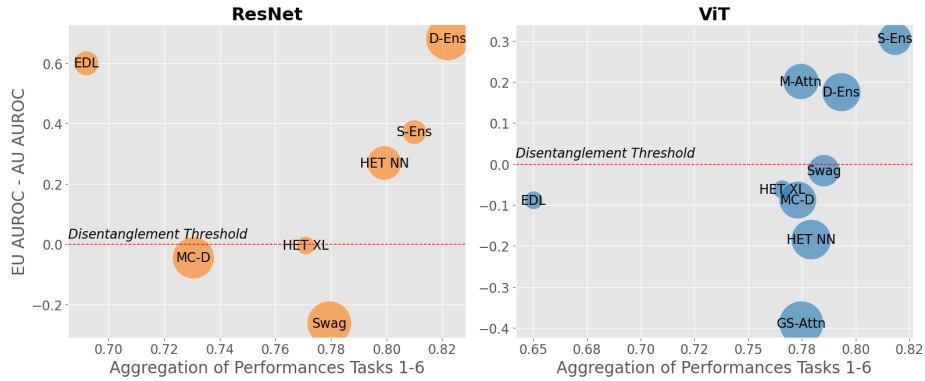
Fig. 3. Comparison of average performance across all tasks (x-axis; see Fig. 1) and disentanglement capability (y-axis), measured as $AUROC(\text{EU}) - AUROC(\text{AU})$ (see Fig. 2). Y-values above the disentanglement threshold indicate that EU and AU are well separated, with EU better aligned to the epistemic task. Values below the threshold suggest an inverse relationship, where AU unexpectedly performs better on the epistemic task. Bubble size reflects the inverse correlation between EU and AU scores; larger bubbles indicate weaker correlation. Methods in the upper-right region combine strong performance with well-separated uncertainty types.

4   Conclusion and Practical Implications

We studied 13 UQ methods across two architectures—ResNet and ViT—and six tasks in a multilabel medical setting, establishing a comprehensive benchmark of UQ performance. Additionally, we assessed distributional methods' ability to decompose uncertainties using the information-theoretical approach. Based on our findings, we offer the following recommendations for clinical practitioners:

Align models and methods with tasks. We observed substantial variability in UQ method performance across tasks and architectures, highlighting the need to align the method with the specific UQ objective. For OOD detection, distributional (bayesian) methods clearly outperform deterministic ones and should be preferred. Notably, Swag with ResNet is highly effective for OOD, while performing poorly on aleatoric tasks—a potentially useful trait if understood. Conversely, for modeling aleatoric uncertainty, simple deterministic methods like LP and CP are surprisingly effective, performing nearly on par with ensembles.

Architectures matter for calibration. For well-calibrated predictions, practitioners should consider both the UQ method and the model architecture. ViTs generally outperform ResNets in calibration (MCE) across methods. SWAG remains a strong choice, while EDL should be used with caution, as it appears consistently ill-calibrated.

Ensembles remain the most reliable. Ensemble based UQ methods (D-Ens, S-Ens) consistently outperform other approaches across tasks while disentangling effectively in comparison and should therefore be preferred whenever possible.

Understand Limits of IT-Based Disentanglement. We showed that disentangling uncertainties via the information-theoretical approach can fall short of its theoretical expectations. If the IT framework holds, EU and AU should be uncorrelated and task-specific. However, in practice disentangled scores do not always align with their intended tasks—e.g., EU scores can counterintuitively perform worse on OOD detection than their AU and PU counterparts. We therefore recommend practitioners to apply the IT framework with caution and encourage further research assessing limitations and practical reliability.

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

References

1. Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., et al.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information fusion **76**, 243–297 (2021)
2. Abdullah, A.A., Hassan, M.M., Mustafa, Y.T.: A review on bayesian deep learning in healthcare: Applications and challenges. IEEe Access **10**, 36538–36562 (2022)
3. Baur, S., Vahidi, A., Wang, M., Zebardast, N., Elze, T., Bischl, B., Rezaei, M., Eslami, M.: Advancing reliability in self-supervised transformer models through hierarchical mask attention heads. In: Medical Imaging 2025: Image Processing. vol. 13406, pp. 152–159. SPIE (2025)
4. Clark, B., Wilming, R., Dox, A., Eschenbach, P., Hached, S., Wodke, D.J., Zewdie, M.T., Bruila, U., Oliveira, M., Schulz, H., Cornils, L.M., Panknin, D., Boubekki, A., Haufe, S.: Exact: Towards a platform for empirically benchmarking machine learning model explanation methods. Measurement: Sensors p. 101794 (2025). `https://doi.org/https://doi.org/10.1016/j.measen.2024.101794`, `https://www.sciencedirect.com/science/article/pii/S2665917424007700`
5. Collier, M., Jenatton, R., Mustafa, B., Houlsby, N., Berent, J., Kokiopoulou, E.: Massively scaling heteroscedastic classifiers. arXiv preprint arXiv:2301.12860 (2023)
6. Collier, M., Mustafa, B., Kokiopoulou, E., Jenatton, R., Berent, J.: Correlated input-dependent label noise in large-scale image classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1551–1560 (2021)
7. Depeweg, S., Hernandez-Lobato, J.M., Doshi-Velez, F., Udluft, S.: Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In: International conference on machine learning. pp. 1184–1193. PMLR (2018)
8. Evans, H., Snead, D.: Understanding the errors made by artificial intelligence algorithms in histopathology in terms of patient impact. NPJ Digital Medicine **7**(1), 89 (2024)
9. Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al.: A survey of uncertainty in deep neural networks. Artificial Intelligence Review **56**(Suppl 1), 1513–1589 (2023)
10. 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)
11. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine learning **110**(3), 457–506 (2021)
12. Johnson, A.E.W., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., ying Deng, C., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs (2019), `https://arxiv.org/abs/1901.07042`
13. Kahl, K.C., Lüth, C.T., Zenk, M., Maier-Hein, K., Jaeger, P.F.: Values: A framework for systematic validation of uncertainty estimation in semantic segmentation. arXiv preprint arXiv:2401.08501 (2024)
14. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems **30** (2017)

15. Lee, S., Purushwalkam, S., Cogswell, M., Crandall, D.J., Batra, D.: Why M heads are better than one: Training a diverse ensemble of deep networks. CoRR **abs/1511.06314** (2015), http://arxiv.org/abs/1511.06314
16. Leibig, C., Allken, V., Ayhan, M.S., Berens, P., Wahl, S.: Leveraging uncertainty information from deep neural networks for disease detection. Scientific reports **7**(1), 1–14 (2017)
17. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017). https://doi.org/https://doi.org/10.1016/j.media.2017.07.005, https://www.sciencedirect.com/science/article/pii/S1361841517301135
18. Maddox, W.J., Izmailov, P., Garipov, T., Vetrov, D.P., Wilson, A.G.: A simple baseline for bayesian uncertainty in deep learning. Advances in neural information processing systems **32** (2019)
19. Mucsányi, B., Kirchhof, M., Oh, S.J.: Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. In: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2024), https://openreview.net/forum?id=x8RgF2xQTj
20. 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. pp. 24384–24394 (2023)
21. Pei, J., Wang, C., Szarvas, G.: Transformer uncertainty estimation with hierarchical stochastic attention (2022)
22. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D.Y., Bagul, A., Langlotz, C.P., Shpanskaya, K.S., Lungren, M.P., Ng, A.Y.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. CoRR **abs/1711.05225** (2017), http://arxiv.org/abs/1711.05225
23. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. Advances in neural information processing systems **31** (2018)
24. Springenberg, M., Frommholz, A., Wenzel, M., Weicken, E., Ma, J., Strodthoff, N.: From modern cnns to vision transformers: Assessing the performance, robustness, and classification strategies of deep learning models in histopathology. Medical Image Analysis **87**, 102809 (2023)
25. 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)
26. U.S. Food and Drug Administration: Evaluation methods for artificial intelligence (ai)-enabled medical devices: Performance assessment and uncertainty quantification (2025), accessed on February 27, 2025
27. Wang, H., Yeung, D.Y.: A survey on bayesian deep learning. ACM computing surveys (csur) **53**(5), 1–37 (2020)
28. Wilson, A.G., Izmailov, P.: Bayesian deep learning and a probabilistic perspective of generalization. Advances in neural information processing systems **33**, 4697–4708 (2020)
29. 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)
30. Zhou, S.K., Greenspan, H., Davatzikos, C., Duncan, J.S., Van Ginneken, B., Madabhushi, A., Prince, J.L., Rueckert, D., Summers, R.M.: A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE **109**(5), 820–838 (2021)