A Systematic Comparison of Bayesian Deep Learning Robustness in Diabetic Retinopathy Tasks

Angelos Filos, Sebastian Farquhar, Aidan N. Gomez, Tim G. J. Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud de Kroon, Yarin Gal
University of Oxford
{angelos.filos, sebastian.farquhar, yarin}@cs.ox.ac.uk

Abstract

Evaluation of Bayesian deep learning (BDL) methods is challenging. We often seek to evaluate the methods' robustness and scalability, assessing whether new tools give 'better' uncertainty estimates than old ones. These evaluations are paramount for practitioners when choosing BDL tools on-top of which they build their applications. Current popular evaluations of BDL methods, such as the UCI experiments, are lacking: Methods that excel with these experiments often fail when used in application such as medical or automotive, suggesting a pertinent need for new benchmarks in the field. We propose a new BDL benchmark with a diverse set of tasks, inspired by a real-world medical imaging application on *diabetic retinopathy diagnosis*. Visual inputs (512×512 RGB images of retinas) are considered, where model uncertainty is used for medical pre-screening—i.e. to refer patients to an expert when model diagnosis is uncertain. Methods are then ranked according to metrics derived from expert-domain to reflect real-world use of model uncertainty in automated diagnosis. We develop multiple tasks that fall under this application, including out-of-distribution detection and robustness to distribution shift. We then perform a systematic comparison of well-tuned BDL techniques on the various tasks. From our comparison we conclude that some current techniques which solve benchmarks such as UCI 'overfit' their uncertainty to the dataset—when evaluated on our benchmark these underperform in comparison to simpler baselines. The code for the benchmark, its baselines, and a simple API for evaluating new BDL tools are made available at `https://github.com/oatml/bdl-benchmarks`.

1 Introduction

Deep learning is continuously transforming intelligent technologies across many fields, from advancing medical diagnostics with complex data, to enabling autonomous driving, to deciding high-stakes economic actions [29]. However, deep learning models struggle to inform their users *when they don't know* – in other words, these models fail to communicate their uncertainty in their predictions. The implications for deep models entrusted with life-or-death decisions are far-reaching: experts in medical domains cannot know whether to trust their auto-diagnostics system, and passengers in self-driving vehicles cannot be alerted to take control when the car does not know how to proceed.

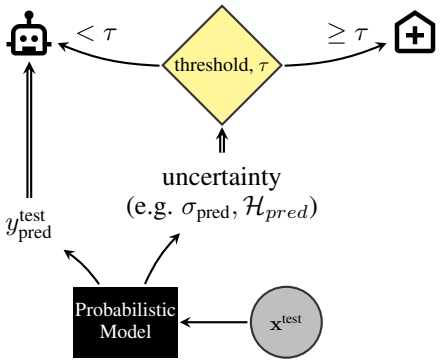


Figure 1: Automated diagnosis: a model provides a classification and an uncertainty estimate. Predictions above an uncertainty threshold are referred to a medical expert, otherwise handled by the model.

4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.

Bayesian deep learning (BDL) offers a pragmatic approach to combining Bayesian probability theory with modern deep learning. BDL is concerned with the development of techniques and tools for quantifying when deep models become uncertain, a process known as *inference* in probabilistic modelling. BDL has already been demonstrated to play a crucial role in applications such as medical diagnostics [30, 16, 5, 49] (see Figure 1), computer vision [21, 18, 17], in the sciences [31, 34], and autonomous driving [1, 15, 21, 18, 19].

Despite BDL's impact on a range of real-world applications and the flourish of recent ideas and inference techniques [11, 3, 8, 48, 50, 37], the development of the field itself is impeded by the lack of realistic benchmarks to guide research. Evaluating new inference techniques on real-world applications often requires expert domain knowledge, and current benchmarks used for the development of new inference tools lack consideration for the cost of development, or for scalability to real-world applications.

Advances in computer vision, natural language and reinforcement learning are usually attributed to the emergence of challenging benchmarks, e.g. ImageNet [6], GLUE [46] and ALE [2], respectively. In contrast, many BDL papers use benchmarks such as the toy UCI datasets [12], which consist of only evaluating root mean square error (RMSE) and negative log-likelihood (NLL) on simple datasets with only a few hundred or thousand data points, each with low input and output dimensionality. Such evaluations are akin to toy MNIST [28] evaluations in deep learning. Due to the lack of alternative standard benchmarks, in current BDL research it is common for researchers developing new inference techniques to evaluate their methods with such toy benchmarks alone, ignoring the demands and constraints of the real-world applications which make use of BDL tools [35]. This means that research in BDL broadly neglects exactly the applications that neural networks have proven themselves most effective for.

In order to make significant progress in the deployment of new BDL inference tools, the tools must scale to real-world settings. And for that, researchers must be able to evaluate their inference and iterate quickly with real-world benchmark tasks without necessarily worrying about the required application-specific expertise. We require benchmarks which test for inference robustness, performance, and accuracy, in addition to cost and effort of development. These benchmarks should include a variety of tasks, assessing different properties of uncertainty while avoiding the pitfalls of overfitting quickly as with UCI. These should asses for scalability to large data and be truthful to real-world applications, capturing their constraints.

Contributions. We build on-top of previous work published at *Nature Scientific Reports* by Leibig et al. [30]. We extend on their methodology and develop an open-source benchmark, building on a downstream task which makes use of BDL in a real-world application—detecting diabetic retinopathy from fundus photos and referring the most uncertain cases for further inspection by an expert (Section 2). We extend this methodology with additional tasks that assess robustness to out-of-distribution and distribution shift, using test datasets which were collected using different medical equipment and for different patient populations. Our implementation is easy to use for machine learning researchers who might lack specific domain expertise, since expert details are abstracted away and integrated into metrics which are exposed through a simple API. Improvement on this benchmark will directly be contributing to the advancement of an important real-world application. We further perform a comprehensive comparison on this new benchmark, contrasting many existing BDL techniques. We develop and tune baselines for the benchmark, including Monte Carlo dropout [8], mean-field variational inference [39, 11, 3] and model ensembling [26], as well as variants of these (Section 3). We conclude by demonstrating the benchmark's usefulness in ranking existing techniques in terms of scalability and effectiveness, and show that despite the fact that some current techniques solve benchmarks such as UCI, they either fail to scale, fail to solve our benchmark, or fail to provide good uncertainty estimates. This shows that an over-reliance on UCI has the potential to badly distort work in the field because researchers prioritize their attention on approaches to Bayesian deep learning that are not suited to large scale applications (Section 5).

It is our hope that the proposed benchmarks will make testing new inference techniques for Bayesian deep learning radically easier, leading to faster development iteration cycles, and rapid development of new tools. Progress on these benchmarks will translate to more *robust* and *reliable* tools for already-deployed decision-making systems, such as automatic medical diagnostics and self-driving car prototypes.

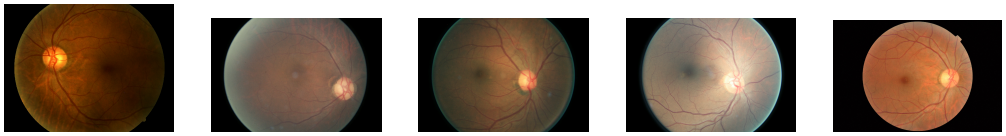
2 Diabetic Retinopathy Benchmark

We describe the dataset, the data processing, as well as the downstream task and metrics used.

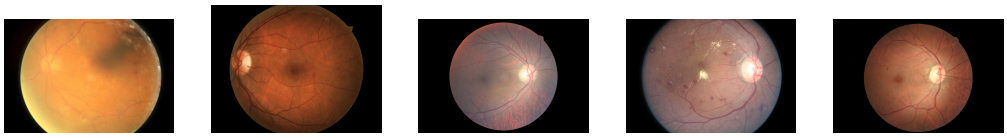
2.1 Dataset

The benchmark is built on the Kaggle Diabetic Retinopathy (DR) Detection Challenge [14] data. It consists of 35,126 training images and 53,576 test images. We hold-out 20% of the training data as a validation set. Each image is graded by a specialist on the following scale: 0 – No DR, 1 – Mild DR, 2 – Moderate DR, 3 – Severe DR and 4 – Proliferative DR. We recast the 5-class classification task as binary classification which is easily applicable to any BDL classification algorithm by asking the user to classify whether each image has sight-threatening DR, which is defined as Moderate DR or greater (classes 2-4) following [30]. Samples from both classes are provided in Figure 2. The data are unbalanced, with only 19.6% of the training set and 19.2% of the test set having a positive label.

Robustness to distribution shift is evaluated by training on the original Kaggle diabetic retinopathy detection challenge dataset [14], and testing on a completely disjoint APTOS 2019 Blindness Detection dataset collected in India with different medical equipment and on a different population.



(a) Healthy samples, $y = 0$.



(b) Unhealthy samples, $y = 1$.

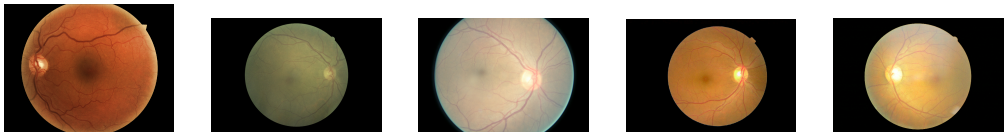
Figure 2: Samples from the two classes, healthy and unhealthy from the raw dataset.

2.2 Data Processing

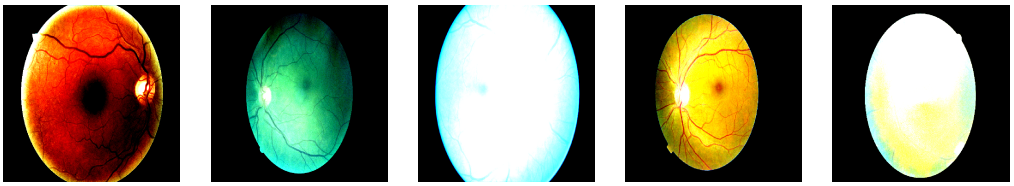
All images are cropped and resized to 512×512, while all three colour channels are used. The data is standard normalized for each colour channel separately, using the empirical statistics of the training data. Similar to Leibig et al. [30], we augment training dataset using affine transformations, including random zooming (by up to $\pm 10\%$), random translations (independent shifts by up to ± 25 pixels) and random rotations (by up to $\pm\pi$). Finally half of the augmented data is also flipped along the vertical and/or the horizontal axis. Examples of original and their corresponding processed images are provided in Figure 3.

2.3 Downstream Task

Machine learning researchers often evaluate their predictions directly on the whole test set. But, in fact, in real-world settings we have additional choices available, like asking for more information when we are uncertain. Because of the importance of accurate diagnosis, it would be unreasonable *not* to ask for further scans of the most ambiguous cases. Moreover, in this dataset, many images feature camera artefacts that distort results. In these cases, it is critically important that a model is able to tell when the information provided to it is not sufficiently reliable to classify the patient. Just like real medical professionals, any diagnostic algorithm should be able to flag cases that require more investigation by medical experts. This task is illustrated in Figure 1, where a threshold, τ, is used to flag cases as certain and uncertain, with uncertain cases referred to an expert. Alternatively, the uncertainty estimates could be used to come up with a priority list, which could be matched to the available resources of a hospital, rather than waste diagnostic resources on patients for whom the diagnosis is clear cut.

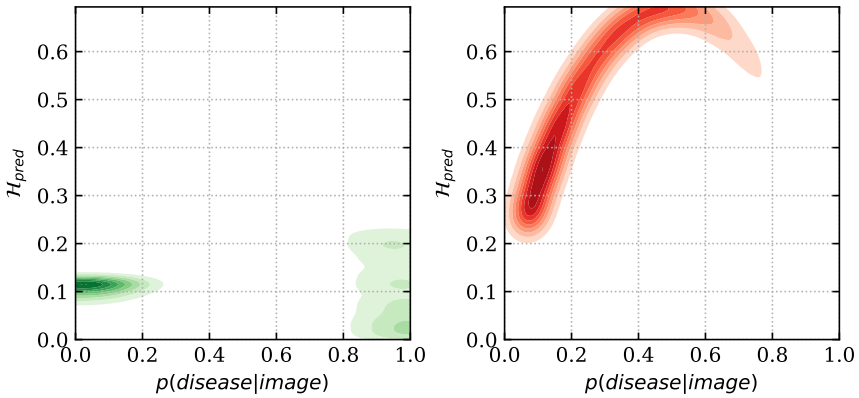


(a) Original samples from the diabetic retinopathy dataset.



(b) Processed samples from the diabetic retinopathy dataset.

Figure 3: Illustrative examples of the pre-processing procedure applied to the original dataset.



(a) Correctly classified test images

(b) Erroneously classified test images

Figure 4: Relation between predictive uncertainty (i.e. entropy), $\mathcal{H}_{\text{pred}}$, of MC Dropout model, and maximum-likelihood, i.e. sigmoid probabilities $p(\text{disease}|\text{image})$. The model has higher uncertainty for the miss-classified images, hence it can be used as an indicator to drive referral.

To get some insight into the dataset, Figure 4 illustrates the relation between predicted probabilities, $p(\text{disease}|\text{image})$, and our estimator for the models' uncertainty about them, the predictive entropy $\mathcal{H}_{\text{pred}}$, for an MC dropout model. Note that the model is correct and certain about most of its predictions, as shown in sub-figure (a), while it is more uncertain when wrong, sub-figure (b).

2.4 Metrics

In order to simulate this process of referring the uncertain cases to experts and relying on the model's predictions for cases it is certain of, we assess the techniques by their diagnostic accuracy and area under receiver-operating-characteristic (ROC) curve, as a function of the referral rate. We expect the models with well-calibrated uncertainty to refer their least confident predictions to experts (see Figure 5), improving their performance as the number of referrals increases.

The accuracy of the binary classifier is defined as the ratio of correctly classified data-points over the size of the population. The receiver-operating-characteristic (ROC) curve (see Figure 6) illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It is created by plotting the true positive rate (a.k.a. sensitivity) against the false positive rate (a.k.a. $1 -$ sensitivity). The quality of such a ROC curve can be summarized by its area under the curve (AUC), which varies between 0.5 (chance level) and 1.0 (best possible value).

3 A Systematic Comparison of BDL Methods

We next present and evaluate various Bayesian deep learning techniques (i.e. baselines) on the diabetic retinopathy diagnosis benchmark. Each method is tuned separately and, in order to obtain statistically significant results, we train nine independent models for each method, using different random number generator seeds. We observe consistency and robustness for our implementations across seeds.

Architecture. Our models are deep convolutional neural networks [27], variants of the well-established VGG architecture [41] (around 2.5 million parameters). The ADAM [22] adaptive optimizer with initial learning rate $4\text{e}{-4}$ and batch size 64 are used for training all models. Leaky rectified linear units (Leaky ReLUs) [51] with $\alpha = 0.2$ are used for the hidden layers, and a sigmoid for the output layer, modelling the probability of a patient having diabetic retinopathy given an image of his retina, $p(\text{disease}|\text{image})$. In contrast to [30] who uses a pre-trained network, we initialize the weights randomly, according to Glorot and Bengio [10] uniform initialization method.

Class imbalance. We compensate for the class imbalance discussed in Section 2.1 by reweighing the cross-entropy part of the cost function, attributing more weight to the minority class, given by the relative class frequencies in each mini-batch, $p(k)_{\text{mini-batch}}$ [30]:

$$\mathcal{L} = -\frac{1}{Kn}\sum_{i=1}^{n}\frac{\mathcal{L}_{\text{cross-entropy}}}{p(k)_{\text{mini-batch}}}. \tag{1}$$

We also tried using a constant class weight, or artificially balancing the two classes by sub-sampling negatively labelled images, but both approaches made training slower and less stable for many baselines.

Uncertainty Estimator. We quantify the uncertainty of our binary classification predictions by predictive entropy [40, 7], which captures the average amount of information contained in the predictive distribution[1]:

$$\mathcal{H}_{\text{pred}}(y|\mathbf{x}) := -\sum_{c} p(y=c|\mathbf{x})\log p(y=c|\mathbf{x}) \tag{2}$$

summing over all possible classes c that y can take, in our case $c \in \{0, 1\}$. This quantity is high when *either* the aleatoric uncertainty is high (the input is ambiguous), *or* when the epistemic uncertainty is high (a probabilistic model has many possible explanations for the input). In practice, we approximate the $p(y = c|\mathbf{x})$ term in (2) by T Monte Carlo samples, $\frac{1}{T}\sum_t p_\theta(y = c|\mathbf{x})$, obtained by stochastic forward passes through the probabilistic networks. Note that this is a biased but consistent estimator of the predictive entropy in (2) [7].

We contrast several methods in BDL which we discuss in more detail next.

3.1 Bayesian Neural Networks

Estimating the uncertainty about a machine learning based prediction on a single observation requires a distribution over possible outcomes, for which a Bayesian perspective is principled. Bayesian approaches to uncertainty estimation have indeed been proposed to assess the reliability of clinical predictions [24, 25, 47, 30] but have only been applied to a handful of large-scale real-world problems [30, 15, 42] that neural networks (NNs) have proven themselves particularly effective for.

Finite NNs with distributions placed over the weights have been studied extensively as Bayesian neural networks (BNNs) [36, 33, 8], providing robustness to over-fitting (i.e. regularization). Exact inference is analytically intractable and hence approximate inference has been applied instead [13, 39, 11, 8].

[1] The predictive uncertainty is the sum of epistemic and aleatoric uncertainty.

Given a dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$, a BNN is defined in terms of a prior $p(\mathbf{w})$ on the weights, as well as the likelihood $p(\mathcal{D}|\mathbf{w})$. Variational Bayesian methods attempt to fit an approximate posterior $q(\mathbf{w})$ to maximize the evidence lower bound (ELBO):

$$\mathcal{L}_q = \mathbb{E}_q[\log p(\mathcal{D}|\mathbf{w})] - \mathrm{KL}[q(\mathbf{w})\|p(\mathbf{w})] \tag{3}$$

We parameterize $q(\mathbf{w})$ with θ parameters and choose prior distribution $p(\mathbf{w})$. The (variational) inference is then recast as the optimization problem $\max_\theta \mathcal{L}_{q_\theta}$. Different methods use different prior distributions and parametric families for the approximate posterior, as well as optimization methods. We discuss these different techniques next.

Mean-field Variational Inference. Mean-field variational inference (MFVI) is an approach to learning an approximate posterior over the weights of a neural network, $q_\theta(\mathbf{w})$, given a prior $p(\mathbf{w})$ [39, 11, 3]. In MFVI, we assume a fully-factorized Gaussian posterior (and prior). This reduces the computational complexity of estimating the evidence lower-bound (ELBO). In addition, we use a Monte Carlo estimate of the KL-divergence term of the ELBO in order to reduce the time complexity of a forward pass to $\mathcal{O}(D)$ in the number of weights. Blundell et al. [3] applied the reparametrization trick from [23] to perform MFVI, which they call Bayes-by-backprop. Instead, we use the Flipout Monte Carlo estimator of the KL-divergence [48], which reduces the variance of the estimator of the gradient. A Monte Carlo estimate of model predictions is made by taking a number of samples from the posterior distribution over the weights and averaging the predictions.

Note that the effective number of trainable parameters is doubled compared to a deterministic NN, since both the mean and scale parameters are now learnable. To allow fair comparison with the other baselines, we reduce the number of channels in the convolutional layers of the MFVI model to reach the model budget of 2.5 million parameters.

Monte Carlo Dropout. Gal and Ghahramani [8] showed that optimising *any* neural network with the standard regularization technique of dropout [43] and L2-regularization is equivalent to a form of variational inference in a probabilistic interpretation of the model, so long as the dropout probability/L2 regularization are appropriately optimized [7]. Monte Carlo samples can be drawn from the dropout NNs by using dropout at *test time*, hence the name of the method Monte Carlo Dropout (MC Dropout). In our implementation, we chose a dropout rate to 0.2 and perform a grid-search to set the L2-regularization coefficient. $5\mathrm{e}{-5}$ was found to be the best value. Better calibration of uncertainties can be obtained by optimizing the dropout rate using convex relaxation methods as in [9], but we leave this as future work.

3.2 Model Ensembling

Lakshminarayanan et al. [26] proposed an alternative to BNNs, termed Deep Ensembles, that is simple to implement, readily parallelizable, requires little hyperparameter tuning, and yields high quality predictive uncertainty estimates. The method quantifies uncertainty by collecting predictions from T independently trained deterministic models (ensemble components). Despite the easy parallelization of the method, the resources for training scale linearly with the required number of ensemble components T, making it prohibitively expensive in some cases.

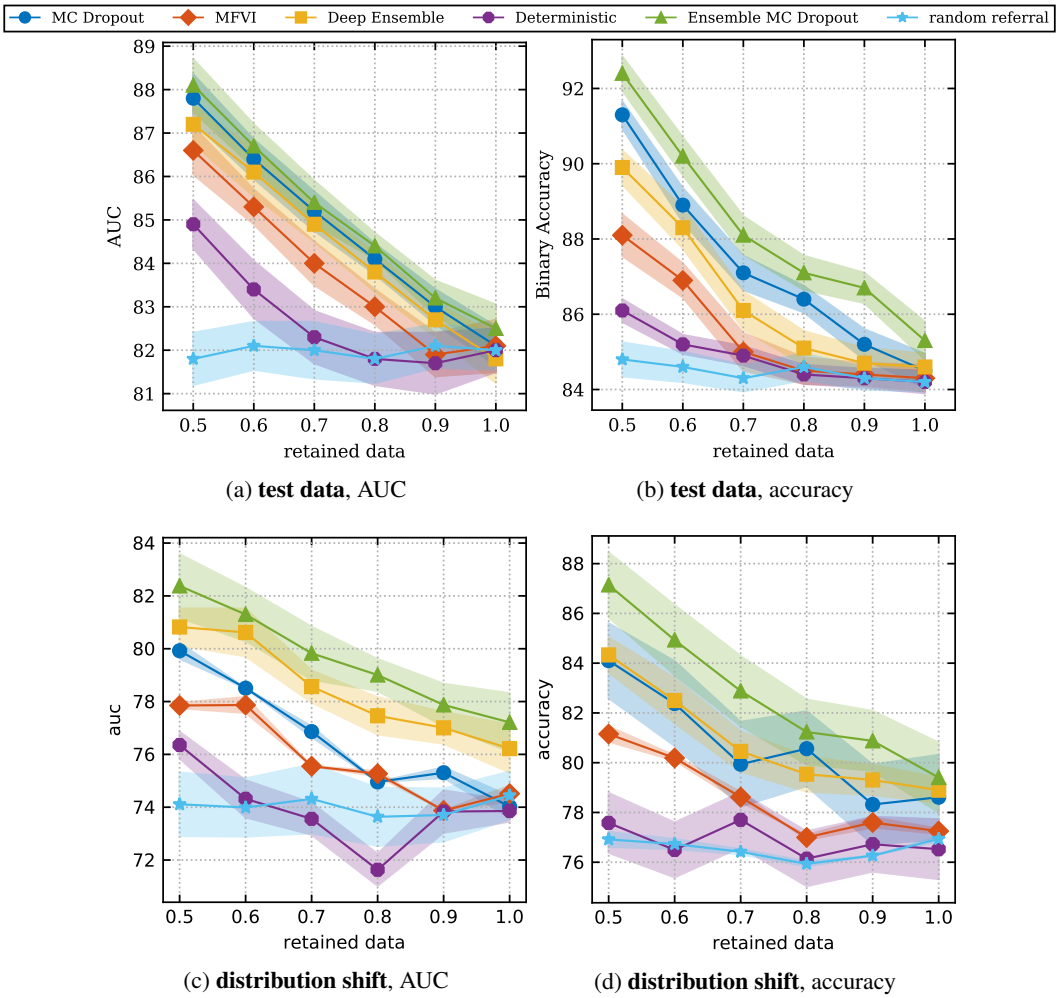
We also report results on an ensemble of MC Dropout models, which performs best of all the other methods, in terms of both accuracy and AUC for all the referral rates, as illustrated in Figure 5 and Table 1. In this technique, several dropout models are separately trained in parallel. Predictions are then made by sampling repeatedly from all models in the ensemble using fresh dropout masks for each sample, before being averaged, to form a Monte Carlo estimate of the ensemble prediction.

3.3 Deterministic

Two naive baselines are evaluated as control, a Deterministic neural network and Random. Both are based on a deep VGG model, trained with dropout and L2-regularization, using exactly the same hyperparameters and set-up as MC Dropout. In fact, because the conditions are identical, we used the same models for the Deterministic and MC Dropout baselines—the only difference is that for MC Dropout we sample dropout mask during evaluation and average over 100 samples from the dropout posterior to estimate uncertainty. In contrast, the Deterministic baseline uses the sigmoid

output $p(\text{disease}|\text{image})$ to quantify uncertainty, and uses the deterministic dropout approximation at test time [43]. That is, a model is assumed to be more confident the closer to 1 or 0 its output is. This is the simplest way a neural network might estimate uncertainty, but it captures only the aleatoric component of uncertainty—it does not capture epistemic uncertainty about the model's knowledge [20]. Figure 4 (right) shows that there is a correlation between the sigmoid output $p(\text{disease}|\text{image})$ and the predictive entropy $\mathcal{H}_{\text{pred}}$, which we use to measure uncertainty. But the overall evidence in Figure 5 and Table 1 suggests that models which also capture the epistemic component of the uncertainty perform much better than the Deterministic baseline.

The Random baseline makes random referrals, without taking any kind of uncertainty (or input) into account. As expected, it has the same accuracy and AUC regardless of how much data is retained vs. referred.



(a) **test data**, AUC

(b) **test data**, accuracy

(c) **distribution shift**, AUC

(d) **distribution shift**, accuracy

Figure 5: Area under the receiver-operating characteristic curve (AUC) and binary accuracy for the different baselines for in-distribution (a) and (b), and out-of-distribution (c) and (d) evaluation. The methods that capture uncertainty score better when less data is retained, referring the least certain patients to expert doctors. The best scoring methods, *MC Dropout*, *mean-field variational inference* and *Deep Ensembles*, estimate and use the predictive uncertainty. The deterministic deep model regularized by *standard dropout* uses only aleatoric uncertainty and performs worse. Shading shows the standard error. The *Ensemble of MC Dropout* method performs consistently better, even under the distribution shift to the APTOS 2019 dataset. However, mean-field variational inference's and MC Dropout's performance degrades in this out-of-distribution.

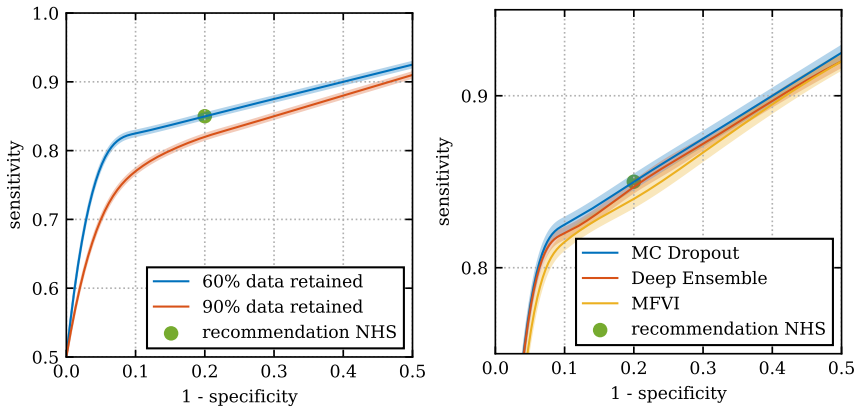


Figure 6: Receiver-operating characteristic curve (ROC) on the diabetic retinopathy diagnosis benchmark and the NHS recommended 85% sensitivity and 80% specifcity ratio: (left) the performance of MC Dropout baseline for 60% and 90% data retained based on predictive entropy; (right) the comparison of the baselines at 60% data retained rate. Note the different y-axes.

	50% data retained		70% data retained		100% data retained	
Method	**AUC** ↑	**Accuracy** ↑	**AUC** ↑	**Accuracy** ↑	**AUC** ↑	**Accuracy** ↑
Kaggle Dataset (out-of-sample)						
MC Dropout	$\mathbf{87.8 \pm 1.1}$	$\mathbf{91.3 \pm 0.7}$	$\mathbf{85.2 \pm 0.9}$	87.1 ± 0.9	82.1 ± 0.9	84.5 ± 0.9
Mean-field VI	86.6 ± 1.1	88.1 ± 1.1	84.0 ± 1.0	85.0 ± 1.0	82.1 ± 1.2	84.3 ± 0.7
Deep Ensembles	87.2 ± 0.9	89.9 ± 0.9	84.9 ± 0.8	86.1 ± 1.0	81.8 ± 1.1	84.6 ± 0.7
Deterministic	84.9 ± 1.1	86.1 ± 0.6	82.3 ± 1.2	84.9 ± 0.5	82.0 ± 1.0	84.2 ± 0.6
Ensemble MC Dropout	$\mathbf{88.1 \pm 1.2}$	$\mathbf{92.4 \pm 0.9}$	$\mathbf{85.4 \pm 1.0}$	$\mathbf{88.1 \pm 1.0}$	82.5 ± 1.1	85.3 ± 1.0
Random	81.8 ± 1.2	84.8 ± 0.9	82.0 ± 1.3	84.3 ± 0.7	82.0 ± 0.9	84.2 ± 0.5
APTOS 2019 Dataset (distribution shift)						
MC Dropout	79.9 ± 0.3	84.1 ± 1.5	76.8 ± 0.2	79.9 ± 1.7	74.0 ± 0.4	78.6 ± 1.7
Mean-field VI	77.8 ± 0.1	81.1 ± 0.3	75.5 ± 0.0	78.6 ± 0.2	74.5 ± 0.3	77.2 ± 0.1
Deep Ensembles	80.8 ± 0.7	84.3 ± 0.7	78.5 ± 0.6	80.4 ± 0.8	76.2 ± 0.9	78.8 ± 0.5
Deterministic	76.3 ± 0.5	77.5 ± 1.2	73.5 ± 0.6	77.7 ± 1.1	73.8 ± 0.4	76.5 ± 1.2
Ensemble MC Dropout	$\mathbf{82.3} \pm 1.2$	$\mathbf{87.1} \pm 1.3$	$\mathbf{79.8 \pm 1.0}$	$\mathbf{82.8 \pm 1.4}$	$\mathbf{77.2 \pm 1.1}$	$\mathbf{79.4 \pm 1.4}$
Random	74.1 ± 1.2	76.9 ± 1.3	74.3 ± 1.3	76.4 ± 1.0	74.4 ± 0.9	76.9 ± 0.0

Table 1: Summary performances of baselines in terms of area under the receiver-operating-characteristic curve (AUC) and classification accuracy as a function of retained data. In the case of no referral (100% data retained), all methods score equally, within standard error bounds. For lower referral rates 'Ensemble MC Dropout' performs best (with MC dropout matching performance in the extreme case of 50% referral rate).

4 Results and Analysis

Table 1 and Figures 5 and 6 summarize the quantitative performance of various methods, described in Section 3. Methods that capture meaningful uncertainty estimates show this by improving performance (i.e. AUC and accuracy) as the rate of referral increases. That is, steeper slopes in Figure 5 are making better estimates of uncertainty, all else equal, because they are able to systematically refer the datapoints where their estimates are less likely to be accurate. Note that all methods perform equally well when all data is retained, conveying that all models have converged to similar overall performance, providing a *fair comparison of* ***uncertainty***.

Benchmarks are often used to compare methodology, e.g. to select which tools we should build on-top. UCI, a popular benchmark in the field, has been used to reproduce such rankings of BDL methods. Importantly, in contrast to the empirical results found in [4] on the toy UCI benchmark and summarised in Table 2, our benchmark suggest a different ranking of methods. While in [4] mean-field variational inference outperforms the other baselines we discuss, Table 1 and Figures 5 and 6 suggest that in the real-world application of diabetic retinopathy diagnosis both ensemble methods (Section 3.2) and Monte Carlo Dropout score consistently higher than MFVI, suggesting that some methods might be 'overfitting' their uncertainty to the simple dataset. That is, extensive

	Log-Likelihood ↑			Root Mean Squared Error ↓		
Datasets	**MC Dropout**	**Mean-field VI**	**Deep Ensembles**	**MC Dropout**	**Mean-field VI**	**Deep Ensembles**
Boston housing	-2.46 ± 0.25	-2.58 ± 0.06	$\mathbf{-2.41 \pm 0.25}$	$\mathbf{2.97 \pm 0.85}$	3.42 ± 0.23	3.28 ± 1.00
Concrete	$\mathbf{-3.04 \pm 0.09}$	-5.08 ± 0.01	$\mathbf{-3.06 \pm 0.18}$	$\mathbf{5.23 \pm 0.53}$	5.71 ± 0.15	6.03 ± 0.58
Energy	-1.99 ± 0.09	$\mathbf{-1.05 \pm 0.01}$	-1.38 ± 0.22	1.66 ± 0.19	$\mathbf{0.81 \pm 0.08}$	2.09 ± 0.29
Kin8nm	$+0.95 \pm 0.03$	$+1.08 \pm 0.01$	$\mathbf{+1.20 \pm 0.02}$	0.10 ± 0.00	0.37 ± 0.00	$\mathbf{0.09 \pm 0.00}$
Naval propulsion plant	$+3.80 \pm 0.05$	-1.57 ± 0.01	$\mathbf{+5.63 \pm 0.05}$	0.01 ± 0.00	0.01 ± 0.00	$\mathbf{0.00 \pm 0.00}$
Power plant	$\mathbf{-2.80 \pm 0.05}$	-7.54 ± 0.00	$\mathbf{-2.79 \pm 0.04}$	$\mathbf{4.02 \pm 0.18}$	$\mathbf{4.02 \pm 0.04}$	4.11 ± 0.17
Protein	-2.89 ± 0.01	-3.67 ± 0.00	$\mathbf{-2.83 \pm 0.02}$	$\mathbf{4.36 \pm 0.04}$	4.40 ± 0.02	4.71 ± 0.06
Wine	$\mathbf{-0.93 \pm 0.06}$	-3.15 ± 0.01	$\mathbf{-0.94 \pm 0.12}$	0.62 ± 0.04	0.65 ± 0.01	0.64 ± 0.04
Yacht	-1.55 ± 0.12	-4.20 ± 0.05	$\mathbf{-1.18 \pm 0.21}$	$\mathbf{1.11 \pm 0.38}$	1.75 ± 0.42	1.58 ± 0.48

Table 2: Results for Deep Ensembles are borrowed from the original paper by Lakshminarayanan et al. [26] and for MC Dropout and Mean-field VI from [35].

tuning on the simple UCI tasks might have resulted in rankings which do not generalise to other tasks. Moreover, Mukhoti et al. [35] show that UCI regression benchmarks are insufficient for drawing conclusions about the effectiveness, and surely the scalability, of the inference techniques.

5 Implications for the Field

Deep learning, as a whole, has had its biggest successes when handling large, high-dimensional data. It is something of a surprise, then, that the standard benchmarks for Bayesian deep learning, UCI, only has input dimensionalities between 4 and 16. Due to the lack of alternative common benchmarks with well tuned baselines, researchers find it hard to publish results in Bayesian deep learning without resorting to a comparison on UCI. As a result, there is an undue focus in Bayesian deep learning on models that perform well with very low numbers of input features and on tiny models with a single layer of only 50 hidden units. UCI plays an important role for a subset of models, but the fact that it is currently the field's main benchmark has a distorting effect on research.

Consider, for example, the ranking of deep learning methods for uncertainty offered by Bui et al. [4]. They compare UCI rankings from multiple papers and calculate the average rankings. They find that Hamiltonian Monte Carlo (average rank 8.80 ± 1.38) and mean-field variational inference (average rank 7.50 ± 1.70) using the reparametrization trick perform best of the neural network models they consider (with the best performer being Deep Gaussian Processes). However, HMC is known not to scale to datasets with large data, a property which is not captured with the benchmark. Further, MC dropout is ranked second-to-last place with average rank 12.10 ± 0.64. Our results show that on a larger-scale dataset, MC dropout has better uncertainty estimates than mean-field variational inference and they have almost identical performance when all datapoints are retained. Moreover, HMC would not scale to this data at all.

By relying too much on UCI as a benchmark, we give a misleading impression of relative performance, which will cause researchers to prioritise the wrong approaches. A number of more computationally intensive extensions to MFVI have emerged since Bui et al. [4] produced their analysis, while less work has gone into building on the methodology of the more computationally parsimonious Bayesian deep learning approaches like deep ensembles or MC dropout [32, 44, 45, 38]. It seems likely that this is partly shaped by the fact that UCI is the predominant benchmark.

Our new benchmark and systematic comparison of BDL tools will offer a way for new methods to demonstrate their effectiveness on large-scale problems, making it easier to publish results that engage with the sorts of problems that deep learning has proven to be effective at, and which downstream users are seeking.

References

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[2] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

[3] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*, 2015.

[4] Thang Bui, Daniel Hernández-Lobato, Jose Hernandez-Lobato, Yingzhen Li, and Richard Turner. Deep gaussian processes for regression using approximate expectation propagation. In *International conference on machine learning*, pages 1472–1481, 2016.

[5] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, et al. Opportunities and obstacles for deep learning in biology and medicine. *bioRxiv*, page 142760, 2018.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.

[7] Yarin Gal. Uncertainty in deep learning. *University of Cambridge*, 2016.

[8] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

[9] Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. In *Advances in Neural Information Processing Systems*, pages 3584–3593, 2017.

[10] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[11] Alex Graves. Practical variational inference for neural networks. In *Advances in neural information processing systems*, pages 2348–2356, 2011.

[12] José Miguel Hernández-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International Conference on Machine Learning*, pages 1861–1869, 2015.

[13] Geoffrey Hinton and Drew Van Camp. Keeping neural networks simple by minimizing the description length of the weights. In *in Proc. of the 6th Ann. ACM Conf. on Computational Learning Theory*. Citeseer, 1993.

[14] Kaggle. Diabetic retinopathy detection challenge, 2015. URL `https://www.kaggle.com/c/diabetic-retinopathy-detection`.

[15] Gregory Kahn, Adam Villaflor, Vitchyr Pong, Pieter Abbeel, and Sergey Levine. Uncertainty-aware reinforcement learning for collision avoidance. *arXiv preprint arXiv:1702.01182*, 2017.

[16] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

[17] Michael Kampffmeyer, Arnt-Børre Salberg, and Robert Jenssen. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2016 IEEE Conference on*, pages 680–688. IEEE, 2016.

[18] Alex Kendall and Roberto Cipolla. Modelling uncertainty in deep learning for camera relocalization. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 4762–4769. IEEE, 2016.

[19] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017.

[20] Alex Kendall and Yarin Gal. What Uncertainties Do WE Need in Bayesian Deep Learning for Computer Vision? *Advances In Neural Information Processing Systems1*, 30, 2017.

[21] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[24] Igor Kononenko. Inductive and bayesian learning in medical diagnosis. *Applied Artificial Intelligence an International Journal*, 7(4):317–337, 1993.

[25] Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109, 2001.

[26] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

[27] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pages 396–404, 1990.

[28] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[29] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[30] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Nature Scientific Reports*, 7(1):17816, 2017.

[31] Laurence Perreault Levasseur, Yashar D Hezaveh, and Risa H Wechsler. Uncertainties in parameters estimated with neural networks: Application to strong gravitational lensing. *The Astrophysical Journal Letters*, 850(1):L7, 2017.

[32] Christos Louizos and Max Welling. Structured and Efficient Variational Deep Learning with Matrix Gaussian Posteriors. *International Conference on Machine Learning*, pages 1708–1716, 2016.

[33] David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

[34] Robert T McGibbon, Andrew G Taube, Alexander G Donchev, Karthik Siva, Felipe Hernández, Cory Hargus, Ka-Hei Law, John L Klepeis, and David E Shaw. Improving the accuracy of møller-plesset perturbation theory with neural networks. *The Journal of chemical physics*, 147 (16):161725, 2017.

[35] Jishnu Mukhoti, Pontus Stenetorp, and Yarin Gal. On the importance of strong baselines in bayesian deep learning. *arXiv preprint arXiv:1811.09385*, 2018.

[36] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 1995.

[37] Kirill Neklyudov, Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. Variance networks: When expectation does not meet your expectations. *arXiv preprint arXiv:1803.03764*, 2018.

[38] Changyong Oh, Kamil Adamczewski, and Mijung Park. Radial and Directional Posteriors for Bayesian Neural Networks. *arXiv*, February 2019. arXiv: 1902.02603.

[39] Carsten Peterson. A mean field theory learning algorithm for neural networks. *Complex systems*, 1:995–1019, 1987.

[40] Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[42] Frank Soboczenski, Michael D Himes, Molly D O'Beirne, Simone Zorzan, Atilim Gunes Baydin, Adam D Cobb, Daniel Angerhausen, Giada N Arney, and Shawn D Domagal-Goldman. Bayesian deep learning for exoplanet atmospheric retrieval. *arXiv preprint arXiv:1811.03390*, 2018.

[43] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[44] Shengyang Sun, Changyou Chen, and Lawrence Carin. Learning Structured Weight Uncertainty in Bayesian Neural Networks. *Artificial Intelligence and Statistics*, pages 1283–1292, 2017.

[45] Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional variational Bayesian Neural Networks. *International Conference on Learning Representations*, 2019.

[46] Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[47] Shijun Wang and Ronald M Summers. Machine learning and radiology. *Medical image analysis*, 16(5):933–951, 2012.

[48] Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. *arXiv preprint arXiv:1803.04386*, 2018.

[49] Daniel E Worrall, Clare M Wilson, and Gabriel J Brostow. Automated retinopathy of prematurity case detection with convolutional neural networks. In *Deep Learning and Data Labeling for Medical Applications*, pages 68–76. Springer, 2016.

[50] Anqi Wu, Sebastian Nowozin, Edward Meeds, Richard E Turner, José Miguel Hernández-Lobato, and Alexander L Gaunt. Deterministic variational inference for robust bayesian neural networks. *International Conference on Learning Representations*, 2019. URL `https://openreview.net/pdf?id=B1l08oAct7`.

[51] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.