A framework for assessing joint human-AI systems based on uncertainty estimation

Emir Konuk[†1], Robert Welch[†1], Filip Christiansen[1,2], Elisabeth Epstein[2], and Kevin Smith[1]

[1] KTH, Stockholm, Sweden
ekonuk@kth.se
[2] Karolinska Institutet, Stockholm Sweden

Abstract. We investigate the role of uncertainty quantification in aiding medical decision-making. Existing evaluation metrics fail to capture the practical utility of joint human-AI decision-making systems. To address this, we introduce a novel framework to assess such systems and use it to benchmark a diverse set of confidence and uncertainty estimation methods. Our results show that certainty measures enable joint human-AI systems to outperform both standalone humans and AIs, and that for a given system there exists an optimal balance in the number of cases to refer to humans, beyond which the system's performance degrades.

Keywords: Uncertainty · Selective Classification · Ultrasound

1 Introduction

Both AI and humans are susceptible to errors, albeit of different types [16]. This diversity in errors presents a unique opportunity: by leveraging notions of confidence or uncertainty, we can synergize AI with human expertise to enhance the joint decision-making process. In a clinical setting, collaborative human-AI systems have the potential to surpass the effectiveness of both doctors or AI working in isolation. But currently we lack meaningful metrics to evaluate the performance of such joint systems.

Consider the scenarios depicted in Figure 1 where the goal is to examine an image x and identify if the patient has cancer. A standalone AI model can make a prediction, $\hat{y}$, as to whether cancer is present or not. Adding some notion of confidence or uncertainty to the model, ψ[3], provides an estimate of the model's reliability (Figure 1b). However, this setup fails to capture the practical impact of a certainty estimate. A pragmatic use case is depicted in Figure 1c, where certainty is used for *selective classification*. Here, the certainty estimate is used

[3] Confidence and uncertainty are related but distinct concepts. Briefly, confidence is an inherent property of virtually any modern probabilistic classifier, *i.e.* the probability output is a form of model confidence. The term "uncertainty" is usually reserved for Bayesian models which can yield multiple stochastic predictions for each sample. We use the term 'certainty', denoted by ψ, to refer to either confidence or uncertainty.

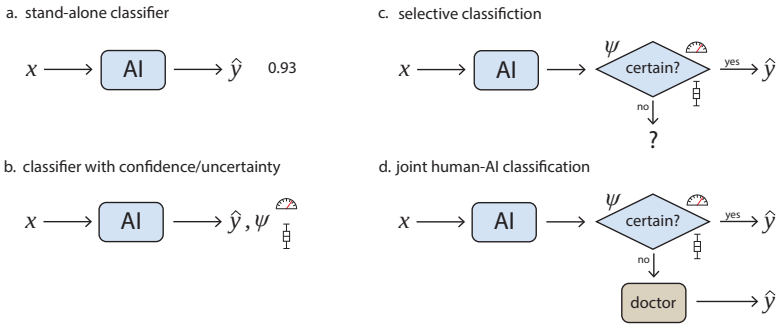

Fig. 1. Different approaches to integrate AI in the diagnostic process.

to decide whether the model should abstain to avoid making a wrong prediction – if the model's confidence or uncertainty value crosses a threshold, it abstains. However, selective classification, which has seen a recent rise in popularity [11], fails to account for what happens to the samples for which the model abstains. Instead, we should consider a joint human-AI system, such as the one depicted in Figure 1d, which refers AI-uncertain cases to a doctor. *Critically, no evaluation framework exists to assess such joint human-AI diagnostic systems.*

In this study, we address this need by proposing a novel evaluation framework for assessing joint human-AI classification systems. Our framework assesses how AI models, equipped with confidence or uncertainty measures, interact with doctors according to clinically informed metrics. By measuring the joint system risk (misclassification rate) and F1, we can assess its performance for various levels of certainty – a variable which can be adjusted according to the availability of doctors. Through extensive evaluation on a unique ultrasound dataset with multiple doctor assessments per case, we find that when an appropriate confidence threshold is chosen, joint systems outperform standalone humans and AIs. We consider different certainty estimation techniques in this joint setting, and while we find some differences between the methods, we do not find one to clearly dominate. Importantly, our experiments show that *there is an optimal balance in the number of cases to refer to humans, beyond which the system's performance degrades – i.e.it is not always beneficial to defer to doctor assessments.* Finally, we apply our proposed framework using various measures of confidence or uncertainty and examine the results.

2 Related Work

There exist two schools of thought on how to evaluate model confidence and uncertainty: (1) to assess the quality of the estimate itself, or (2) to assess the utility of the estimate, as depicted in Figure 1. In the first approach [15,17], metrics such as the Expected Calibration Error (ECE) [9], Negative Log-Likelihood (NLL), and Brier score [3] serve as valuable indicators of the AI's reliability.

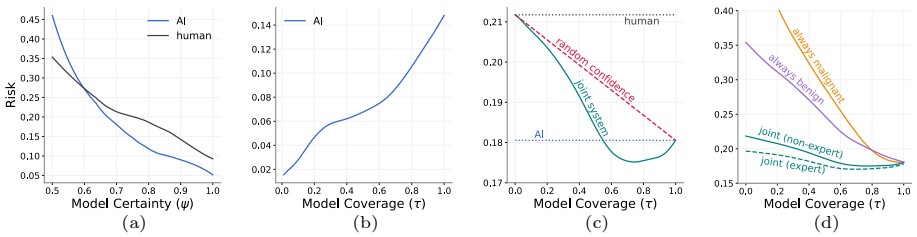

Fig. 2. Certainty estimates in joint human-AI systems. *(a)* AI and doctor performance varies according to model certainty – doctors outperform the model when its certainty is low. *(b)* Naively computing risk coverage without considering humans in the system (as in AURC) is problematic – it appears optimal for the model to always abstain. *(c)* The risk coverage of the joint system has a clear optimal operating point. *(d)* The joint system's optimal point can vary depending on the quality of human predictions.

Better scores indicate the model's alignment with the true data distribution. Proponents of the utility evaluation approach point out that these metrics do not capture the estimate's impact on performance [6,11,8], which is often of most practical interest. They argue for measuring the area under the receiver operating characteristic curve of a misclassification detector, AUC_{mis},or the selective classification performance using the area under the risk coverage curve, AURC. However, AUC_{mis} is blind to the actual classification performance [6,11] and AURC is an aggregate metric that treats all possible abstain rates the same [8] when only certain ranges are of practical interest. But more importantly, these approaches fail to account for what happens to the samples the model abstains from predicting.

Although informative, these metrics are not fully appropriate for evaluating certainty estimates in clinical contexts, as seen in conflicting studies. Bungert *et al.* [4] claimed current methods fall short in clinical reliability based on AURC evaluations, whereas Alves *et al.* [1] found ensemble-based uncertainty estimates promising for misclassification detection. These discrepancies underscore the need for a comprehensive evaluation framework for joint human-AI systems, a gap in the literature which our study aims to address.

3 Joint Human-AI Evaluation Framework

Intuitively, we understand that AI with low certainty, *i.e.* low ψ, correlates with higher error rates. We observe in Figure 2a that a similar correlation is also true for humans. Importantly, there is a crossover point at which the human error rate is lower than the AI's. This can be exploited to achieve better clinical outcomes by referring patients with low certainty ψ to a human reader. But naively trying to accomplish this using risk coverage for the selective classification model, like AURC, is problematic because it fails to consider the abstained samples. This results in a monotonically increasing risk function, incorrectly suggesting that the optimal strategy for the model is to always abstain (Figure 2b).

The evaluation framework should consider humans and AI as a joint system. The AI is comprised of a classifier and a certainty estimator,

$$\hat{y} = \arg\max_{y} P_m(y|x,\theta) \quad \text{and} \quad \psi = \Phi[P_m(y|x,\theta)] \tag{1}$$

where $\hat{y}$ is the predicted class, θ are the weights of the classifier P_m, and Φ is a function that estimates the confidence or uncertainty of a prediction, ψ. For the AI-as-first-reader scenario in Figure 1d, if the certainty ψ is higher than a preset threshold, $\psi > \alpha$, a prediction is made by the AI. Otherwise, the patient is sent to a human reader who makes the prediction. The certainty threshold α is a hyperparameter of the joint-human AI system. It determines the model coverage, τ, the proportion of instances for which the model makes predictions as opposed to abstaining because of high uncertainty or low confidence.

Crucially, we evaluate the correctness of *all* predictions, irrespective of who or what made them. We measure the risk-coverage of the joint system,

$$\text{JRC}(\tau) = R_m(\tau) + R_h(\tau) \tag{2}$$

where $R_m(\tau)$ and $R_h(\tau)$ are the risk, *i.e.* error-rate, of the model and humans at a model coverage τ. The result is a *joint risk-coverage curve*, as seen in Figure 2c. When $\tau = 1$, only the AI makes predictions and the joint risk reduces to the risk of the model alone, $R_m(1)$. When $\tau = 0$, only humans make predictions, so the joint risk reduces to $R_h(0)$. A poor certainty estimate, which assigns random values to confidence or uncertainty, yields a straight line over different values of τ (the dashed red line). A good certainty estimate sorts cases to ensure low risk for both AI and humans. It has the potential to perform better than either AI or humans alone, as depicted by the green curve. The joint system performance can be improved by enhancing either the classifier, the humans, or the uncertainty estimate. Figure 2d shows the effects of improving human performance.

Furthermore, clinical tasks must consider if the errors are Type-I or Type-II. Thus, we measure the joint true positives and false positives and negatives to calculate the joint F1, *i.e.*, the harmonic mean of precision and recall, at coverage τ,

$$\text{JF1C}(\tau) = \frac{2 * \text{TP}_j(\tau)}{2 * \text{TP}_j(\tau) + \text{FP}_j(\tau) + \text{FN}_j(\tau)} \tag{3}$$

Finally, we provide summary metrics that compute the partial area under the JRC and JF1C curves. These consider a range of pertinent coverage values $\tau \in [\gamma, 1.0]$, with suggested values of $\gamma = \{0.5, 0.75, 0.9\}$.

$$\text{pAUJF1C}_\gamma = \int_\gamma^{1.0} \frac{2 * \text{TP}_j(\tau)}{2 * \text{TP}_j(\tau) + \text{FP}_j(\tau) + \text{FN}_j(\tau)} d\tau \tag{4}$$

and partial area under the JRC curve score, given by

$$\text{pAUJRC}_\gamma = \int_\gamma^{1.0} R_m(\tau) + R_h(\tau) d\tau \tag{5}$$

4 Experiments

Data. Our study utilized Ovarian tumour Machine Learning Collaboration - Retrospective Study (OMLC-RS) dataset comprising 17,119 ultrasound images from 3,652 patients, gathered from 20 centers across eight countries using 21 different ultrasound systems. The task is to classify ovarian tumors in ultrasound images. The ground truth for model training and evaluation was histological diagnosis. A total of 66 doctors, including 33 experts with over five years of experience, assessed the exams. Each exam was assessed by approximately 14 doctors.

Confidence estimates. We consider two approaches for estimating the certainty of AI: confidence-based and entropy-based. These approaches can be applied to the various model types described below. The confidence measure ψ is simply the maximum probability assigned by the model P_m,

$$\psi = \Phi[P_m(y|x,\theta)] = \max_y P_m(y|x,\theta) \tag{6}$$

Uncertainty estimates. Alternatively, uncertainty can be estimated using the entropy of the predicted probabilities,

$$\psi = \Phi[P_m(y|x,\theta)] = -\mathcal{H}[P_m(y|x,\theta)] = \sum_y P_m(y|x,\theta) \ln P_m(y|x,\theta) \tag{7}$$

Higher entropy indicates lower certainty ψ. In both Eq. 6 and Eq. 7, point estimates are used to estimate certainty.

Some of the models we consider are Bayesian NN approximations, *i.e.* a distribution (or a set) of weights from which we can sample a model in order to make a prediction, $\theta \sim p(\theta|\mathcal{D})$, where $\mathcal{D}$ is the dataset. For these models, the entropy of the average of multiple predictions is called the total predictive entropy and we use its negative as our certainty estimate,

$$\psi = \Phi[P_m(y|x,\theta)] = -\mathcal{H}[\mathbb{E}_{p(\theta|\mathcal{D})}[P_m(y|x,\theta)]] \tag{8}$$

Under some strong assumptions, we can further decompose predictive entropy into epistemic and aleatoric components [7]. We again use their negatives.

$$\psi = \Phi[P_m(y|x,\theta)] = -\mathbb{E}_{p(\theta|\mathcal{D})}[\mathcal{H}[P_m(y|x,\theta)]] \tag{9}$$

$$\psi = \Phi[P_m(y|x,\theta)] = \mathbb{E}_{p(\theta|\mathcal{D})}[\mathcal{H}[P_m(y|x,\theta)]] - \mathcal{H}[\mathbb{E}_{p(\theta|\mathcal{D})}[P_m(y|x,\theta)]] \tag{10}$$

AI models. We apply the certainty estimates described above to a variety of neural networks: some standard classifiers, and some built with uncertainty estimation in mind. All use an ImageNet [5] pretrained ResNet50 [10] backbone. We train the models image-wise but evaluate patient-wise. We consider the following models: *(1)* Maximum Likelihood Estimation (MLE) training, *i.e.* the standard deep learning approach using cross-entropy loss; *(2)* Temperature Scaling (TS) [9] where the model's probabilities are calibrated using temperature scaling on the validation set; *(3)* Test Time Augmentation (TTA) [2] where we apply 128

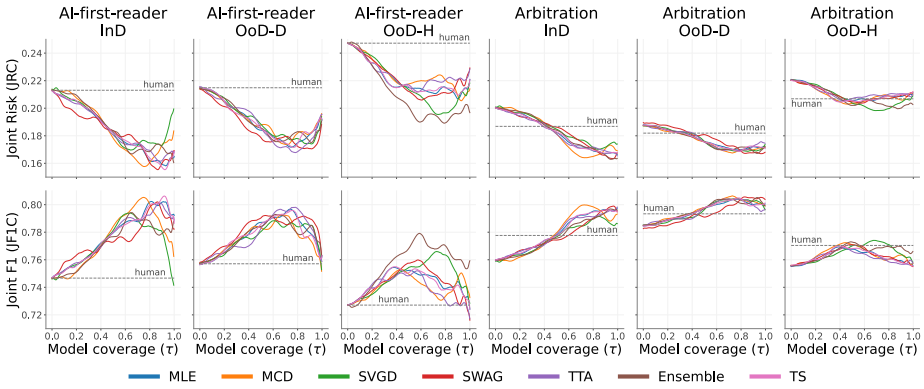

Fig. 3. Joint coverage curves for assessing the performance of human-AI systems.

different image augmentations at inference and use the median prediction; *(4)* Monte Carlo Dropout (MCD) [7]–a Bayesian NN approximation which makes multiple predictions using different dropout configurations; *(5)* Deep Ensembles [12] which uses an ensemble of NNs[4]; *(6)* Stein Variational Gradient Descent (SVGD) [13], a modified gradient descent method for training a diverse ensemble of NNs jointly; *(7)* Stochastic Weight Averaging Gaussian (SWAG) [14], we fit a normal distribution over weights using model checkpoints.

InD and OoD experiments. Using the models, certainty measures, and data described above, we conduct the following experiments. First, we assess how the joint human-AI systems perform when deployed in-domain (InD). For this, the dataset was divided into train, validation and test sets (3097/277/278 samples) using stratified sampling across all centers on a patient-wise basis. We also investigate how the joint human-AI systems perform when deployed in out-of-domain (OoD) settings. We consider both unseen hospitals (OoD-H) and unseen ultrasound devices (OoD-D). For OoD-H, we leave out 2 hospitals as the test set, and for OoD-D we leave out 15 devices as the test set (GE Voluson devices dominate our dataset, and are used for training and validation). For all experiments, for each of the 7 AI model types we report results using the confidence or uncertainty measure that yields the lowest joint risk for any threshold α.

Arbitration experiments. In addition to the AI-as-first-reader strategy shown in Figure 1d, we investigate an *arbitration* scenario where an AI and a non-expert doctor concurrently assess a patient (Figure 5 in Supplementary). In case of disagreements or AI abstaining, an expert 2$^{\text{nd}}$ human reader makes the final decision. For this setting, we consider InD, OoD-H, and OoD-D, along with the various AI models and certainty measures described above.

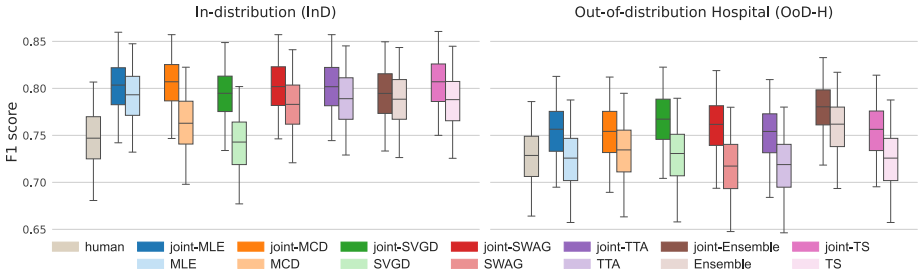

Fig. 4. Performance of humans, AI-standalone, and AI-first-reader joint system at optimal coverage τ. Boxplots indicate 95% confidence intervals through bootstrapping.

5 Results and Discussion

In Figure 3 we present the main results, comparing joint human-AI systems using different certainty methods against standalone human performance (subdivided by human expertise in Supplementary Figure 6). These experiments cover different reading strategies over situations where the system is deployed in environments identical to the training context (InD), and in novel contexts, including out-of-domain deployments characterized by different hospitals (OoD-H) and imaging devices (OoD-D). We find that all joint systems surpass both standalone doctors and standalone models. For some settings, we observe strong performance and efficiency gains from the AI-as-first-reader strategy. For example, for images from unseen devices, joint systems achieve a boost in F1 score of $\approx 4\%$ over both AI and doctors, while demanding 75% less of doctors' time. Looking at arbitration as a reading strategy, we observe the same general trends as for AI-as-first-reader, noting that the stability of joint system performance across different model types and model coverage is improved by having more doctors involved in the process and by having experienced doctors make the final decision. In Figure 4, we present the peak performance of the AI-first-reader joint system at the ideal model coverage τ. This perspective highlights the improvements achieved by the joint system and AI in isolation over traditional human-only approaches in the in-domain (InD) scenario. However, for out-of-domain (OoD) deployments, the AI-standalone's advantage disappears. Looking back at Figure 3, we observe that, for OoD data, the joint system relies on more doctor assessments to perform well, indicating the ability of the AI to defer to human judgment when encountering unfamiliar data patterns. Finally, in Table 1, we provide our summary metrics from Eq. 4 alongside established confidence and uncertainty metrics for comparison.

Which certainty estimate is the best? While our study did not identify a single certainty estimate as being definitively superior, we observed that integrating any of the evaluated models within a collaborative human-AI framework consistently enhanced diagnostic accuracy beyond what could be achieved

[4] We enforce diversity by minimizing cosine similarity between the ensemble weights.

Table 1. Established measures of confidence and uncertainty along with pAUJF1C_γ. A complete list of metrics on InD, OoD-H and OoD-D is given in the Supplementary.

Method	F1 ↑		NLL ↓		ECE ↓		AURC ↓		AUC_{mis} ↑		$\text{pAUJF1C}_{0.5}$ ↑		$\text{pAUJF1C}_{0.9}$ ↑	
	InD	OoD-H	InD	OoD-H	InD	OoD-H	InD	OoD-H	InD	OoD-H	InD	OoD-H	InD	OoD-H
MLE	**0.793**	0.723	0.410	0.452	0.047	0.039	0.079	0.092	0.727	0.760	**0.394**	0.368	**0.075**	0.069
TTA	0.789	0.717	0.422	0.489	0.065	0.073	0.076	0.100	0.732	0.746	0.393	0.364	0.074	0.068
TS	0.787	0.723	0.409	0.451	0.053	0.041	0.077	0.092	0.740	0.760	**0.394**	0.368	**0.075**	0.068
MCD	0.759	0.732	0.411	0.470	0.043	0.070	0.080	0.102	0.771	0.719	0.393	0.366	0.073	0.069
SVGD	0.741	0.729	0.417	0.442	0.051	**0.032**	0.085	0.087	**0.798**	**0.781**	0.388	0.374	0.071	0.069
SWAG	0.783	0.717	**0.388**	0.466	0.060	0.057	**0.067**	0.090	0.773	**0.781**	0.390	0.369	0.074	0.068
Ensemble	0.787	**0.760**	0.392	**0.428**	**0.027**	0.037	0.068	**0.079**	0.747	0.752	0.390	**0.380**	0.073	**0.071**

by either humans or AI systems operating independently. The performance of these methods varies significantly depending on the level of coverage, τ, chosen. For example, when using AI-as-first-reader for InD cases, the joint F1 score for the SVGD method peaks at a coverage level of $\tau = 0.6$. In contrast, the TS method reaches its highest performance with considerably less reliance on human input, at a coverage level of $\tau = 0.95$. Our summary metric pAUJF1C_γ ($\gamma \in \{0.5, 0.75, 0.9\}$) shows very little difference between certainty measures on InD data. For OoD-H data, however, Deep Ensembles seem to offer a slight advantage on average. Remarkably, MLE, the simplest approach to certainty estimation, is competitive with more sophisticated methods – particularly for InD data. This suggests that even straightforward certainty estimation methods can be valuable in decision-making processes of joint human-AI system.

Is there a need for a new certainty metric to evaluate joint human-AI systems? We argue that *yes*, there is, for two main reasons. *(1)* Metrics that assess the quality of the estimate, such as NLL or ECE, fail to capture the practical utility of certainty estimates – a consideration of particular relevance for clinical applications. Metrics that do consider utility, such as AURC and AUC_{mis}, overlook the implications of model abstention on system performance. As seen in Table 1, these shortcomings can lead to misleading conclusions. For example, TS [2] would seem like a poor certainty estimation method for InD based on the F1, NLL, ECE, AURC, or AUC_{mis} metrics. But in reality, we find TS outperforms the other methods when used to support doctors, as shown in Figure 4-InD. *(2)* Existing metrics do not consider the efficiency of healthcare resource utilization. By inspecting the joint coverage curves, hospital administrators can tailor AI model usage to align with available resources, optimizing both performance and resource allocation. For these reasons, we argue that joint coverage curves should be the primary analysis tool when evaluating joint human-AI systems. When benchmarking certainty estimates, utility in joint human-AI systems, as measured by pAUJF1C_γ, should be reported alongside other established measures.

Other considerations. Our findings reveal a 'sweet spot' of human involvement in the joint human-AI system, beyond which additional human oversight harms performance. When applied to new settings, such as unfamiliar hospitals or imaging devices, the system's optimal balance shifts to require more human oversight. This highlights the need for good certainty estimates to allow the AI

to defer when it is uncertain. In principle, our framework can be extended to various reading strategies, including unblinded double reading. However, a proper assessment of many strategies would require either prospective data involving human decisions or a reliable simulation, which presents its own set of challenges. Finally, we note that determining an optimal threshold (α) for model abstention based on validation set data may not effectively translate to scenarios involving out-of-distribution (OoD) samples, highlighting the need to collect data to calibrate the model to its new setting.

Acknowledgments. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Data providers : A. Czekierdowski, F. P. G. Leone, L. A. Haak, R. Fruscio, A. Gaurilcikas, D. Franchi, D. Fischerova, E. Mor, L. Savelli, M. À. Pascual, M. Kudla, S. Guerriero, F. Buonomo, K. Liuba, N. Montik, J. L. Alcázar, E. Domali, N. C. P. Pangilinan, C. Carella, M. Munaretto, P. Šašková, D. Verri, C. Visenzi .

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

References

1. Alves, N., Bosma, J.S., Venkadesh, K.V., Jacobs, C., Saghir, Z., de Rooij, M., Hermans, J., Huisman, H.: Prediction variability to identify reduced ai performance in cancer diagnosis at mri and ct. Radiology **308**(3), e230275 (2023)
2. Ayhan, M.S., Berens, P.: Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In: Medical Imaging with Deep Learning (2022)
3. Brier, G.W.: Verification of forecasts expressed in terms of probability. Monthly weather review **78**(1), 1–3 (1950)
4. Bungert, T.J., Kobelke, L., Jaeger, P.F.: Understanding silent failures in medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 400–410. Springer (2023)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Ding, Y., Liu, J., Xiong, J., Shi, Y.: Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 4–5 (2020)
7. 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)
8. Galil, I., Dabbah, M., El-Yaniv, R.: What can we learn from the selective prediction and uncertainty estimation performance of 523 imagenet classifiers. arXiv preprint arXiv:2302.11874 (2023)
9. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networksdvi. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)

10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. Jaeger, P.F., Lüth, C.T., Klein, L., Bungert, T.J.: A call to reflect on evaluation practices for failure detection in image classification. arXiv preprint arXiv:2211.15259 (2022)
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. Liu, Q., Wang, D.: Stein variational gradient descent: A general purpose bayesian inference algorithm. Advances in neural information processing systems **29** (2016)
14. 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)
15. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Advances in neural information processing systems **32** (2019)
16. Rezazade Mehrizi, M.H., Mol, F., Peter, M., Ranschaert, E., Dos Santos, D.P., Shahidi, R., Fatehi, M., Dratsch, T.: The impact of ai suggestions on radiologists' decisions: a pilot study of explainability and attitudinal priming interventions in mammography examination. Scientific Reports **13**(1), 9230 (2023)
17. Seligmann, F., Becker, P., Volpp, M., Neumann, G.: Beyond deep ensembles: A large-scale evaluation of bayesian deep learning under distribution shift. Advances in Neural Information Processing Systems **36** (2024)