Quantifying Aleatoric and Epistemic Uncertainty in Machine Learning: Are Conditional Entropy and Mutual Information Appropriate Measures? (Supplementary Material)

Lisa Wimmer[1,3] **Yusuf Sale**[2,3] **Paul Hofman**[2,3] **Bernd Bischl**[1,3] **Eyke Hüllermeier**[2,3]

[1]Department of Statistics, LMU Munich, Germany
[2]Institute of Informatics, LMU Munich, Germany
[3]Munich Center for Machine Learning (MCML), Germany

1 EXPERIMENTAL DETAILS

In the following, we list the most important training configurations used to generate our results. The full experimental code is hosted in a public repository[1].

Software Our codebase is written in `Python`. It chiefly relies on the `PyTorch` [Paszke et al., 2019], `PyTorch Lightning` [Lightning AI, 2023], `Laplace Redux` [Daxberger et al., 2021], and `scikit-learn` [Pedregosa et al., 2011] libraries.

Datasets The real-world computer vision tasks are `CIFAR10` [Krizhevsky, 2009] and `MNIST` [LeCun et al., 1998]. Both contain ten balanced classes. We further synthesize rectangles (white-on-black), where the class label is determined by whether height $>$ width or *vice versa*, and random non-convex polygons (white-on-black) with 3–5 vertices. These datasets comprise 60k (10k) training (test) samples. The tabular classification problem is created via `scikit-learn`'s `make_classification` function, using two features (and four classes. Here, we generate 6k (1k) training (test) samples.

Base learners Our probabilistic classifiers all combine some base learners into an explicit (deep ensemble, random forest) or implicit (Laplace approximation) ensemble. We train `EfficientNet-B7` (approx. 64m parameters; Tan and Le [2019]) for `CIFAR10` and a small convolutional network (three convolutional layers with ReLU activation; approx. 62k parameters) for `MNIST` and the rectangle/polygon images. In the tabular classification problem, we use a random forest with a maximum tree depth of ten as well as single-hidden-layer MLPs with a hidden layer size of ten, adopting the default parameters from `scikit-learn` unless stated otherwise. Ensemble size is set to $M = 10$.

[1]`https://github.com/lisa-wm/entropybaseduq`

Training Configurations We use an SGD optimizer (momentum 0.9), a learning rate schedule with cosine annealing, where the initial learning rate is set to 10^{-2}, and weight decay (5×10^{-4}). Training runs for a maximum of 200 epochs at batch size 256 with early stopping if validation loss does not improve over five consecutive epochs (evaluated on a validation set containing 10% of the training data).

2 ADDITIONAL RESULTS

2.1 INCREASING DATA NOISE

Compared to the ensemble of MLPs[2], the random forest (Fig. 1) reacts in both uncertainty components when class overlap is increased.

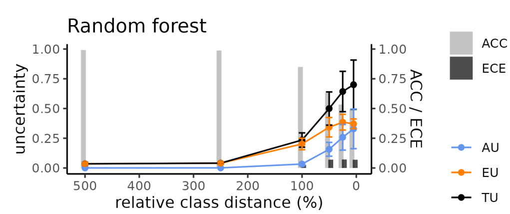

Figure 1: Entropy-based uncertainty for increasing class overlap (tabular data).

In order to simulate label noise, we randomly change classes for a varying share (1%–75%) of observations in the tabular classification task, leading to datasets as depicted in Fig. 2.

[2]In the tabular classification task, we bootstrap the data for the MLP ensemble to make it directly comparable to the random forest that relies on this technique.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

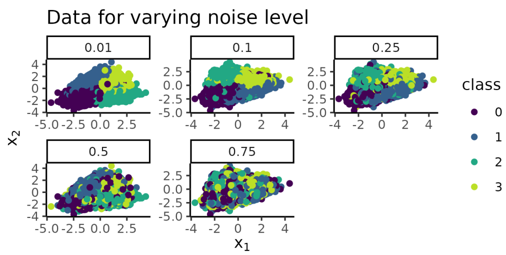

Figure 2: Tabular data with two features and four classes for increasing noise level.

Expected Behavior AU picks up with increasing noise level. Since learner capacity remains fixed, it is reasonable to assume that EU also rises to some extent when the decision boundaries become more complex with mounting degree of dataset contamination.

Observed Behavior As observed in the experiments modifying image resolution and class overlap, we find that AU duly increases for a rising noise level, though it remains moderate for the random forest even in the most extreme scenario (Fig. 3), where three out of four labels are assigned randomly. EU goes up slightly for the random forest, as presumed, but remains ultra-low for every value of the ablation with the MLP ensemble.

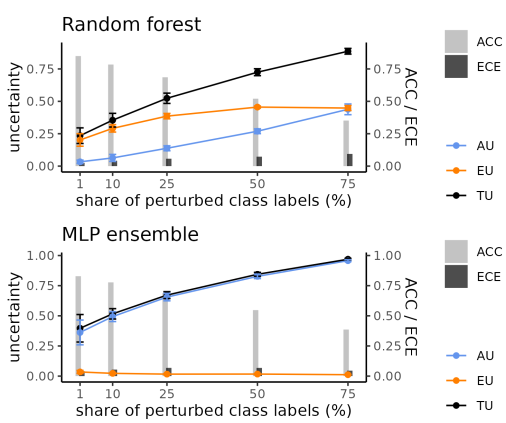

Figure 3: Entropy-based uncertainty for increasing label noise (tabular data).

2.2 NUMBER OF ENSEMBLE MEMBERS

We study for the tabular classification problem how different ensemble sizes (2–50) affect the uncertainty estimates.

Expected Behavior There should be no systematic pattern except for possible volatility for very small ensemble sizes, where the finite-ensemble estimator might have larger bias.

Observed Behavior The results are indeed fairly stable for different values of M (Fig. 4). Again, the overall levels of reported uncertainty differ considerably between the learners.

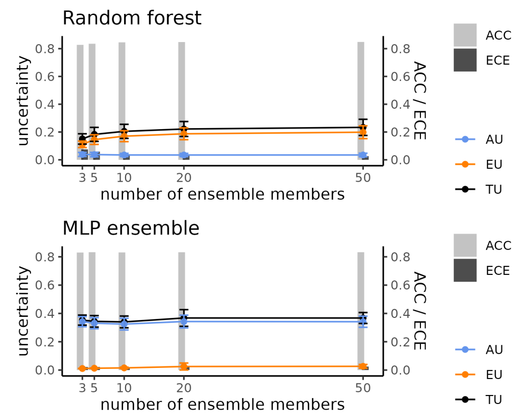

Figure 4: Entropy-based uncertainty for increasing number of ensemble members (tabular data).

We also compute the uncertainty measures for the computer vision tasks, where such large ensembles are prohibitively expensive, that result from using $M = 5$. Tables 1–8 show the uncertainty values as an average over all possible five-member ensembles that can be constructed from the ten original predictions (we can compute this *ex post* since ensemble size does not affect training for either of the used probabilistic learners: deep ensembles are trained in parallel with no shared loss propagation, and Laplace approximation is an inherently *ex-post* approach anyway). The results are quite robust here as well (with some exceptions for the particularly noisy settings, such as 1% sample size).

2.3 BASE LEARNER COMPLEXITY

Lastly, we investigate the effect of changing the base learner's capacity in the random forest and ensemble of MLPs. As a a proxy for capacity, we use maximum tree depth and hidden-layer size, respectively.

Expected Behavior Initially, AU should decrease when base learners get more capacity so they can fit more varied distributions, express their confidence more adequately and achieve better calibration. Similarly, the additional complexity might result in higher EU because the base learners have more freedom for disagreement.

Observed Behavior We find that AU indeed reduces considerably for more complex base learners (Fig. 5), especially for the random forest, which appears to overstate AU when the base learners are very simple (resulting in high calibration error). The strong effect is quite striking and might be overlooked as performance is relatively stable, again underlining that accuracy, calibration and uncertainty must be considered jointly. EU, on the other hand, does not change much – apparently, relation between capacity and reported AU is quite consistent across base learners and does not provoke more conflict when the ensemble members obtain more freedom.

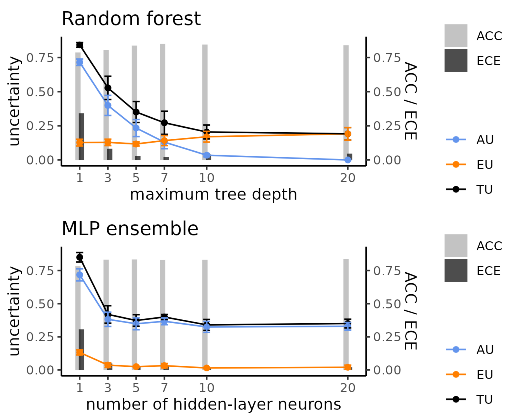

Figure 5: Increasing base learner complexity

Table 1: Results for sample size with ensemble of $M = 5$. Mean and standard deviation are obtained by aggregating over all possible ensembles of size five that can be sampled from the ten predictions of the original experiment.

Experiment	**Case**	**Probabilistic learner**	**Dataset**	**Measure**	**Mean**	**Standard deviation**	$M = 10$
sample size	1	Laplace approximation	MNIST	TU	0.5918	0.0451	0.7754
sample size	1	Laplace approximation	MNIST	AU	0.0091	0.0012	0.0091
sample size	1	Laplace approximation	MNIST	EU	0.5827	0.0449	0.7663
sample size	2	Laplace approximation	MNIST	TU	0.5794	0.0436	0.7419
sample size	2	Laplace approximation	MNIST	AU	0.0386	0.0037	0.0388
sample size	2	Laplace approximation	MNIST	EU	0.5408	0.0416	0.7031
sample size	5	Laplace approximation	MNIST	TU	0.5370	0.0416	0.6716
sample size	5	Laplace approximation	MNIST	AU	0.0631	0.0053	0.0634
sample size	5	Laplace approximation	MNIST	EU	0.4738	0.0391	0.6083
sample size	10	Laplace approximation	MNIST	TU	0.4578	0.0364	0.5760
sample size	10	Laplace approximation	MNIST	AU	0.0447	0.0039	0.0449
sample size	10	Laplace approximation	MNIST	EU	0.4131	0.0341	0.5311
sample size	50	Laplace approximation	MNIST	TU	0.1756	0.0247	0.2147
sample size	50	Laplace approximation	MNIST	AU	0.0354	0.0033	0.0355
sample size	50	Laplace approximation	MNIST	EU	0.1402	0.0221	0.1791
sample size	100	Laplace approximation	MNIST	TU	0.0793	0.0116	0.0947
sample size	100	Laplace approximation	MNIST	AU	0.0223	0.0022	0.0224
sample size	100	Laplace approximation	MNIST	EU	0.0570	0.0096	0.0723

Table 2: Results for sample size with ensemble of $M = 5$. Mean and standard deviation are obtained by aggregating over all possible ensembles of size five that can be sampled from the ten predictions of the original experiment.

Experiment	**Case**	**Probabilistic learner**	**Dataset**	**Measure**	**Mean**	**Standard deviation**	$M = 10$
sample size	1	deep ensemble	MNIST	TU	0.0487	0.0110	0.0518
sample size	1	deep ensemble	MNIST	AU	0.0343	0.0065	0.0344
sample size	1	deep ensemble	MNIST	EU	0.0144	0.0049	0.0174
sample size	2	deep ensemble	MNIST	TU	0.0381	0.0042	0.0404
sample size	2	deep ensemble	MNIST	AU	0.0257	0.0032	0.0258
sample size	2	deep ensemble	MNIST	EU	0.0124	0.0016	0.0146
sample size	5	deep ensemble	MNIST	TU	0.0212	0.0017	0.0223
sample size	5	deep ensemble	MNIST	AU	0.0152	0.0012	0.0153
sample size	5	deep ensemble	MNIST	EU	0.0060	0.0011	0.0070
sample size	10	deep ensemble	MNIST	TU	0.0137	0.0011	0.0144
sample size	10	deep ensemble	MNIST	AU	0.0098	0.0007	0.0099
sample size	10	deep ensemble	MNIST	EU	0.0039	0.0007	0.0045
sample size	50	deep ensemble	MNIST	TU	0.0082	0.0007	0.0087
sample size	50	deep ensemble	MNIST	AU	0.0051	0.0004	0.0051
sample size	50	deep ensemble	MNIST	EU	0.0031	0.0004	0.0036
sample size	100	deep ensemble	MNIST	TU	0.0067	0.0008	0.0072
sample size	100	deep ensemble	MNIST	AU	0.0042	0.0004	0.0042
sample size	100	deep ensemble	MNIST	EU	0.0025	0.0004	0.0030

Table 3: Results for sample size with ensemble of $M = 5$. Mean and standard deviation are obtained by aggregating over all possible ensembles of size five that can be sampled from the ten predictions of the original experiment.

Experiment	**Case**	**Probabilistic learner**	**Dataset**	**Measure**	**Mean**	**Standard deviation**	$M = 10$
sample size	1	Laplace approximation	CIFAR10	TU	0.8171	0.0629	0.8507
sample size	1	Laplace approximation	CIFAR10	AU	0.6339	0.0587	0.6364
sample size	1	Laplace approximation	CIFAR10	EU	0.1832	0.0290	0.2143
sample size	2	Laplace approximation	CIFAR10	TU	0.5742	0.0368	0.6030
sample size	2	Laplace approximation	CIFAR10	AU	0.4337	0.0277	0.4354
sample size	2	Laplace approximation	CIFAR10	EU	0.1405	0.0092	0.1676
sample size	5	Laplace approximation	CIFAR10	TU	0.3823	0.0247	0.4114
sample size	5	Laplace approximation	CIFAR10	AU	0.2449	0.0158	0.2459
sample size	5	Laplace approximation	CIFAR10	EU	0.1374	0.0089	0.1655
sample size	10	Laplace approximation	CIFAR10	TU	0.3000	0.0199	0.3268
sample size	10	Laplace approximation	CIFAR10	AU	0.1776	0.0117	0.1783
sample size	10	Laplace approximation	CIFAR10	EU	0.1224	0.0083	0.1485
sample size	50	Laplace approximation	CIFAR10	TU	0.1327	0.0089	0.1460
sample size	50	Laplace approximation	CIFAR10	AU	0.0724	0.0048	0.0727
sample size	50	Laplace approximation	CIFAR10	EU	0.0603	0.0043	0.0733
sample size	100	Laplace approximation	CIFAR10	TU	0.0690	0.0047	0.0736
sample size	100	Laplace approximation	CIFAR10	AU	0.0460	0.0030	0.0461
sample size	100	Laplace approximation	CIFAR10	EU	0.0231	0.0018	0.0275

Table 4: Results for sample size with ensemble of $M = 5$. Mean and standard deviation are obtained by aggregating over all possible ensembles of size five that can be sampled from the ten predictions of the original experiment.

Experiment	**Case**	**Probabilistic learner**	**Dataset**	**Measure**	**Mean**	**Standard deviation**	$M = 10$
sample size	1	deep ensemble	CIFAR10	TU	0.9022	0.0715	0.9425
sample size	1	deep ensemble	CIFAR10	AU	0.7073	0.1223	0.7100
sample size	1	deep ensemble	CIFAR10	EU	0.1949	0.0770	0.2324
sample size	2	deep ensemble	CIFAR10	TU	0.7189	0.0652	0.7750
sample size	2	deep ensemble	CIFAR10	AU	0.4410	0.0676	0.4428
sample size	2	deep ensemble	CIFAR10	EU	0.2778	0.0404	0.3322
sample size	5	deep ensemble	CIFAR10	TU	0.4204	0.0273	0.4523
sample size	5	deep ensemble	CIFAR10	AU	0.2519	0.0173	0.2529
sample size	5	deep ensemble	CIFAR10	EU	0.1685	0.0113	0.1995
sample size	10	deep ensemble	CIFAR10	TU	0.2826	0.0183	0.3072
sample size	10	deep ensemble	CIFAR10	AU	0.1545	0.0106	0.1551
sample size	10	deep ensemble	CIFAR10	EU	0.1281	0.0082	0.1521
sample size	50	deep ensemble	CIFAR10	TU	0.1318	0.0131	0.1458
sample size	50	deep ensemble	CIFAR10	AU	0.0617	0.0076	0.0620
sample size	50	deep ensemble	CIFAR10	EU	0.0701	0.0057	0.0838
sample size	100	deep ensemble	CIFAR10	TU	0.1064	0.0176	0.1187
sample size	100	deep ensemble	CIFAR10	AU	0.0480	0.0101	0.0482
sample size	100	deep ensemble	CIFAR10	EU	0.0585	0.0078	0.0706

Table 5: Results for image resolution with ensemble of $M = 5$. Mean and standard deviation are obtained by aggregating over all possible ensembles of size five that can be sampled from the ten predictions of the original experiment.

Experiment	**Case**	**Probabilistic learner**	**Dataset**	**Measure**	**Mean**	**Standard deviation**	$M = 10$
image resolution	5	Laplace approximation	MNIST	TU	0.7660	0.0571	0.8540
image resolution	5	Laplace approximation	MNIST	AU	0.3781	0.0387	0.3796
image resolution	5	Laplace approximation	MNIST	EU	0.3879	0.0418	0.4744
image resolution	10	Laplace approximation	MNIST	TU	0.6475	0.0463	0.6996
image resolution	10	Laplace approximation	MNIST	AU	0.3847	0.0348	0.3862
image resolution	10	Laplace approximation	MNIST	EU	0.2628	0.0317	0.3134
image resolution	25	Laplace approximation	MNIST	TU	0.1149	0.0099	0.1261
image resolution	25	Laplace approximation	MNIST	AU	0.0634	0.0048	0.0636
image resolution	25	Laplace approximation	MNIST	EU	0.0516	0.0057	0.0624
image resolution	50	Laplace approximation	MNIST	TU	0.0787	0.0087	0.0923
image resolution	50	Laplace approximation	MNIST	AU	0.0259	0.0024	0.0260
image resolution	50	Laplace approximation	MNIST	EU	0.0528	0.0065	0.0663
image resolution	100	Laplace approximation	MNIST	TU	0.0703	0.0108	0.0833
image resolution	100	Laplace approximation	MNIST	AU	0.0212	0.0024	0.0213
image resolution	100	Laplace approximation	MNIST	EU	0.0491	0.0085	0.0620

Table 6: Results for image resolution with ensemble of $M = 5$. Mean and standard deviation are obtained by aggregating over all possible ensembles of size five that can be sampled from the ten predictions of the original experiment.

Experiment	**Case**	**Probabilistic learner**	**Dataset**	**Measure**	**Mean**	**Standard deviation**	$M = 10$
image resolution	5	deep ensemble	MNIST	TU	0.7635	0.0487	0.7672
image resolution	5	deep ensemble	MNIST	AU	0.7585	0.0484	0.7615
image resolution	5	deep ensemble	MNIST	EU	0.0050	0.0008	0.0056
image resolution	10	deep ensemble	MNIST	TU	0.5378	0.0350	0.5413
image resolution	10	deep ensemble	MNIST	AU	0.5272	0.0344	0.5292
image resolution	10	deep ensemble	MNIST	EU	0.0107	0.0013	0.0121
image resolution	25	deep ensemble	MNIST	TU	0.0519	0.0039	0.0537
image resolution	25	deep ensemble	MNIST	AU	0.0421	0.0030	0.0423
image resolution	25	deep ensemble	MNIST	EU	0.0098	0.0012	0.0114
image resolution	50	deep ensemble	MNIST	TU	0.0088	0.0008	0.0093
image resolution	50	deep ensemble	MNIST	AU	0.0058	0.0004	0.0059
image resolution	50	deep ensemble	MNIST	EU	0.0030	0.0004	0.0035
image resolution	100	deep ensemble	MNIST	TU	0.0074	0.0006	0.0079
image resolution	100	deep ensemble	MNIST	AU	0.0046	0.0004	0.0046
image resolution	100	deep ensemble	MNIST	EU	0.0028	0.0003	0.0033

Table 7: Results for image resolution with ensemble of $M = 5$. Mean and standard deviation are obtained by aggregating over all possible ensembles of size five that can be sampled from the ten predictions of the original experiment.

Experiment	**Case**	**Probabilistic learner**	**Dataset**	**Measure**	**Mean**	**Standard deviation**	$M = 10$
image resolution	5	Laplace approximation	CIFAR10	TU	0.7137	0.0451	0.7171
image resolution	5	Laplace approximation	CIFAR10	AU	0.7093	0.0448	0.7122
image resolution	5	Laplace approximation	CIFAR10	EU	0.0043	0.0003	0.0049
image resolution	10	Laplace approximation	CIFAR10	TU	0.2676	0.0169	0.2697
image resolution	10	Laplace approximation	CIFAR10	AU	0.2593	0.0164	0.2603
image resolution	10	Laplace approximation	CIFAR10	EU	0.0083	0.0005	0.0095
image resolution	25	Laplace approximation	CIFAR10	TU	0.1151	0.0073	0.1176
image resolution	25	Laplace approximation	CIFAR10	AU	0.1007	0.0064	0.1011
image resolution	25	Laplace approximation	CIFAR10	EU	0.0144	0.0010	0.0165
image resolution	50	Laplace approximation	CIFAR10	TU	0.0835	0.0055	0.0890
image resolution	50	Laplace approximation	CIFAR10	AU	0.0545	0.0036	0.0547
image resolution	50	Laplace approximation	CIFAR10	EU	0.0290	0.0021	0.0343
image resolution	100	Laplace approximation	CIFAR10	TU	0.0690	0.0047	0.0736
image resolution	100	Laplace approximation	CIFAR10	AU	0.0460	0.0030	0.0461
image resolution	100	Laplace approximation	CIFAR10	EU	0.0231	0.0018	0.0275

Table 8: Results for image resolution with ensemble of $M = 5$. Mean and standard deviation are obtained by aggregating over all possible ensembles of size five that can be sampled from the ten predictions of the original experiment.

Experiment	**Case**	**Probabilistic learner**	**Dataset**	**Measure**	**Mean**	**Standard deviation**	$M = 10$
image resolution	5	deep ensemble	CIFAR10	TU	0.7351	0.0467	0.7425
image resolution	5	deep ensemble	CIFAR10	AU	0.7012	0.0446	0.7040
image resolution	5	deep ensemble	CIFAR10	EU	0.0339	0.0027	0.0386
image resolution	10	deep ensemble	CIFAR10	TU	0.3727	0.0248	0.3900
image resolution	10	deep ensemble	CIFAR10	AU	0.2752	0.0197	0.2763
image resolution	10	deep ensemble	CIFAR10	EU	0.0975	0.0067	0.1137
image resolution	25	deep ensemble	CIFAR10	TU	0.2084	0.0187	0.2265
image resolution	25	deep ensemble	CIFAR10	AU	0.1134	0.0126	0.1138
image resolution	25	deep ensemble	CIFAR10	EU	0.0950	0.0068	0.1127
image resolution	50	deep ensemble	CIFAR10	TU	0.1174	0.0088	0.1302
image resolution	50	deep ensemble	CIFAR10	AU	0.0527	0.0045	0.0529
image resolution	50	deep ensemble	CIFAR10	EU	0.0648	0.0046	0.0773
image resolution	100	deep ensemble	CIFAR10	TU	0.1045	0.0137	0.1160
image resolution	100	deep ensemble	CIFAR10	AU	0.0480	0.0084	0.0482
image resolution	100	deep ensemble	CIFAR10	EU	0.0565	0.0055	0.0679